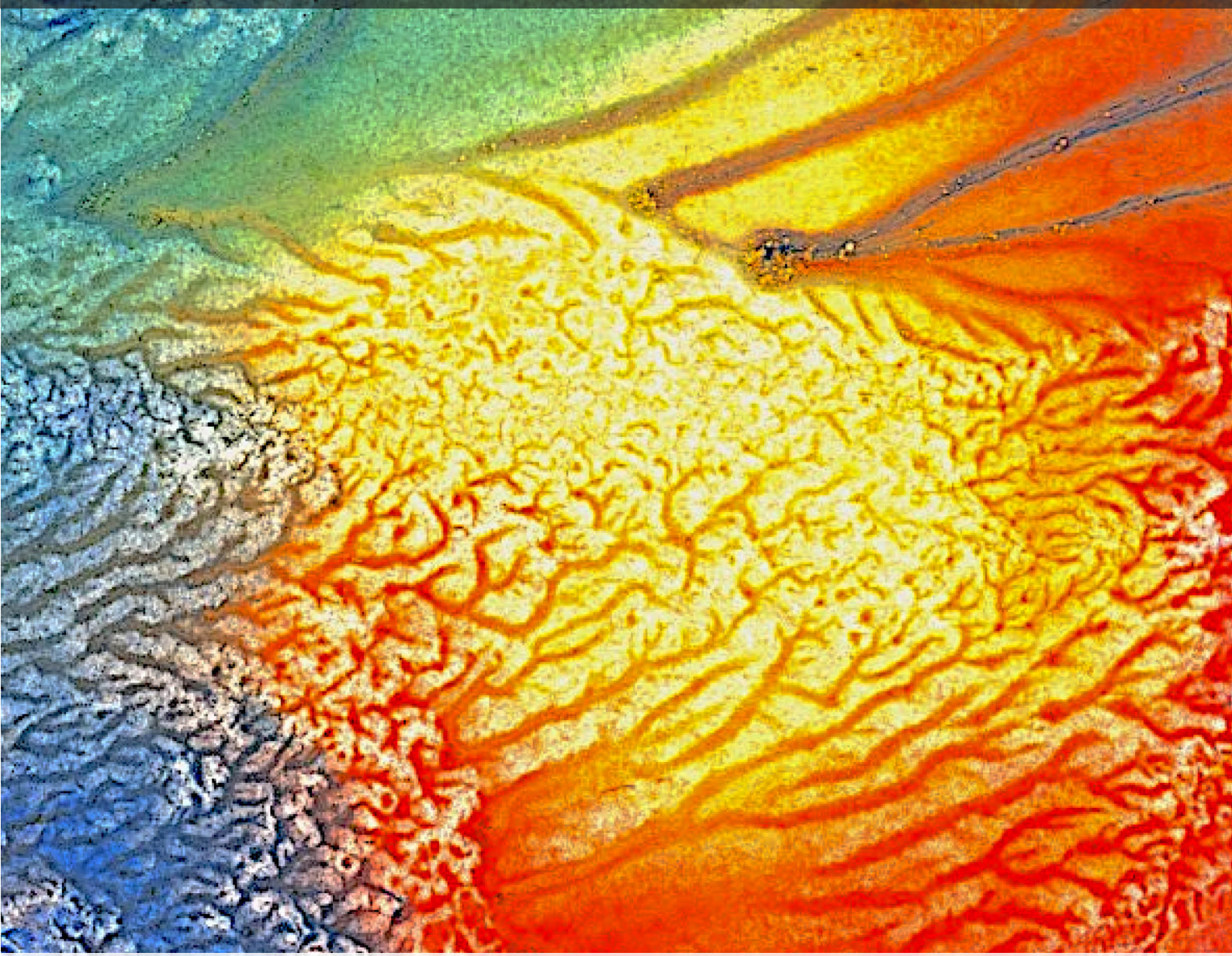
GENOME INVADING RNA NETWORKS

EDITED BY: Guenther Witzany and Luis Villarreal

PUBLISHED IN: Frontiers in Microbiology and Frontiers in Plant Science



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-477-8
DOI 10.3389/978-2-88945-477-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

GENOME INVADING RNA NETWORKS

Topic Editors:
Guenther Witzany, Telos–Philosophische Praxis, Bürmoos, Austria
Luis Villarreal, University of California, Irvine, United States

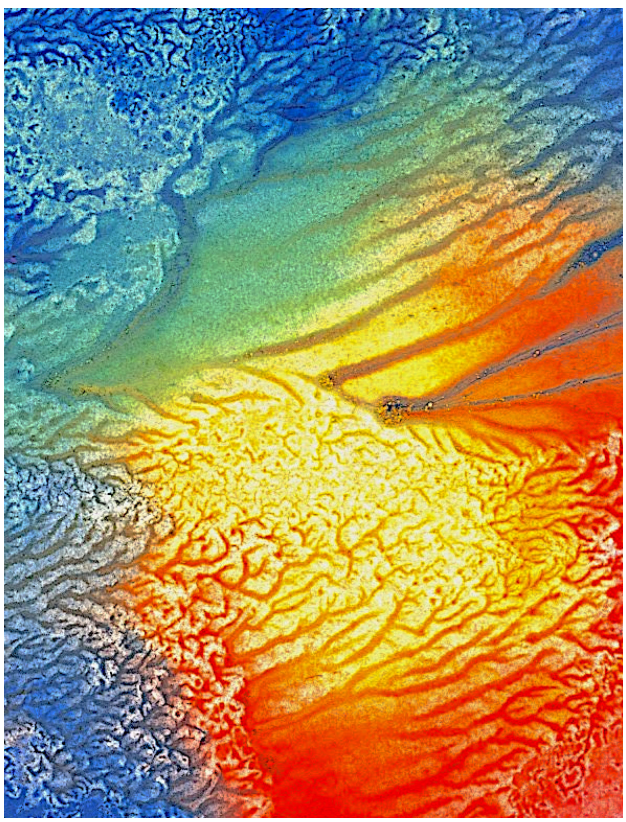


Image: Helga Oman

A new paradigmatic understanding of evolution, genetic novelty, code-generating, genome-formatting factors, infectious RNA Networks, viruses and other natural genetic content operators.

Citation: Witzany, G., Villarreal, L., eds. (2018). Genome Invading RNA Networks. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-477-8

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 27 March 2018
doi: 10.3389/fmicb.2018.00581

Editorial: Genome Invading RNA Networks

Luis P. Villarreal[1] *and Guenther Witzany*[2]*

[1] Center for Virus Research, University of California, Irvine, Irvine, CA, United States, [2] Telos–Philosophische Praxis, Buermoos, Austria

Keywords: RNA Networks, genetic identities, regulatory RNAs, infectious agents, Natural Genetic Content Operators

Editorial on the Research Topic

Genome Invading RNA Networks

It has been long accepted that newly acquired biological information is mostly derived from random, error-based' events (Eigen, 1971). However, the serial nature of acquiring such random events makes it very difficult to account for the origin or modification of regulatory networks. There is now abundant empirical evidence establishing the crucial role of non-coding DNA (acting through the expression of RNA with its complex biology) to create regulatory control (Mattick, 2003; Atkins et al., 2011). Along with the parallel comeback of regulatory RNA in virology, RNA is now at center stage in how we think about complex organisms (Koonin et al., 2006; Atkins et al., 2011).

Regulatory RNAs derive from infectious events and can co-operate, build communities, generate nucleotide sequences de novo and insert/delete themselves into host genetic content (Villarreal, 2005; Koonin, 2009). In this sense genome invading RNA networks determine host genetic identities (self-recognition) throughout all kingdoms including the virosphere (Britten, 2004; Marraffini and Sontheimer, 2010; Villarreal, 2011a). But inclusion of a transmissive viral RNA biology differs fundamentally from conventional thinking in that it represents a vertical domain of life providing vast amounts of linked information not derived from direct ancestors (Villarreal, 2014). Interestingly single RNA stem loops react as physico-chemical entities exclusively, whereas with the network-cooperation of various RNA stem-loops in a module-like manner biological selection emerges (Manrubia and Briones, 2007; Vaidya, 2012; Higgs and Lehman, 2015). Additionally co-operating RNAs outcompete selfish genetic parasites (Hayden and Lehman, 2006; Vaidya et al., 2012).

Thus, we can argue, that for DNA based organisms, the introduction of infective collectives of RNA groups are a central driving force of evolution. Such RNA groups are co-adapted from persistent infectious agents and now serve as regulatory tools in nearly all cellular processes (Witzany, 2016) as documented in several retrovirus derived mobile genetic elements (Brosius, 1999; Villarreal, 2011b; Chuong et al., 2016). Additionally, the resulting productive RNA networks constantly produce new sequence space (i.e., complex regulation) which not only further serve as adaptation tools for their cell-based host organisms but also provides crucial roles in evolutionary novelty (Villarreal, 2011b). This RNA productivity results out of the empirical fact that a single RNA sequence can fold into different and unrelated secondary structures with different functions in a (environmentally determined) context-depending way (Schultes and Bartel, 2000).

Infection derived RNAs serve as the agents of regulatory networks in the cellular transcriptome (Feschotte, 2008; Briones et al., 2009; Koonin, 2009; Villarreal and Witzany, 2010). Without transcription from the genetic storage medium of DNA into the living world of such RNA agents, no relevant genetic process in the cellular transcriptome can be initiated (Volff, 2006). RNAs,

Edited by:
David Gilmer,
Université de Strasbourg, France

Reviewed by:
Cristina Romero-López,
Institute of Parasitology and Biomedicine "López-Neyra" (CSIC), Spain
Roland Marquet,
Architecture et Réactivité de l'ARN, France

****Correspondence:***
Guenther Witzany
witzany@sbg.at

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *11 February 2018*
Accepted: *14 March 2018*
Published: *27 March 2018*

Citation:
Villarreal LP and Witzany G (2018) Editorial: Genome Invading RNA Networks. Front. Microbiol. 9:581. doi: 10.3389/fmicb.2018.00581

with their inherent repeat syntax, format the expression of coding sequences and organize the coherent line-up of timely coordinated steps of replication (Shapiro and von Sternberg, 2005). The transport of genetic information to the progeny cells is also coordinated by these agents (Spadafora, 2017). Furthermore, they are crucial for the cooperation between networks of RNA-stem loops to constitute important nucleoprotein complexes such as ribosome, spliceosome, and editosome (Witzany, 2011). Therefore, such RNA groups are essential for complex order of genome constructions (Witzany, 2014).

Additionally of interest is that infectious non-coding RNAs insert preferentially in non-coding DNA areas, whereas coding DNA usually is not the target (Bushman, 2003; Mitchell et al., 2004; Bartel, 2009). In this perspective the non-coding DNA is the preferred habitat to settle down by infectious RNAs, e.g., y-chromosome in human genomes (Shapiro, 2002; Villarreal, 2009; Lambowitz and Zimmerly, 2011). This may indicate that the preferred change in evolutionary processes occurs in regulatory sections and not in the information storage coding for proteins, the main source for "mutations" in previous theoretical concepts of evolution (Villarreal and Witzany, 2013).

Frontiers Research Topic Genome Invading RNA Networks highlights various RNA networks being active in host genomes.

Sablok et al. discussed classification, identification and roles of tRNA derived smallRNAs across plants and their potential involvement in abiotic and biotic stresses. Wang et al. investigated how retrotransposon insertion polymorphisms can impact human health and disease. Moelling et al. demonstrated that RNase H-like activities of retroviruses, TEs, and phages, have built up innate and adaptive immune systems throughout all domains of life. Liu at al. summarize recent advances in understanding the roles of miRNAs involved in the plant defense against viruses and viral counter-defense. Malicki et al. review three retrotransposon classes that might represent a domestication of the selfish elements. Habibi and Salmani exemplified direct action of RNA networks in shaping the genome. Pecman et al. compared two different approaches for detection and discovery of plant viruses and viroids. Nagata et al. found that sequence changes in the RNase H domain and the reverse transcriptase connection domain are responsible for subtype classification. Zinad et al. suggest that natural antisense transcripts interfere with their corresponding sense transcript to elicit concordant and discordant regulation. Ottesen et al. describe how the abundance of Alu-like sequences may contribute toward Survival Motor Neuron gene pathogenesis. Ariza-Mateos and Gómez show how RNA viruses mimic key factors of the host cell. Spadafora found that spermatozoa act as collectors of somatic information and as delivering vectors to the next generation. Frías-Lasserre et al. demonstrate how current epigenetic advances on non-coding RNAs has changed the perspective on evolutionary relevant variations. Scolaro et al. demonstrate that evolutive processes for viruses are now interpreted as coordinated phenomenon that leads to global non-random remodeling of the population. Seligmann and Raoult found that ribosomal RNA stem-loop hairpins resemble those formed by viruses and short parasitic repeats infesting bacterial genomes. Fu et al. provide deep insights into the molecular mechanisms of influenza virus infection.

More and more empirical evidence establishes the crucial role of natural genetic content editors such as viruses and RNA networks to create genetic novelty, complex regulatory control, epigenetics, genetic identity, immunity, inheritance vectors, new sequence space, evolution of complex organisms and evolutionary transitions (Villarreal and Witzany, 2015; Chuong et al., 2016, Spadafora, this issue).

Genetic identities of RNA networks such as e.g., group I introns, group II introns, viroids, RNA viruses, retrotransposons, LTRs, non-LTRs, SINEs, LINEs, Alus invade and even persist in host genomes (Villarreal, 2009). Also mixed networks of RNA- and DNA viruses derived parts that integrate into host genomes have been found (Stedman, 2015), not forgetting persistent retroviral infections and the essential roles of reverse transcriptases and related RNase H endonucleases (Moelling and Broecker, 2015).

Highly dynamic RNA-Protein networks such as ribosome, editosome and spliceosome together with several context-dependent sequence modificating interactions, such as pseudo-knotting, frame-shifting, loop-kissing, by-passing translation generate a large variety of RNA regulatory functions out of a given DNA content (Cao et al., 2014; Denzler et al., 2014; Peselis and Serganov, 2014; Samatova et al., 2014; Keam and Hutvagner, 2015; Atkins et al., 2016).

There are reasonable expectations that this new empirically based perspective on the evolution of genetic novelty and biological information will have more explanatory power in the future than the "error-replication" narrative of the last century.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Atkins, J. F., Gesteland, R. F., and Cech, T. R. (eds.). (2011). *RNA Worlds. From Life's Origin to Diversity in Gene Regulation*. New York, NY: Cold Spring Harbor Laboratory Press.

Atkins, J. F., Loughran, G., Bhatt, P. R., Firth, A. E., and Baranov, P. V. (2016). Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* 44, 7007–7078. doi: 10.1093/nar/gkw530

Briones, C., Stich, M., and Manrubia, S. C. (2009). The dawn of the RNA world: toward functional complexity through ligation of random RNA oligomers. *RNA* 15, 743–749. doi: 10.1261/rna.1488609

Britten, R. J. (2004). Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16825–16830. doi: 10.1073/pnas.0406985101

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 238, 115–134. doi: 10.1016/S0378-1119(99)00227-9

Bushman, F. D. (2003). Targeting survival: Integration site selection by retroviruses and LTR-retrotransposons. *Cell* 115, 135–138. doi: 10.1016/S0092-8674(03)00760-8

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

Cao, S., Xu, X., and Chen, S. J. (2014). Predicting structure and stability for RNA complexes with intermolecular loop-loop base-pairing. *RNA* 20, 835–845. doi: 10.1261/rna.043976.113

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351, 1083–1087. doi: 10.1126/science.aad5497

Denzler, R., Agarwal, V., Stefano, J., Bartel, D., and Stoffel, M. (2014). Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol. Cell* 54, 766–776. doi: 10.1016/j.molcel.2014.03.045

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58, 465–523. doi: 10.1007/BF00623322

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405. doi: 10.1038/nrg2337

Hayden, E. J., and Lehman, N. (2006). Self-assembly of a group I intron from inactive oligonucleotide fragments. *Chem. Biol.* 13, 909–918. doi: 10.1016/j.chembiol.2006.06.014

Higgs, P. G., and Lehman, N. (2015). The RNA World: molecular cooperation at the origins of life. *Nat. Rev. Genet.* 16, 7–17. doi: 10.1038/nrg3841

Keam, S. P., and Hutvagner, G. (2015). tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. *Life* 5, 1638–1651. doi: 10.3390/life5041638

Koonin, E. V. (2009). On the origin of cells and viruses: primordial virus world scenario. *Ann. N.Y. Acad. Sci.* 1178, 47–64. doi: 10.1111/j.1749-6632.2009.04992.x

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Direct.* 19, 1–29. doi: 10.1186/1745-6150-1-1

Lambowitz, A. M., and Zimmerly, S. (2011). Group II introns: Mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* 3:a003616 doi: 10.1101/cshperspect.a003616

Manrubia, S. C., and Briones, C. (2007). Modular evolution and increase of functional complexity in replicating RNA molecules. *RNA* 13, 97–107. doi: 10.1261/rna.203006

Marraffini, L. A., and Sontheimer, E. J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* 463, 568–571. doi: 10.1038/nature08703

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25, 930–939. doi: 10.1002/bies.10332

Mitchell, R. S., Beitzel, B. F., Schroder, A. R., Shinn, P., Chen, H., Berry, C. C., et al. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* 2:e234. doi: 10.1371/journal.pbio.0020234

Moelling, K., and Broecker, F. (2015). The reverse transcriptase-RNase H: from viruses to antiviral defense. *Ann. N.Y. Acad. Sci.* 1341, 126–135. doi: 10.1111/nyas.12668

Peselis, A., and Serganov, A. (2014). Structure and function of pseudoknots involved in gene expression control. *Wiley Interdiscip. Rev. RNA* 5, 803–822. doi: 10.1002/wrna.1247

Samatova, E., Konevega, A. L., Wills, N. M., Atkins, J. F., and Rodnina, M. V. (2014). High-efficiency translational bypassing of non-coding nucleotides specified by mRNA structure and nascent peptide. *Nat. Commun.* 5, 4459. doi: 10.1038/ncomms5459

Schultes, E. A., and Bartel, D. P. (2000). One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* 289, 448–452. doi: 10.1126/science.289.5478.448

Shapiro, J. A. (2002). Repetitive DNA, genome system architecture and genome reorganization. *Res. Microbiol.* 153, 447–453. doi: 10.1016/S0923-2508(02)01344-X

Shapiro, J. A., and von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biol. Rev. Camb. Philos. Soc.* 80, 227–250. doi: 10.1017/S1464793104006657

Spadafora, C. (2017). The "evolutionary field" hypothesis. Non-Mendelian transgenerational inheritance mediates diversification and evolution. *Prog. Biophys. Mol. Biol.* 134, 27–37. doi: 10.1016/j.pbiomolbio.2017.12.001

Stedman, K. M. (2015). Deep Recombination: RNA and ssDNA Virus Genes in DNA Virus and Host Genomes. *Annu. Rev. Virol.* 2, 203–217. doi: 10.1146/annurev-virology-100114-055127

Vaidya, N. (2012). *Spontaneous Cooperative Assembly of Replicative Catalytic RNA Systems*. Dissertations and Theses. Portland state University

Vaidya, N., Manapat, M. L., Chen, I. A., Xulvi-Brunet, R., Hayden, E. J., and Lehman, N. (2012). Spontaneous network formation among cooperative RNA replicators. *Nature* 491, 72–77. doi: 10.1038/nature11549

Villarreal, L. P. (2005). *Viruses and the Evolution of Life*. Washington, DC: ASM Press.

Villarreal, L. P. (2009). *Origin of Group Identity. Viruses, Addiction and Cooperation*, New York, NY: Springer.

Villarreal, L. P. (2011a). Viruses and host evolution: virus-mediated self identity. *Adv. Exp. Med. Biol.* 738, 185–217. doi: 10.1007/978-1-4614-1680-7_12

Villarreal, L. P. (2011b). Viral ancestors of antiviral systems. *Viruses* 3, 1933–1958. doi: 10.3390/v3101933

Villarreal, L. P. (2014). Force for ancient and recent life: viral and stem-loop RNA consortia promote life. *Ann. N.Y. Acad. Sci.* 1341, 25–34. doi: 10.1111/nyas.12565

Villarreal, L. P., and Witzany, G. (2010). Viruses are essential agents within the roots and stem of the tree of life. *J. Theor. Biol.* 262, 698–710. doi: 10.1016/j.jtbi.2009.10.014

Villarreal, L. P., and Witzany, G. (2013). Rethinking quasispecies theory: from fittest type to cooperative consortia. *World J. Biol. Chem.* 4, 79–90. doi: 10.4331/wjbc.v4.i4.79

Villarreal, L. P., and Witzany, G. (2015). When competing viruses unify: evolution, conservation, and plasticity of genetic identities. *J. Mol. Evol.* 80, 305–318. doi: 10.1007/s00239-015-9683-y

Volff, J. N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* 28, 913–922. doi: 10.1002/bies.20452

Witzany, G. (2011). The agents of natural genome editing. *J. Mol. Cell. Biol.* 3, 181–189. doi: 10.1093/jmcb/mjr005

Witzany, G. (2014). RNA sociology: group behavioral motifs of RNA consortia. *Life (Basel).* 4, 800–818. doi: 10.3390/life4040800

Witzany, G. (2016). Crucial steps to life: From chemical reactions to code using agents. *Biosystems* 140, 49–57. doi: 10.1016/j.biosystems.2015.12.007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 02 August 2017
doi: 10.3389/fmicb.2017.01418

Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes

Lu Wang[1,2,3], *Emily T. Norris*[1,2,3] *and I. K. Jordan*[1,2,3]*

[1] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, United States, [2] PanAmerican Bioinformatics Institute, Cali, Colombia, [3] Applied Bioinformatics Laboratory, Atlanta, GA, United States

The human genome hosts several active families of transposable elements (TEs), including the Alu, LINE-1, and SVA retrotransposons that are mobilized via reverse transcription of RNA intermediates. We evaluated how insertion polymorphisms generated by human retrotransposon activity may be related to common health and disease phenotypes that have been previously interrogated through genome-wide association studies (GWAS). To address this question, we performed a genome-wide screen for retrotransposon polymorphism disease associations that are linked to TE induced gene regulatory changes. Our screen first identified polymorphic retrotransposon insertions found in linkage disequilibrium (LD) with single nucleotide polymorphisms that were previously associated with common complex diseases by GWAS. We further narrowed this set of candidate disease associated retrotransposon polymorphisms by identifying insertions that are located within tissue-specific enhancer elements. We then performed expression quantitative trait loci analysis on the remaining set of candidates in order to identify polymorphic retrotransposon insertions that are associated with gene expression changes in B-cells of the human immune system. This progressive and stringent screen yielded a list of six retrotransposon insertions as the strongest candidates for TE polymorphisms that lead to disease via enhancer-mediated changes in gene regulation. For example, we found an SVA insertion within a cell-type specific enhancer located in the second intron of the *B4GALT1* gene. *B4GALT1* encodes a glycosyltransferase that functions in the glycosylation of the Immunoglobulin G (IgG) antibody in such a way as to convert its activity from pro- to anti-inflammatory. The disruption of the *B4GALT1* enhancer by the SVA insertion is associated with down-regulation of the gene in B-cells, which would serve to keep the IgG molecule in a pro-inflammatory state. Consistent with this idea, the *B4GALT1* enhancer SVA insertion is linked to a genomic region implicated by GWAS in both inflammatory conditions and autoimmune diseases, such as systemic lupus erythematosus and Crohn's disease. We explore this example and the other cases uncovered by our genome-wide screen in an effort to illuminate how retrotransposon insertion polymorphisms can impact human health and disease by causing changes in gene expression.

Keywords: transposable elements, retrotransposons, Alu, L1, SVA, gene expression, gene regulation, GWAS

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos - Philosophische Praxis, Austria

Reviewed by:
Dixie Mager,
BC Cancer Agency, Canada
Jianwei Wang,
China Academy of Chinese Medical Sciences, China

***Correspondence:**
I. K. Jordan
king.jordan@biology.gatech.edu

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *12 May 2017*
Accepted: *13 July 2017*
Published: *02 August 2017*

Citation:
Wang L, Norris ET and Jordan IK (2017) Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. Front. Microbiol. 8:1418. doi: 10.3389/fmicb.2017.01418

INTRODUCTION

At least one half of the human genome sequence is derived from the replication and insertion of retrotransposons – RNA agents that transpose among chromosomal locations via the reverse transcription of RNA intermediates (Lander et al., 2001; de Koning et al., 2011). The vast majority of retrotransposon-related sequences in the human genome are derived from ancient insertion events and are no longer capable of transposition. Nevertheless, there are several families of human retrotransposons that remain active. The most abundant active families of human retrotransposons are the Alu (Batzer and Deininger, 1991; Batzer et al., 1991), LINE-1 (L1) (Kazazian et al., 1988; Brouha et al., 2003), and SVA (Ostertag et al., 2003; Wang et al., 2005) retrotransposons; recent evidence indicates that a smaller number of HERV-K endogenous retroviruses also remain capable of transposition (Wildschutte et al., 2016).

Sequences from active retrotransposon families generate insertional polymorphisms within and between human populations by means of germline transposition events. In this way, ongoing retrotranspositional activity of these RNA agents serves as an important source of human genetic variation. Retrotransposons are further distinguished by the fact that they are known to impact the regulation of human genes in a number of different ways (Feschotte, 2008; Rebollo et al., 2012; Chuong et al., 2017). Nevertheless, the joint phenotypic implications of retrotransposon generated human genetic variation, coupled with their capacity for genome regulation, have yet to be fully explored. We previously studied the implications of somatic retrotransposition for the etiology of cancer vis a vis retrotransposon induced regulatory changes in tumor suppressor genes (Clayton et al., 2016). For the current study, we were curious to understand how insertion polymorphisms generated by human retrotransposon activity may be related to commonly expressed health and disease phenotypes.

In one sense, a link between retrotransposon activity and disease is already well established. Active human retrotransposons were originally discovered due to the deleterious effects of element insertions (Kazazian et al., 1988). There are 124 genetic diseases that have been demonstrated to be caused by retrotransposon insertions, including cystic fibrosis (Alu), hemophilia A (L1) and X-linked dystonia-parkinsonism (SVA) (Hancks and Kazazian, 2012; Hancks and Kazazian, 2016). However, these cases represent so-called Mendelian diseases caused by very deleterious mutations that are expressed with high penetrance. Disease causing mutations of this kind are extremely rare and do not segregate among populations as common genetic polymorphisms. Complex multi-factorial diseases, on the other hand, are associated with more common genetic variants that exert their effects in a probabilistic as opposed to a deterministic manner. The contribution of common retrotransposon polymorphisms to complex health and disease related phenotypes has yet to be systematically explored.

Given the known connection between retrotransposon activity and genetic disease, we hypothesized that retrotransposon insertion polymorphisms may also contribute to inter-individual phenotypic differences that are associated with common diseases that have complex, multi-factorial genetic etiology. Since we previously showed that retrotransposon insertions contribute to inter-individual and population-specific differences in human gene regulation (Wang L. et al., 2016), we also hypothesized that the impact of retrotransposon insertion polymorphisms on human health could be mediated by gene regulatory effects.

Previously, it has only been possible to investigate the impact of retrotransposon polymorphisms on disease phenotypes for a limited number of individuals owing to the number of genomes that were available (Rishishwar et al., 2017). For the current study, we leveraged the accumulation of whole genome sequence and expression datasets, along with data on single nucleotide polymorphism (SNP) disease-associations, in order to perform a population level genome-wide screen for retrotransposon polymorphisms that are linked to complex health- and disease-related phenotypes.

MATERIALS AND METHODS

Polymorphic Transposable Element (PolyTE) and SNP Genotypes

Human polymorphic TE (polyTE) insertion presence/absence genotypes for whole genome sequences of 445 individuals from five human populations were accessed from the phase 3 variant release of the 1000 Genomes Project (1KGP) (Altshuler et al., 2015). Whole genome SNP genotypes were taken for the same set of individuals. The phase 3 variant release corresponds to the human genome reference sequence build GRCh37/hg19, and the 5 human populations are YRI: Yoruba in Ibadan, Nigeria from Africa, CEU: Utah Residents (CEPH) with Northern and Western European Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland and TSI: Toscani in Italia from Europe. We chose these genome sequence datasets because they have matching RNA-seq data for the same individuals [see Expression Quantitative Trait Locus (eQTL) Analysis section]. The YRI population was taken to represent the African continental population group (AFR), and the four populations from Europe (CEU, FIN, GBR, and TSI) were grouped together as the European (EUR) continental population group for downstream analysis (**Figure 1**). PolyTE insertion genotypes were characterized by the 1KGP Structural Variation Group using the program MELT as previously described (Sudmant et al., 2015). Previously, we performed an independent validation the performance of this program for human polyTE insertion variant calling from whole genome sequences (Rishishwar et al., 2016). The polyTE genotype data were downloaded via the 1000 Genomes Project ftp hosted by the NCBI: http://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/ For a given polyTE insertion site in the genome, there are three possible presence/absence genotype values for an individual genome: 0-no polyTE insertion (homozygote absent), 1-a single polyTE insertion (heterozygote), and 2-two polyTE insertions (homozygote present). PolyTE genotypes were used for eQTL analysis as described in section "Expression Quantitative Trait Locus (eQTL) Analysis". For each of the two continental population groups, only polyTE insertions

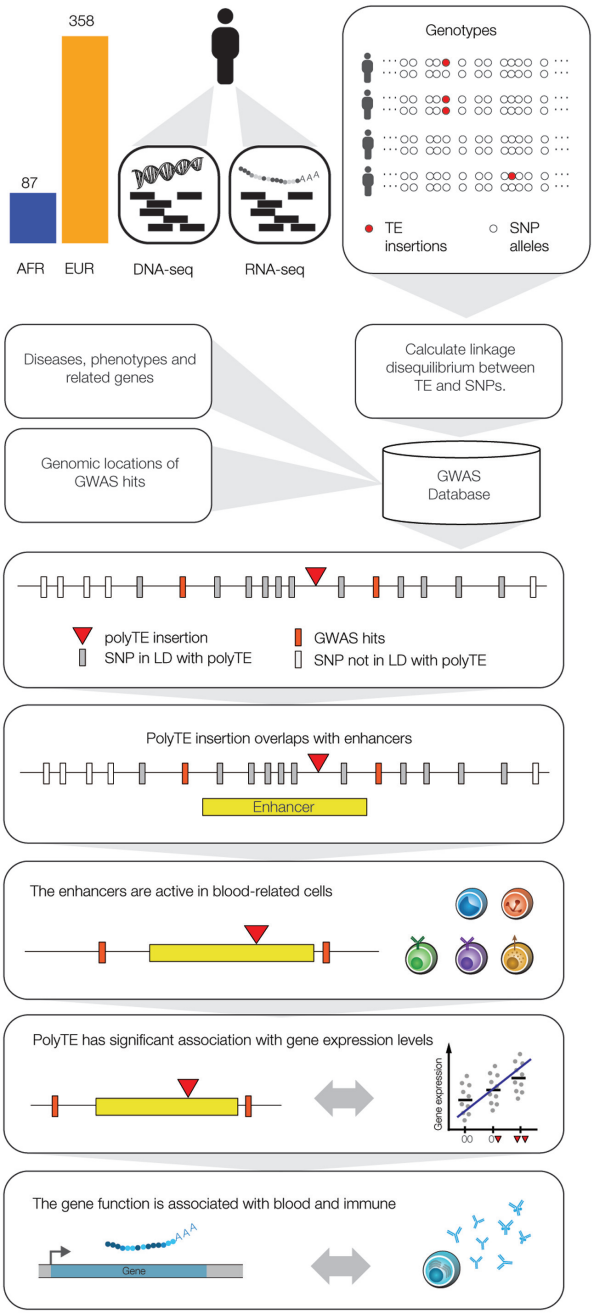


FIGURE 1 | Integrative data analysis used to screen for polyTE disease-associations. Matched DNA-seq and RNA-seq samples were analyzed for 445 individuals, 87 Africans (AFR – blue) and 358 Europeans (EUR – orange). Genome-wide genotypes were characterized for polyTE insertions (presence/absence) and SNPs, and the linkage disequilibrium (LD) structure for polyTE insertions and SNP alleles was characterized for all samples. The NHGRI-EBI GWAS database was mined to extract SNP disease-associations and related information on diseases, phenotypes, genes and SNP genomic locations. A series of filters was applied to screen for a set of high-confidence polyTE disease-associations. PolyTEs were evaluated for: (1) minor allele frequency (MAF), (2) linkage to disease-associated SNPS (i.e., GWAS hits), (3) overlap with tissue-specific enhancers, (4) associations with gene expression, and (5) functional relevance for blood- and immune system-related diseases.

and SNPs with greater than 5% minor allele frequencies (MAF) were used for the downstream analysis to ensure both the confidence of genotype calls and the reliability of the association analyses. Minor alleles for TEs are assumed to be the insertion present allele, since the ancestral state for any polyTE insertion site corresponds to the absence of an insertion (Rishishwar et al., 2015).

PolyTE-SNP Linkage Analysis

The GCTA program (version 1.25.0) was used to estimate the linkage disequilibrium (LD) structure for polyTEs and SNPs in genomic regions centered at each polyTE insertion site. For each polyTE insertion site, pairwise correlations (r) between the target polyTE insertion alleles and all SNP alleles in the same LD block were computed across all individual genome samples. A correlation (r) significance P-value threshold of 0.05 was used to identify all SNPs considered to be in LD with each polyTE insertion. Pairwise distances between polyTE insertion sites and linked SNPs were calculated as the number of base pairs between each polyTE insertion site and all linked SNP locations.

Genome Wide Association Studies (GWAS) for Disease

Associations between human genetic variants (SNPs) and health- or disease-related phenotypes were explored using the NHGRI-EBI GWAS database (MacArthur et al., 2017). GWAS database SNPs with genome-wide association values of $P < 10^{-5}$ were taken for analysis, and the genomic location, specific health- or disease-related phenotype, identity of the risk allele and original reporting publications were recorded for each associated SNP. The GWAS SNPs were screened for LD with polyTE insertions as described in section "PolyTE-SNP Linkage Analysis" to yield a set of candidate disease-linked polyTE insertions for further analysis.

Evaluating polyTE Regulatory Potential

The regulatory potential for polyTE insertions was evaluated by considering their co-location with known enhancer sequences. Active enhancers for 127 cell-types and tissues were characterized by the Roadmap Epigenomics Project using the ChromHMM program (Ernst and Kellis, 2012; Roadmap Epigenomics et al., 2015). ChromHMM integrates multiple genome-wide chromatin datasets (i.e., epigenomes), such as ChIP-seq of various histone modifications, using a multivariate Hidden Markov Model to identify the locations of tissue-specific enhancers based on their characteristic chromatin states. The data files with genomic locations for enhancers across all 127 epigenomes were accessed through the project website at http://mitra.stanford.edu/kundaje/leepc12/web_portal/chr_state_learning.html. The genomic locations of polyTE insertions that are in LD with disease-associated SNPs were compared with the genomic locations for enhancers from the 127 epigenomes, and polyTE insertions found to be located within active enhancer elements were considered to have regulatory potential. A subset of 27 epigenomes characterized for cells and tissues related to

blood and the immune system – such as T cells, B cells and hematopoietic stem cells – were selected for downstream eQTL analysis [see Expression Quantitative Trait Locus (eQTL) Analysis].

The overall relative regulatory potential of polyTE insertions in a given epigenome *i* is quantified as:

$$r_{\mathrm{i}} = \frac{t_{\mathrm{i}}}{s_{\mathrm{i}}}$$

where t_{i} is the proportion of polyTEs that are co-located with an enhancer element in a given epigenome *i*, and s_{i} is the proportion of SNPs from polyTE LD blocks that overlap with an enhancer element in the same epigenome *i*.

Physical associations between TE-enhancer insertions and nearby gene promoter regions were evaluated with chromatin-chromatin interaction map data, based on several different data sources, including 4C, 5C, ChIA-PET, and Hi-C, using the Chromatin Chromatin Space Interaction (CCSI) database at http://songyanglab.sysu.edu.cn/ccsi/ (Xie et al., 2016).

Expression Quantitative Trait Locus (eQTL) Analysis

Associations between polyTE insertion genotypes and tissue-specific gene expression levels were characterized using eQTL analysis (**Figure 1**). PolyTE insertion presence/absence genotypes were characterized as described in section "Polymorphic Transposable Element (polyTE) and SNP Genotypes". RNA-seq gene expression data for the same 445 individual genome samples used for polyTE genotype characterization were taken from the GEUVADIS RNA sequencing project. Genome-wide expression levels were measured for the same lymphoblastoid cell lines, i.e., Epstein–Barr virus (EBV) transformed B-lymphocytes (B cells), as used for DNA-seq analysis in the 1KGP. RNA isolation, library preparation, sequencing and read-to-genome mapping were performed as previously described (Lappalainen et al., 2013). As with the polyTE genotype data, the RNA-seq reads were mapped to the human genome build GRCh37/hg19. The process of gene expression normalization and quantification based on these RNA-seq data has been extensively validated as part of the GEUVEDIS project (t Hoen et al., 2013). The GEUVADIS RNA-seq data were used to compute gene expression levels for ENSEMBL gene models as previously described (Flicek et al., 2013). Normalization of gene expression levels was done using a combination of a modified reads per kilobase per million mapped reads (RPKM) approach followed by the probabilistic estimation of expression residuals (PEER) method as previously described (Stegle et al., 2012). This procedure has been shown to eliminate batch effects among different RNA-seq samples and to reduce the overall variance across samples, thereby ensuring the most accurate and comparable gene expression level inferences among samples. The normalized gene expression levels were accessed from the GUEVADIS project ftp server hosted at the EBI: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/.

PolyTE insertions that are (1) linked to at least one disease-associated SNP, and (2) located within a blood- or immune system-related enhancer were taken as a candidate set for eQTL analysis with the lymphoblastoid cell line RNA-seq data. PolyTE insertion presence/absence genotypes were regressed against gene expression levels to identify eQTLs (TE-eQTLs) using the program Matrix eQTL (Shabalin, 2012). Matrix eQTL was run using the additive linear (least squares model) option with gender and population used as covariates. This was done for all possible pairs of polyTE insertion sites from the candidate set and all genes. *Cis* vs. *trans* TE-eQTLs were defined later as polyTE insertion sites that fall inside (*cis*) or outside (*trans*) 1 megabase from gene boundaries. *P*-values were calculated for all TE-eQTL associations, and FDR *q*-values were then calculated to correct for multiple statistical tests. The genome-wide significant TE-eQTL association threshold was set at FDR $q < 0.05$, corresponding to $P = 4.7 \times 10^{-7}$ (AFR) and $P = 2.6 \times 10^{-7}$ (EUR).

Interrogation of Disease-Associated Gene Function and Association Consistency

The potential functional impacts of disease-associated TE-eQTL were evaluated via comparison of annotated gene functions and reported GWAS phenotypes for polyTE-linked SNPs. Gene functions were taken from the NCBI Entrez gene summaries, and GWAS phenotypes were taken from the original literature where the associations were reported. Genes that were found to be functionally related to GWAS reported health- or disease-related phenotypes were further checked for the direction of association. If the GWAS SNP-gene pair shows the same direction of association as the polyTE-gene pair, then the pair was included in the final set of significant gene-polyTEs association pairs (**Table 1**). For each gene in the final set, its tissue-specific expression levels across 18 tissues, including 4 blood- and immune-related tissues, were taken from the Illumina BodyMap and GTEx projects (Flicek et al., 2013; Mele et al., 2015; Rivas et al., 2015).

RESULTS

We used a genome-scale data analysis approach to explore the potential impact of human genetic variation generated by the activity of TEs on health and disease (**Figure 1**). This approach entailed an integrative analysis of (1) TE insertion polymorphisms, (2) SNPs, (3) SNP-disease associations, (4) tissue-specific enhancers, (5) eQTL, and (6) gene function/expression profiles. The rationale behind this approach was to employ a series of successive genome-wide filters, which would converge on a set of high-confidence TE insertion polymorphisms that are most likely to impact health- or disease-related phenotypes. Our analysis started with 5,845 polyTE insertions, with MAF > 0.05 for two continental population groups (European and African), and converged on a final set of seven high-confidence TE disease-association candidates (**Figure 2**). The final set of seven health/disease-implicated TE insertion polymorphisms that we found are distinguished by their linkage to disease-associated SNPs as well as their regulatory and functional properties. We describe the results and

TABLE 1 | PolyTE disease associations.

PolyTE[a]	Chr[b]	Pos[b]	GWAS SNP[c]	GWAS Phenotype[c]	GWAS gene[c]	GWAS *P*-value[c]	#Enhancer Overlaps[d]	eGene[d]	eQTL *P*-value[d]	eQTL Type[d]
Alu-2829	3	154966214	rs13064954	Diabetic retinopathy	*LINC00881, CCNL1*	7.00E-07	1	*LILRA1*	5.94e-10	*Trans*
Alu-5072	6	32589834	rs4530903	Lymphoma	*TRNAI25*	2.00E-08	20	*HLA-DRB5*	8.49e-13	*Cis*
Alu-5075	6	32657952	rs2858870	Nodular sclerosis Hodgkin lymphoma	*TRNAI25*	8.00E-18	15	*HLA-DQB1-AS1*	1.36e-11	*Cis*
SVA-282	6	33030313	rs3077	Chronic hepatitis B infection	*HLA-DPA1*	5.00E-39	6	*HLA-DPB2*	1.05e-13	*Cis*
SVA-401	9	33130564	rs10758189	IgG glycosylation	*B4GALT1*	2.00E-06	4	*B4GALT1*	4.47e-20	*Cis*
SVA-438	10	17712792	rs6602203	Glucose homeostasis traits	*ST8SIA6-AS1, PRPF38AP2*	5.00E-06	7	*TMEM236*	1.30e-07	*Cis*

[a]*PolyTE insertion identifier following the nomenclature of the 1KGP structural variation group.* [b]*Genomic location of the polyTE.* [c]*Information on the linked disease-associated GWAS SNP, disease phenotype and gene.* [d]*Information on the regulatory potential of the polyTE insertion based on co-location with enhancers and eQTL analysis.*

implications for each step in our TE disease-association screen in the sections below.

Linkage Disequilibrium for PolyTEs and Disease-Associated SNPs

The genomic locations of polyTE insertions were characterized for 445 individuals from one African (AFR) and four European (EUR) populations as described in section "Polymorphic Transposable Element (polyTE) and SNP Genotypes" of the Materials and Methods. This was done for the most common families of active human TEs: Alu, L1, and SVA. For each polyTE insertion location, individual genotypes were characterized as homozygous absent (0), heterozygous (1), or homozygous present (2). The distributions of polyTE insertion genotypes among individuals from each population were used to screen for polyTEs that are found at relatively high MAF > 0.05. The LD structure of the resulting common polyTE insertions, with adjacent common SNPs (also at MAF > 0.05), was then defined using correlation analysis across individual genome samples (see Materials and Methods section PolyTE-SNP Linkage Analysis). In addition, the genomic locations of common polyTE insertion variants and their linked SNPs were compared to the locations of disease-associated SNPs reported in the NHGRI-EBI GWAS database [see Materials and Methods Genome Wide Association Studies (GWAS) for Disease]. Linkage correlation coefficients between all polyTE insertions analyzed here and GWAS SNPs are shown in **Supplementary Table S1**.

Distributions of LD correlations between polyTEs and adjacent SNPs were compared separately for non-disease-associated vs. disease-associated SNPs. For all three families of active human TEs, in both the AFR and EUR population groups, polyTEs are found in significantly higher LD with disease-associated SNPs compared to non-disease-associated SNPs (**Figure 3A**). In addition, polyTE variants are located closer to disease-associated SNPs than non-disease-associated SNPs for the EUR population group (**Figure 3B**). A similar enrichment was not seen for the AFR population group, which may be attributed to the lower number of samples available for analysis for this group (AFR = 87 vs. EUR = 358). Indeed, when a larger number of AFR samples, which do not have matched RNA-seq data, were used for the same linkage analysis, the results were qualitatively identical to those seen for the EUR samples analyzed here. Taken together, these results indicate that polyTEs are more likely to be tightly linked to disease-associated SNPs compared to adjacent linked SNPs from the same LD blocks, suggesting

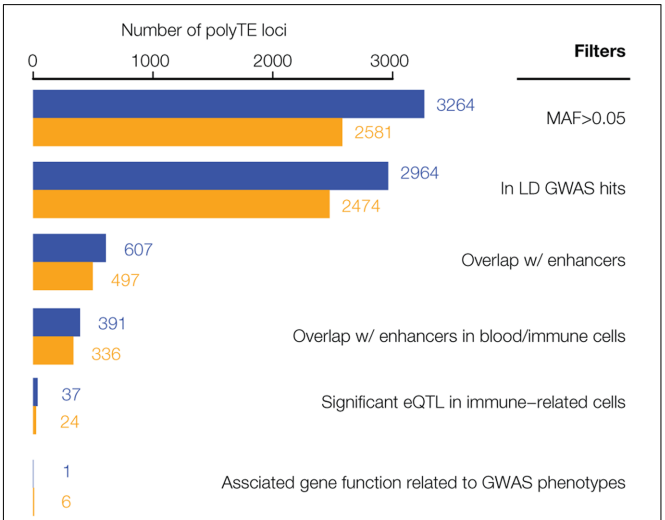


FIGURE 2 | Results of the genome-wide screen for polyTE disease-associations. As illustrated in **Figure 1**, a series of filters was applied to screen for a final set of high-confidence polyTE disease-associations. The number of polyTE insertions that remain after the application of each successive filter is shown for the African (AFR – blue) and European (EUR – orange) population groups.

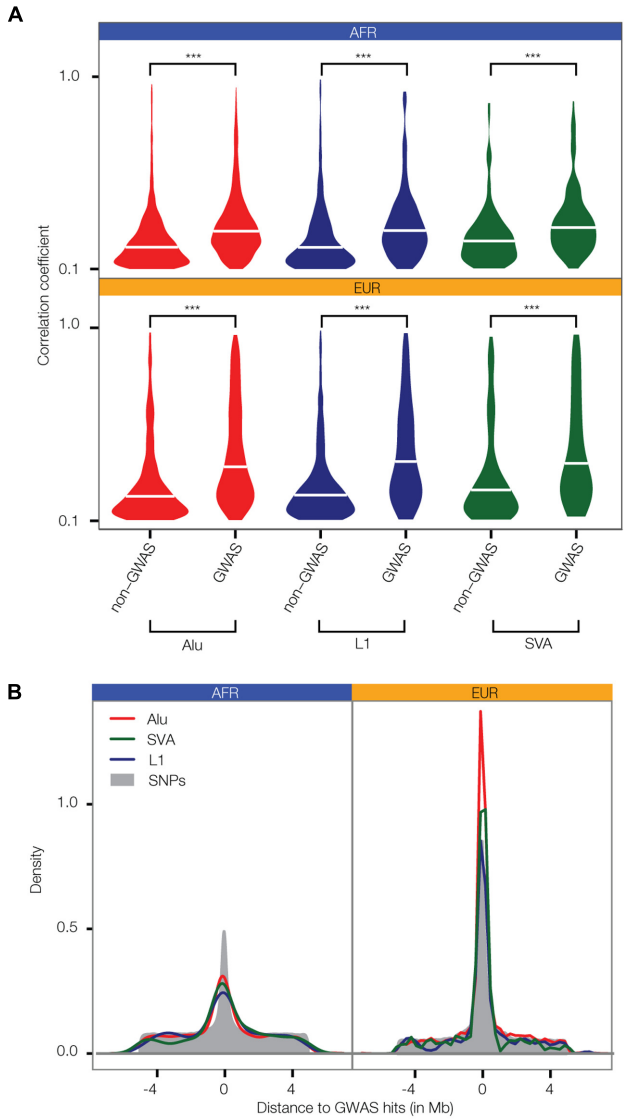


FIGURE 3 | Linkage between polyTE insertions and SNP disease-associations from GWAS. **(A)** Distributions of the allele correlation coefficients (*r*) are shown for (1) polyTE insertions and non-GWAS SNPs, and (2) polyTE insertions and GWAS SNPs. Higher correlation coefficients indicate tighter linkage. The significance of the differences for the non-GWAS SNP vs. the GWAS SNP correlation coefficient distributions are indicated (Kolmogorov–Smirnov test; $***P < 2.4 \times 10^{-5}$). **(B)** Density distributions of the distance between polyTE insertions and SNPs to the nearest GWAS disease-associated SNP. Correlation coefficient **(A)** and density **(B)** distributions are shown separately for the Alu (red), L1 (blue), and SVA (green) TE families in the African (AFR – blue) and European (EUR – orange) population groups.

a possible role in disease etiology for some TE variants. This is notable in light of the facts that (1) the TE genotypes were not considered in the initial association studies, and (2) TE insertions entail substantially larger-scale genetic variants than SNPs. Thus, polyTEs found on the same haplotypes as disease-associated SNPs may be expected to have an even greater impact on health- and disease-outcomes in some cases.

Co-location of Disease-Linked PolyTEs with Tissue-Specific Enhancers

Given the fact that TEs are known to participate in human gene regulation via a wide variety of mechanisms (Feschotte, 2008; Rebollo et al., 2012; Chuong et al., 2017), we hypothesized that polyTEs may impact disease by virtue of gene regulatory effects. The regulatory potential of polyTEs linked to disease-associated SNPs was first evaluated by searching for insertions that are co-located with tissue-specific enhancers. The locations of enhancers were characterized for 127 cell- and tissue-types based on their chromatin signatures as described in section "Evaluating polyTE Regulatory Potential" of the Materials and Methods. An example of an enhancer co-located with a disease-linked polyTE is shown for an Alu element that is inserted 5′ to the Immunoglobulin Heavy Variable 2-26 (*IGHV2-26*) encoding gene (**Figure 4A**). We found a total 607 disease-linked polyTEs co-located with enhancers in the AFR population group and 437 in EUR group; 391 (AFR) and 336 (EUR) of those enhancers correspond to blood- or immune-related tissues (**Figure 2**). Details on the co-localization of polyTE insertions and the enhancers characterized for each epigenome are shown in **Supplementary Table S2**.

We estimated the overall regulatory potential for disease-linked polyTEs in all cell- and tissue-types by computing the relative ratio of enhancer co-located insertions as described in section "Evaluating polyTE Regulatory Potential" of the Materials and Methods. The results of this analysis are considered separately for the blood- and immune-related tissues (**Figure 4B**) and all other tissues from which enhancers were characterized. Enhancer co-located disease-linked polyTEs from blood/immune cell-types show higher overall regulatory potential than ones that are co-located with enhancers characterized for the other tissue-types (**Figures 4C,D**). These results suggest that the set of disease-linked polyTEs studied here may have a disproportionate impact on immune-related diseases, and we focused our subsequent efforts on this subset of health conditions.

Expression Associations for Disease-Linked and Enhancer Co-located PolyTEs

We further evaluated the regulatory potential of the polyTEs that were found to be both disease-linked and co-located with blood- and immune-related enhancers using an eQTL approach [see Materials and Methods section Expression Quantitative Trait Locus (eQTL) analysis]. Genotypes for this subset of polyTEs from the 445 genome samples analyzed here were regressed against gene expression levels characterized from lymphoblastoid cell lines for the same individuals. Quantile–quantile (Q–Q) plots comparing the observed vs. expected *P*-values for the e-QTL analysis in the AFR and EUR population groups are shown in **Figure 5A**, revealing a number of statistically significant associations that are likely to be true-positives. There are 83 (AFR) and 42 (EUR) genome-wide significant TE-eQTL (**Supplementary Table S3**), and they are enriched in genomic

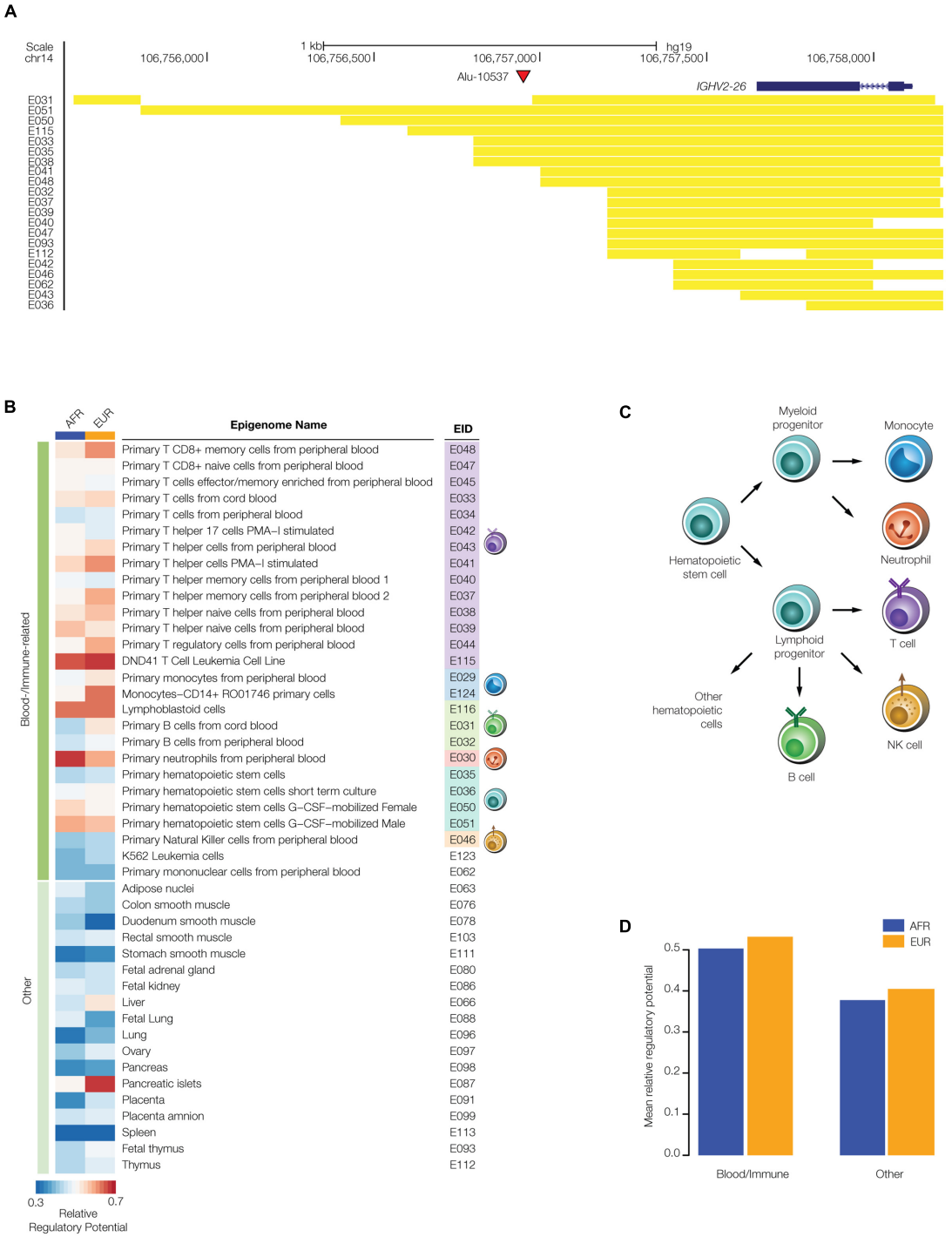


FIGURE 4 | Regulatory potential of disease-linked polyTE insertions. PolyTE insertions linked to disease-associated SNPs were evaluated for their co-location with tissue-specific enhancers. **(A)** UCSC Genome Browser screen capture showing an example of a polyTE insertion – Alu-10537 – that overlaps with a number of tissue specific enhancers. The genomic location of the Alu insertion on chromosome 14, downstream of the *IGHV2-26* gene (blue gene model), is indicated with a red arrow. The genomic locations of co-located enhancers, characterized based on chromatin signatures from a variety of tissue-specific epigenomes, are indicated with yellow bars. **(B)** Heatmap showing the relative regulatory potential (see Materials and Methods section Evaluating polyTE Regulatory Potential) of polyTE insertions for a variety of tissue-specific epigenomes. Blood- and immune-related tissues are shown separately from examples of the other tissue types analyzed here. **(C)** Developmental lineage of immune-related cells for which enhancer genomic locations were characterized. **(D)** The mean relative regulatory potential for disease-linked polyTE insertions is shown for blood- and immune-related tissues compared to all other tissue-types analyzed here. Values for the African (AFR – blue) and European (EUR – orange) population groups are shown separately.

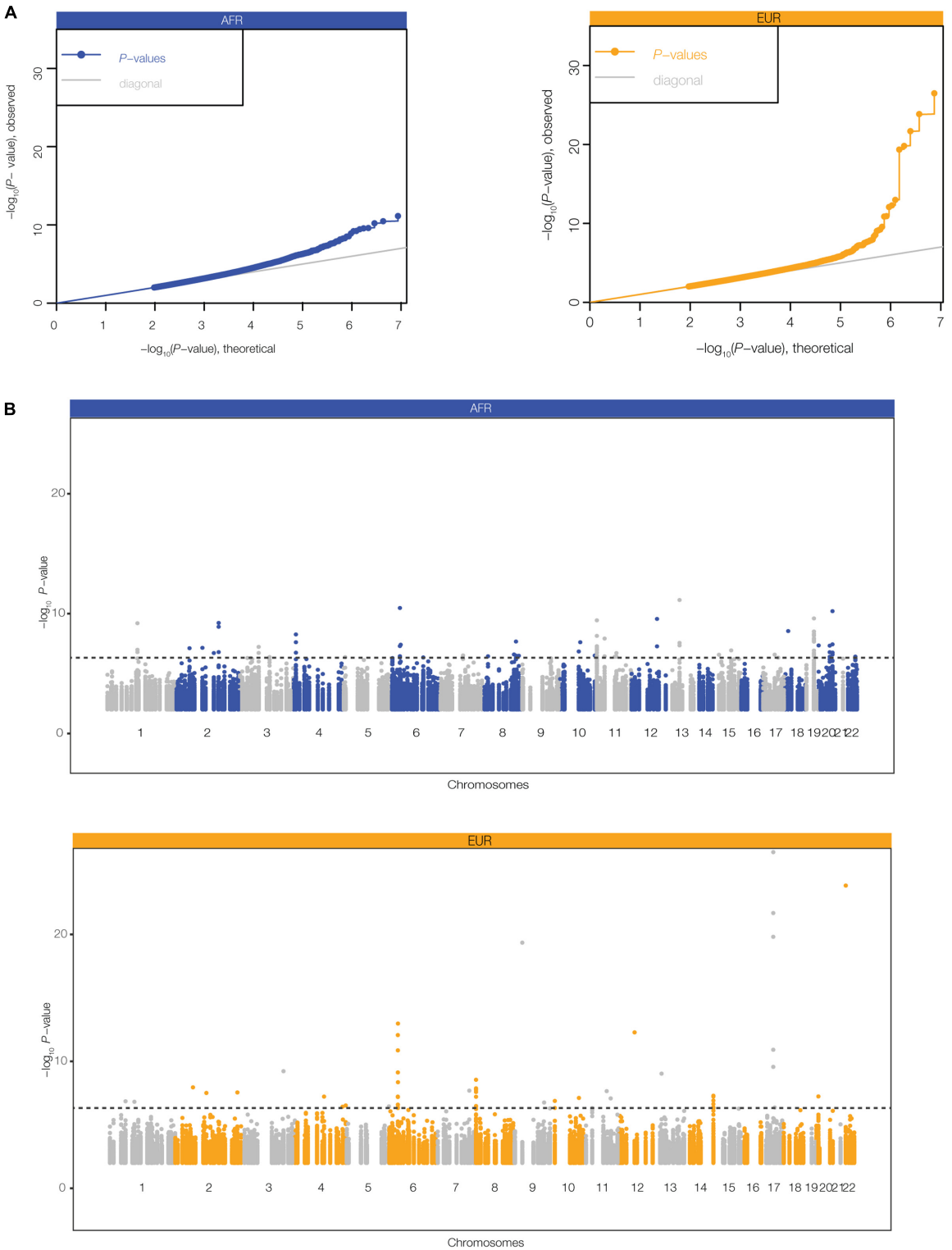


FIGURE 5 | Expression quantitative trait (eQTL) analysis for disease-linked polyTEs. eQTL analysis was performed by regressing lymphoblastoid gene expression levels against polyTE insertion genotypes for the for the African (AFR – blue) and European (EUR – orange) individuals analyzed here. **(A)** Quantile–quantile (Q–Q) plots relating the observed (*y*-axis) to the expected (*x*-axis) TE-eQTL log transformed *P*-values. **(B)** Manhattan plots showing the genomic distributions of TE-eQTL log transformed *P*-values. The dashed line corresponds to a false discovery rate (FDR) threshold of $q < 0.05$, corresponding to $P = 4.7 \times 10^{-7}$ (AFR) and $P = 2.6 \times 10^{-7}$ (EUR).

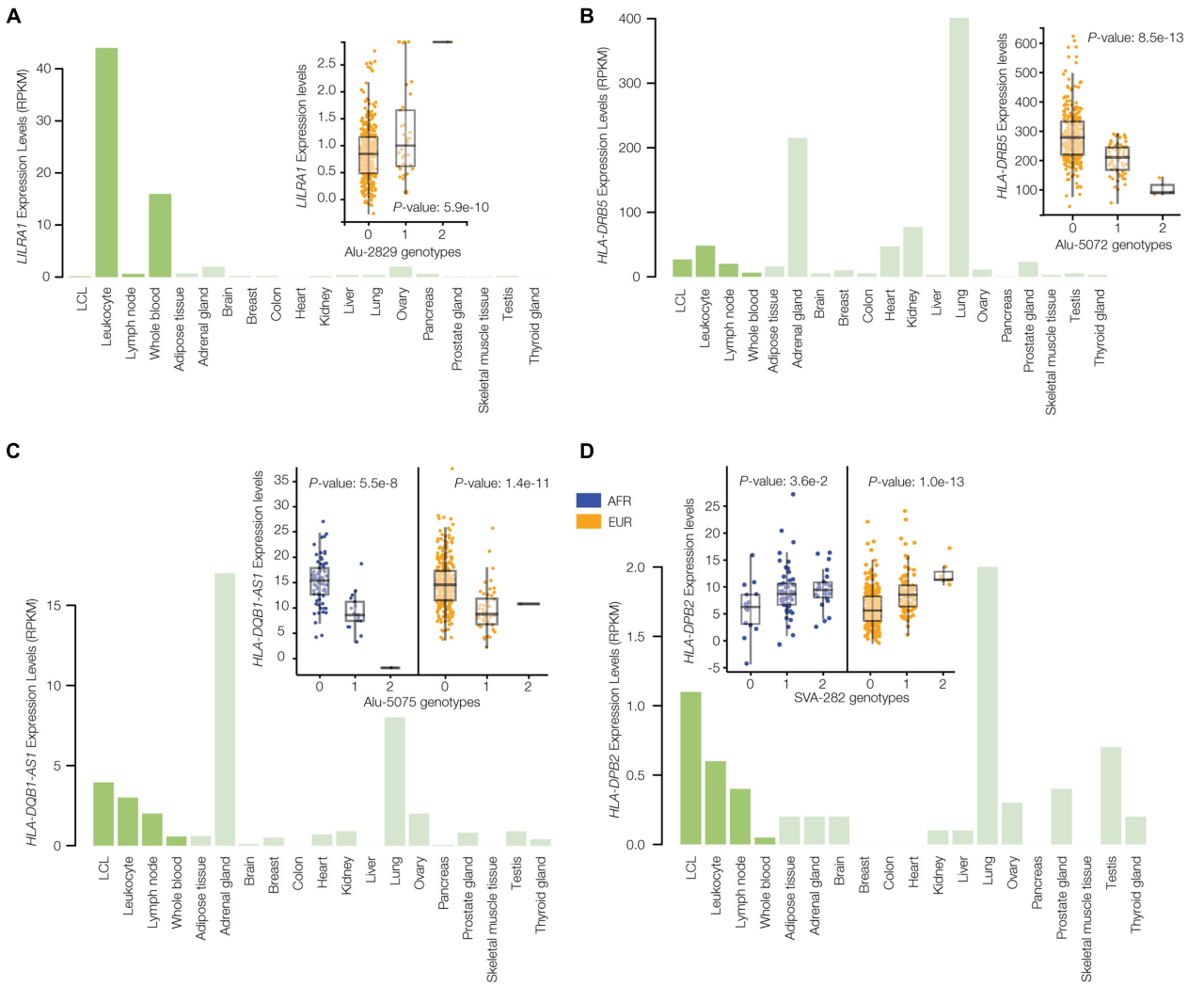


FIGURE 6 | Gene expression profiles and eQTL results for disease-associated polyTE insertions. Bar-charts of tissue-specific expression levels and box-pots of eQTL analyses are shown for four examples of TE-eQTLs corresponding to disease-linked and enhancer co-located polyTE insertions that regulate immune-related genes: **(A)** Alu-2829 and *LILRA1*, **(B)** Alu-5072 and *HLA-DRB5*, **(C)** Alu-5075 and *HLA-DQB1-AS1*, and **(D)** SVA-282 and *HLA-DPB2*. Bar-charts show tissue-specific expression levels as normalized RPKM values (green). The inset eQTL box-plots show individual sample gene expression levels (*y*-axis) regressed against polyTE insertion presence/absence genotypes (*x*-axis): 0-homozygous absent, 1-heterozygous, 2-homozygous present. Each dot represents a single individual from the African (AFR – blue) and/or European (EUR – orange) population groups.

regions that encode immune-related genes (**Figure 5B**). We narrowed this list further by selecting the strongest TE-eQTL association for each individual polyTE, resulting in a final list of 37 (AFR) and 24 (EUR) immune-related TE-eQTL (**Figure 2**).

The results of the TE-eQTL analysis further underscore the regulatory potential of the disease-linked polyTEs characterized here and also allowed us to narrow down the list of candidate insertions. Starting with the list of TE-eQTL, we searched for 'consistent' examples where the disease-linked polyTE is associated with the expression of a gene that is functionally related to the annotated disease phenotype. This allowed us to converge on a final set of seven high-confidence disease-associated TE insertion polymorphisms (**Figure 2** and **Table 1**). Four of these disease-associated polyTEs are illustrated in **Figure 6**, and we provide additional information on two examples in the following section "Effects of polyTE Insertions on Immune- and Blood-Related Conditions." The four examples shown in **Figure 6** all correspond to polyTEs that are linked to disease-associated SNPs and co-located with enhancers characterized from blood- or immune system-related tissues; in addition, the genes that these polyTE insertions regulate are all known to function in the immune system.

Six of the seven disease-associated polyTE insertions are considered to be population-specific, based on significant eQTL results in only one population, whereas a single case is shared between both the AFR and EUR population groups (**Figure 6C**).

However, two of the six cases considered to be population-specific using the eQTL criterion do show consistent trends across populations but failed to reach genome-wide significance when controls for multiple statistical tests were implemented (**Figures 6D**, **7B**).

Effects of PolyTE Insertions on Immune- and Blood-Related Conditions

Here, we described two specific examples of the effects that polyTE insertions can exert on immune- and blood-related disease phenotypes. **Figure 7A** shows the SVA-401 insertion that is co-located with a cell-type specific enhancer found in the second intron of the Beta-1,4-Galactosyltransferase 1 (*B4GALT1*) encoding gene, which is normally expressed at high levels in immune-related tissues. Chromatin interaction maps characterized for several different cell types – CD34, GM12878, and Mcf7 – show that this *B4GALT1* intronic enhancer physically associates with the gene's promoter region. The disruption of the *B4GALT1* enhancer by the SVA insertion is associated with down-regulation of the gene in B-cells, for both AFR and EUR population groups (**Figure 7B**). *B4GALT1* encodes a glycosyltransferase that functions in the glycosylation of the Immunoglobulin G (IgG) antibody in such a way as to convert its activity from pro- to anti-inflammatory (**Figure 7C**) (Kaneko et al., 2006; Anthony and Ravetch, 2010; Lauc et al., 2013). Down-regulation of this gene in individuals with the enhancer SVA insertion should thereby serve to keep the IgG molecule in a pro-inflammatory state. Consistent with this idea, the *B4GALT1* enhancer SVA insertion is linked to a genomic region implicated by GWAS in both inflammatory conditions and autoimmune diseases such as systemic lupus erythematosus and Crohn's disease (Lauc et al., 2013).

Another example of an SVA insertion into an enhancer element is shown for the adjacently located Signal Transducing Adaptor Molecule (*STAM*) and Transmembrane Protein 236 (*TMEM236*) encoding genes. The SVA-438 insertion is co-located with an enhancer in the first intron of the *STAM* gene (**Figure 7D**), but its presence is associated with changes in expression of the nearby *TMEM236* gene (**Figure 7E**). *TMEM236* is located ∼100 kbp downstream of the SVA-438 insertion and is most highly expressed in pancreatic islet α-cells (**Figure 7F**) (Ackermann et al., 2016; Wang Y.J. et al., 2016). Islet α-cells function to secrete glucagon, a peptide hormone that elevates glucose levels in the blood (Quesada et al., 2008). The SVA-438 insertion is associated with increased expression of *TMEM236*, which would be expected to lead to increased blood glucose levels. This expectation is consistent with the fact that the SVA-438 insertion is also linked to the risk allele (T) of the SNP rs6602203, which is associated with a reduced metabolic clearance rate of insulin (MCRI), an endophenotype that is associated with the risk of type 2 diabetes (Palmer et al., 2015). In other words, up-regulation of *TMEM236* by the SVA-438 insertion may be mechanistically linked to insulin resistance by virtue of increasing blood sugar and decreasing insulin clearance.

DISCUSSION

The results reported here underscore the influence that retrotransposon insertion polymorphisms can exert on human health- and disease-related phenotypes. The integrative data analysis approach that we took for this study also revealed how polyTE disease-associations are mediated by the gene regulatory properties of retrotransposon insertions. We adopted a conservative approach to screen for the potential regulatory effects of retrotransposon insertions by choosing candidate elements as those that were inserted into regions previously defined as tissue-specific enhancers in blood/immune cells. Retrotransposons that insert into enhancer sequences could entail loss-of-function mutants by virtue of disrupting enhancer sequences, or they could serve as gain-of-function mutants by altering enhancer activity. Our results can be considered to show instances of both loss- and gain-of-function enhancer mutations with respect to the decrease or increase, respectively, of gene expression levels that are associated with element insertion genotypes (**Figures 6**, **7**). Nevertheless, it is worth noting that our conservative approach could be prone to false negatives as it would not uncover novel enhancer activity provided by element insertions at new locations in the genome.

The TE regulatory findings that we report here are consistent with previous studies showing that TE-derived sequences have contributed a wide variety of gene regulatory elements to the human genome (Feschotte, 2008; Rebollo et al., 2012; Chuong et al., 2017), including promoters (Jordan et al., 2003; Marino-Ramirez et al., 2005; Conley et al., 2008), enhancers (Bejerano et al., 2006; Kunarso et al., 2010; Chuong et al., 2013, 2016; Notwell et al., 2015), transcription terminators (Conley and Jordan, 2012) and several classes of small RNAs (Weber, 2006; Piriyapongsa et al., 2007; Kapusta et al., 2013). Human TEs can also influence gene regulation by modulating various aspects of chromatin structure throughout the genome (Lander et al., 2001; Pavlicek et al., 2001; Schmidt et al., 2012; Jacques et al., 2013; Sundaram et al., 2014; Wang et al., 2015).

It is important to note that the research efforts which have uncovered the regulatory properties of human TEs, including a number of our own studies, have dealt exclusively with sequences derived from relatively ancient insertion events. These ancient TE insertions are present at the same (fixed) locations in the genome sequences of all human individuals. In other words, previously described TE-derived regulatory sequences are uniformly present among individual human genomes and thereby do not represent a source of structural genetic variation. Such fixed TE-derived regulatory sequences may not be expected to provide for gene regulatory variation among individuals or for that matter to contribute to inter-individual differences related to health and disease.

Nevertheless, we recently showed that TE insertion polymorphisms also exert regulatory effects on the human genome (Wang L. et al., 2016). Specifically, polyTE insertions were shown to contribute to both inter-individual and population-specific differences in gene expression and to facilitate the re-wiring of transcriptional networks. The results

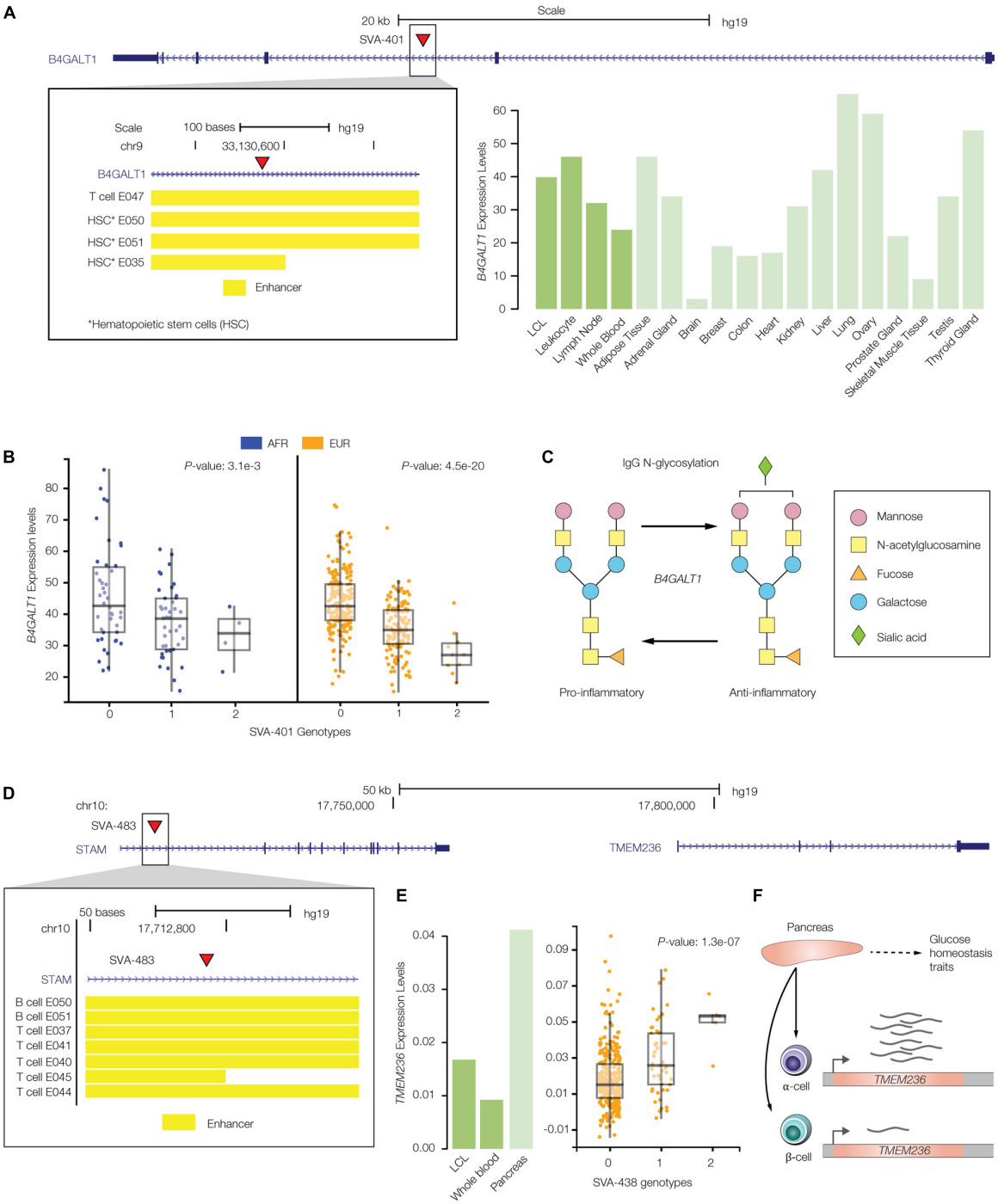


FIGURE 7 | PolyTE insertions associated with immune- and blood-related conditions. **(A)** UCSC Genome Browser screen capture showing the location of the SVA-401 insertion (red arrow) on chromosome 19 within the second exon of the *B4GALT1* gene. The inset shows the genomic locations of co-located enhancers, characterized based on chromatin signatures from a variety of tissue-specific epigenomes locations, as yellow bars. The bar-chart shows *B4GALT1* tissue-specific expression levels as normalized RPKM values (green). **(B)** eQTL box-plots show individual sample gene expression levels (*y*-axis) regressed against SVA-401 insertion presence/absence genotypes (*x*-axis): 0-homozygous absent, 1-heterozygous, 2-homozygous present. Each dot represents a single individual from the African (AFR – blue) or European (EUR – orange) population groups. **(C)** *B4GALT1* catalyzed glycosylation of the Immunoglobulin G (IgG) antibody, resulting in conversion from pro- to anti-inflammatory activity. **(D)** UCSC Genome Browser screen capture showing the location of the SVA-438 insertion (red arrow) on chromosome 10 within the first exon of the *STAM* gene, upstream of the regulated *TMEM236* gene. The inset shows the genomic locations of co-located enhancers (yellow bars). **(E)** Bar-chart of *TMEM236* tissue-specific expression levels and box-pot of the SVA-438 *TMEM236* eQTL analyses. **(F)** Functional role and cell-type specific expression profile for *TMEM236*.

reported here extend those findings up the hierarchy of human biological organization by revealing potential mechanistic links between polyTE-induced gene regulatory changes and the endophenotypes that underlie human health and disease.

AUTHOR CONTRIBUTIONS

LW performed all of the analyses described in the study. EN contributed all GWAS data used in the study. IJ conceived of, designed and supervised the study. All authors contributed to the drafting and revision of the manuscript.

FUNDING

LW was supported by the Georgia Tech Bioinformatics Graduate Program. EN and IJ were supported by the IHRC-Georgia Tech Applied Bioinformatics Laboratory (ABiL).

ACKNOWLEDGMENT

The authors would like to thank Jianrong Wang for his advice on the analysis of tissue-specific enhancer elements.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01418/full#supplementary-material

TABLE S1 | Linkage disequilibrium (LD) structure between polyTEs and GWAS disease-associated SNPs.

TABLE S2 | Genomic co-localization of polyTEs with tissue-specific enhancers.

TABLE S3 | Expression quantitative trait loci (eQTL) results for disease-linked and enhancer co-located polyTEs.

REFERENCES

Ackermann, A. M., Wang, Z., Schug, J., Naji, A., and Kaestner, K. H. (2016). Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* 5, 233–244. doi: 10.1016/j.molmet.2016.01.002

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Anthony, R. M., and Ravetch, J. V. (2010). A novel role for the IgG Fc glycan: the anti-inflammatory activity of sialylated IgG Fcs. *J. Clin. Immunol.* 30(Suppl. 1), S9–S14. doi: 10.1007/s10875-010-9405-6

Batzer, M. A., and Deininger, P. L. (1991). A human-specific subfamily of Alu sequences. *Genomics* 9, 481–487. doi: 10.1016/0888-7543(91)90414-A

Batzer, M. A., Gudi, V. A., Mena, J. C., Foltz, D. W., Herrera, R. J., and Deininger, P. L. (1991). Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res.* 19, 3619–3623. doi: 10.1093/nar/19.13.3619

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., et al. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441, 87–90. doi: 10.1038/nature04696

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., et al. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5280–5285. doi: 10.1073/pnas.0831042100

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351, 1083–1087. doi: 10.1126/science.aad5497

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18, 71–86. doi: 10.1038/nrg.2016.139

Chuong, E. B., Rumi, M. A., Soares, M. J., and Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.* 45, 325–329. doi: 10.1038/ng.2553

Clayton, E. A., Wang, L., Rishishwar, L., Wang, J., McDonald, J. F., and Jordan, I. K. (2016). Patterns of transposable element expression and insertion in cancer. *Front. Mol. Biosci.* 3:76. doi: 10.3389/fmolb.2016.00076

Conley, A. B., and Jordan, I. K. (2012). Cell type-specific termination of transcription by transposable element sequences. *Mob. DNA* 3:15. doi: 10.1186/1759-8753-3-15

Conley, A. B., Piriyapongsa, J., and Jordan, I. K. (2008). Retroviral promoters in the human genome. *Bioinformatics* 24, 1563–1567. doi: 10.1093/bioinformatics/btn243

de Koning, A. P. J., Gu, W. J., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7:e1002384. doi: 10.1371/journal.pgen.1002384

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405. doi: 10.1038/nrg2337

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., et al. (2013). Ensembl 2013. *Nucleic Acids Res.* 41, D48–D55. doi: 10.1093/nar/gks1236

Hancks, D. C., and Kazazian, H. H. Jr. (2016). Roles for retrotransposon insertions in human disease. *Mob. DNA* 7:9. doi: 10.1186/s13100-016-0065-9

Hancks, D. C., and Kazazian, H. H. (2012). Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* 22, 191–203. doi: 10.1016/j.gde.2012.02.006

Jacques, P. E., Jeyakani, J., and Bourque, G. (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* 9:e1003504. doi: 10.1371/journal.pgen.1003504

Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72. doi: 10.1016/S0168-9525(02)00006-9

Kaneko, Y., Nimmerjahn, F., and Ravetch, J. V. (2006). Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation. *Science* 313, 670–673. doi: 10.1126/science.1129594

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X. Y., Ramsay, L., Bourque, G., et al. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* 9:e1003470. doi: 10.1371/journal.pgen.1003470

Kazazian, H. H. Jr., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., and Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* 332, 164–166. doi: 10.1038/332164a0

Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., et al. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* 42, 631–634. doi: 10.1038/ng.600

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Lauc, G., Huffman, J. E., Pucic, M., Zgaga, L., Adamczyk, B., Muzinic, A., et al. (2013). Loci associated with N-glycosylation of human immunoglobulin G

show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet.* 9:e1003225. doi: 10.1371/journal.pgen.1003225

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Marino-Ramirez, L., Lewis, K. C., Landsman, D., and Jordan, I. K. (2005). Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* 110, 333–341. doi: 10.1159/000084965

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Notwell, J. H., Chung, T., Heavner, W., and Bejerano, G. (2015). A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat. Commun.* 6:6644. doi: 10.1038/ncomms7644

Ostertag, E. M., Goodier, J. L., Zhang, Y., and Kazazian, H. H. Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* 73, 1444–1451. doi: 10.1086/380207

Palmer, N. D., Goodarzi, M. O., Langefeld, C. D., Wang, N., Guo, X., Taylor, K. D., et al. (2015). Genetic variants associated with quantitative glucose homeostasis traits translate to type 2 diabetes in Mexican Americans: the GUARDIAN (genetics underlying diabetes in Hispanics) consortium. *Diabetes* 64, 1853–1866. doi: 10.2337/db14-0732

Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J. V., and Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 276, 39–45. doi: 10.1016/S0378-1119(01)00645-X

Piriyapongsa, J., Marino-Ramirez, L., and Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics* 176, 1323–1337. doi: 10.1534/genetics.107.072553

Quesada, I., Tuduri, E., Ripoll, C., and Nadal, A. (2008). Physiology of the pancreatic alpha-cell and glucagon secretion: role in glucose homeostasis and diabetes. *J. Endocrinol.* 199, 5–19. doi: 10.1677/JOE-08-0290

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* 46, 21–42. doi: 10.1146/annurev-genet-110711-155621

Rishishwar, L., Marino-Ramirez, L., and Jordan, I. K. (2016). Benchmarking computational tools for polymorphic transposable element detection. *Brief. Bioinform.* doi: 10.1093/bib/bbw072 [Epub ahead of print].

Rishishwar, L., Tellez Villa, C. E., and Jordan, I. K. (2015). Transposable element polymorphisms recapitulate human evolution. *Mob. DNA* 6:21. doi: 10.1186/s13100-015-0052-6

Rishishwar, L., Wang, L., Clayton, E. A., Marino-Ramirez, L., McDonald, J. F., and Jordan, I. K. (2017). Population and clinical genetics of human transposable elements in the (post) genomic era. *Mob. Genet. Elements* 7, 1–20. doi: 10.1080/2159256X.2017.1280116

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., et al. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348, 666–669. doi: 10.1126/science.1261877

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., et al. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148, 335–348. doi: 10.1016/j.cell.2011.11.058

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358. doi: 10.1093/bioinformatics/bts163

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7, 500–507. doi: 10.1038/nprot.2011.457

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., et al. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* 24, 1963–1976. doi: 10.1101/gr.168872.113

t Hoen, P. A., Friedlander, M. R., Almlof, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* 31, 1015–1022. doi: 10.1038/nbt.2702

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A., et al. (2005). SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* 354, 994–1007. doi: 10.1016/j.jmb.2005.09.085

Wang, J., Vicente-Garcia, C., Seruggia, D., Molto, E., Fernandez-Minan, A., Neto, A., et al. (2015). MIR retrotransposon sequences provide insulators to the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 112, E4428–E4437. doi: 10.1073/pnas.1507253112

Wang, L., Rishishwar, L., Marino-Ramirez, L., and Jordan, I. K. (2016). Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res.* 45, 2318–2328. doi: 10.1093/nar/gkw1286

Wang, Y. J., Schug, J., Won, K. J., Liu, C., Naji, A., Avrahami, D., et al. (2016). Single-Cell transcriptomics of the human endocrine pancreas. *Diabetes* 65, 3028–3038. doi: 10.2337/db16-0405

Weber, M. J. (2006). Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.* 2:e205. doi: 10.1371/journal.pgen.0020205

Wildschutte, J. H., Williams, Z. H., Montesion, M., Subramanian, R. P., Kidd, J. M., and Coffin, J. M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl. Acad. Sci. U.S.A.* 113, E2326–E2334. doi: 10.1073/pnas.1602336113

Xie, X., Ma, W., Songyang, Z., Luo, Z., Huang, J., Dai, Z., et al. (2016). CCSI: a database providing chromatin-chromatin spatial interaction information. *Database* 2016:bav124. doi: 10.1093/database/bav124

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 14 September 2017
doi: 10.3389/fmicb.2017.01745

RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense

Karin Moelling[1,2†], Felix Broecker[3†], Giancarlo Russo[4] and Shinichi Sunagawa[5]*

[1] Institute of Medical Microbiology, University of Zurich, Zurich, Switzerland, [2] Max Planck Institute for Molecular Genetics, Berlin, Germany, [3] Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [4] Functional Genomics Center Zurich, ETH Zurich/University of Zurich, Zurich, Switzerland, [5] Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland

Retroviral infections are 'mini-symbiotic' events supplying recipient cells with sequences for viral replication, including the reverse transcriptase (RT) and ribonuclease H (RNase H). These proteins and other viral or cellular sequences can provide novel cellular functions including immune defense mechanisms. Their high error rate renders RT-RNases H drivers of evolutionary innovation. Integrated retroviruses and the related transposable elements (TEs) have existed for at least 150 million years, constitute up to 80% of eukaryotic genomes and are also present in prokaryotes. Endogenous retroviruses regulate host genes, have provided novel genes including the syncytins that mediate maternal-fetal immune tolerance and can be experimentally rendered infectious again. The RT and the RNase H are among the most ancient and abundant protein folds. RNases H may have evolved from ribozymes, related to viroids, early in the RNA world, forming ribosomes, RNA replicases and polymerases. Basic RNA-binding peptides enhance ribozyme catalysis. RT and ribozymes or RNases H are present today in bacterial group II introns, the precedents of TEs. Thousands of unique RTs and RNases H are present in eukaryotes, bacteria, and viruses. These enzymes mediate viral and cellular replication and antiviral defense in eukaryotes and prokaryotes, splicing, R-loop resolvation, DNA repair. RNase H-like activities are also required for the activity of small regulatory RNAs. The retroviral replication components share striking similarities with the RNA-induced silencing complex (RISC), the prokaryotic CRISPR-Cas machinery, eukaryotic V(D)J recombination and interferon systems. Viruses supply antiviral defense tools to cellular organisms. TEs are the evolutionary origin of siRNA and miRNA genes that, through RISC, counteract detrimental activities of TEs and chromosomal instability. Moreover, piRNAs, implicated in transgenerational inheritance, suppress TEs in germ cells. Thus, virtually all known immune defense mechanisms against viruses, phages, TEs, and extracellular pathogens require RNase H-like enzymes. Analogous to the prokaryotic CRISPR-Cas anti-phage defense possibly originating from TEs termed casposons, endogenized retroviruses ERVs and amplified TEs can be regarded as related forms of inheritable immunity in eukaryotes. This survey suggests that RNase H-like activities of retroviruses, TEs, and phages, have built up innate and adaptive immune systems throughout all domains of life.

Keywords: RNase H, reverse transcriptase, retroviruses, (Retro)-transposons, ribozymes, evolution, antiviral defense, immune systems

OPEN ACCESS

Edited by:
Guenther Witzany, Telos-Philosophische Praxis, Austria

Reviewed by:
Yize Li, University of Pennsylvania, United States
Junji Xing, Houston Methodist Research Institute, United States

***Correspondence:**
Karin Moelling
moelling@molgen.mpg.de

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *26 July 2017*
Accepted: *28 August 2017*
Published: *14 September 2017*

Citation:
Moelling K, Broecker F, Russo G and Sunagawa S (2017) RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. Front. Microbiol. 8:1745. doi: 10.3389/fmicb.2017.01745

RT AND RNase H OF RETROVIRUSES AND RETROVIRUS-LIKE ELEMENTS

The discovery of the reverse transcriptase (RT), initially described as replication enzyme of retroviruses in 1970 (Baltimore, 1970; Temin and Mizutani, 1970), was so unexpected that it was awarded a Nobel prize in 1975. Shortly after the RT, the retroviral ribonuclease H (RNase H) was identified in retrovirus particles as essential component for the replication of viral RNA via an RNA-DNA hybrid intermediate to double stranded DNA (dsDNA) (Mölling et al., 1971; Hansen et al., 1988; Tisdale et al., 1991). Historically, the RNase H has often been considered as part of the RT in retroviruses. However, the enzyme has its proper role, impact on evolution and importance for the degradation of nucleic acids in various biological processes. Similarly, the RT is of much more general importance than just replicating retroviral genomes. One prominent example is the telomerase, the RT or TERT, that elongates chromosomal ends in embryonic tissue and stem cells. Here, RNase H activity is not involved, since the template RNA needs to be copied repeatedly. Both the RT and the RNase H are among the most abundant proteins on our planet (Ma et al., 2008; Caetano-Anollés et al., 2009; Majorek et al., 2014).

RNases H are also present in pararetroviruses such as hepatitis B viruses, cauliflower mosaic viruses that infect plants, the monkey-specific spuma or foamy viruses, or the sheep lentivirus Visna Maedi Virus. They all require an RNase H to cooperate with the RT. Pararetroviruses follow the same life cycle as retroviruses except that the replication intermediates packaged into the virion are at a different stage so that they contain dsDNA, not ssRNA (Flint et al., 2015).

Contrary to Francis Crick's 'central dogma' from 1958 the RT, in concert with the RNase H, allows for the flow of information to occur from RNA to DNA. This was regarded as the reverse orientation, a historical view rooted in the discovery that DNA was the carrier of genetic information (**Figure 1A**). Reverse transcription of RNA occurring in concert with an RNase H is not restricted to retroviruses and pararetroviruses but is important also for retrotransposons that amplify via a 'copy-and-paste' mechanism such as the Ty elements of yeast. This mode of replication is closely related to that of retroviruses. However, there is no particle formation, release, and exogenous infection, since an envelope gene is missing. Genetic information is retrotransposed and can amplify within genomes.

An appreciation of the biological importance of RNA is increasing. RTs, as well as RNases H can exert crucial roles in the biogenesis and degradation of RNA molecules independently of each other. RNases H are involved whenever processing, trimming, or removal of genetic information is required – which happens as frequently as does the *de novo* synthesis of nucleic acid polymers. In theory, synthesis and degradation of nucleic acids should be in a well-balanced equilibrium. The RNase H-like structure is involved in numerous cleavage enzymes such as the retroviral integrase. The retroviral life cycle requires an integrase, which allows for inserting the DNA provirus into the cellular genome. Integrases adopt an RNase H-like core structure. Similarly, the 'cut-and-paste' replicative mechanism of transposable elements (TEs) also requires an integrase-like enzyme termed transposase (similarly with an RNase H fold), independently of an RT. The RT itself can also act independently of an RNase H, as in the case of telomerase, the enzyme that extends the ends of chromosomes. Telomerase depends on a short RNA molecule that is copied repeatedly – template degradation by an RNase H must not occur. In contrast, DNA-dependent DNA polymerases require RNases H for the removal of RNA primers after they have served their function, whereby the RNase H, in this case, is not fused to the polymerase as in retroviruses but is a separate molecule.

It came as a surprise when sequencing of the human genome revealed that almost 50% of its sequence is composed of retrovirus-like elements such as long and short interspersed nuclear elements (LINEs and SINEs), endogenous retroviruses (ERVs) often shortened to solitary LTRs, and Alu elements (a subclass of SINEs) that are common source of mutation in humans (Lander et al., 2001). Human ERVs (HERVs) populate the human genome and result from former germ line cell infections up to 150 Mio or more years ago.

The RNase H was first discovered in lysates of calf thymus, with unknown functions for a long time (Stein and Hausen, 1969). RNase H activity was also early described in the yeast *Saccharomyces cerevisiae*, where it was found to be involved in DNA replication (Karwan and Wintersberger, 1986). DNA replication requires RNA primers to initiate lagging strand DNA synthesis and their subsequent removal by the RNase H. The abundance of RT enzymes in bacteria, with 1021 different types identified, is surprising, yet their functions remain poorly understood (Simon and Zimmerly, 2008).

RNase H IN VARIOUS SPECIES

RNases H are essential for degrading the RNA template during reverse transcription of the retroviral genome, thereby generating short RNA primers that initiate DNA synthesis (Flint et al., 2015). Primers are subsequently removed by the RNase H. In the 1980s/1990s, three prokaryotic RNases H, RNase HI, HII and HIII with roman numbers, and two eukaryotic enzymes, RNase H1 and H2 with arabic numbers, have been characterized and classified based on differences in amino acid sequences. RNase HI/H1 and retroviral RNase H are classified as type 1, whereas RNase HII/H2 are type 2 RNases H (Cerritelli and Crouch, 2009; Tadokoro and Kanaya, 2009). Furthermore, there are two yeast enzymes RNase H1 and RNase H2 that resemble the mammalian and prokaryotic enzymes (Karwan and Wintersberger, 1986). RNase H enzymes are also found in archaea (Ohtani et al., 2004). Cellular RNase H enzymes share the common activity of degrading the RNA moiety of RNA-DNA hybrids necessary, for instance, to remove RNA primers during DNA replication. Mammalian RNase H2 is composed of three subunits A, B, and C that form a functional complex with only the A subunit exhibiting enzymatic activity and the other two supplying scaffold and

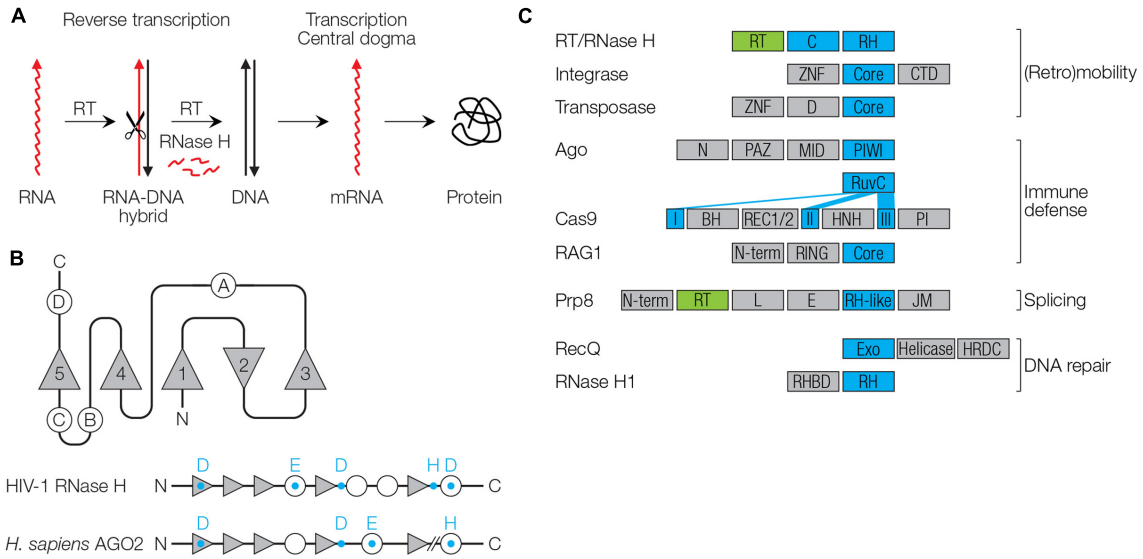


FIGURE 1 | Function and structure of RNase H-like proteins. **(A)** Reverse transcriptase (RT), in concert with the RNase H catalyzes retroviral replication. **(B)** Conserved RNase H fold. β-sheets are shown as triangles (numbered 1–5), α-helices as circles (numbered A–D). Conserved catalytic residues are shown as blue letters. **(C)** Domain architectures of proteins with RNase H-like (RH) domains (blue) and RT domains (green). RT/RNase H (HIV-1): C, connection/tether/linker domain (Artymiuk et al., 1993). Integrase (HIV-1): ZNF, zinc finger domain; Core, catalytic core; CTD, C-terminal domain with DNA binding activity (Hindmarsh and Leis, 1999). Transposase, D, dimerization domain (Jiang et al., 2016). Argonaute (*Homo sapiens* AGO2): N, N-terminal domain; PAZ, PIWI/Argonaute/Zwille domain recognizing the 3′ end of small RNAs; MID, middle domain recognizing the 5′ end of small RNAs; PIWI, RNase H-like domain (Song et al., 2004). Cas9 (*Streptococcus pyogenes*): The RuvC domain with RNase H fold is separated into three parts (I–III); BH, bridge helix domain; REC1/2, recognition domains (REC1 binds the guide RNA); HNH, nuclease domain; PI, PAM-interacting domain (Jinek et al., 2014; Nishimasu et al., 2014). Rag-1 (*H. sapiens* RAG1): N-term, N-terminal domain with ubiquitin ligase activity; RING, Really Interesting New Gene domain with zinc finger motif; Core, catalytic core domain with endonuclease that also contains a zinc finger motif (Bassing et al., 2002; Yurchenko et al., 2003). Prp8 (*Saccharomyces cerevisiae*): N-term, N-terminal; L, linker; E, endonuclease-like; JM, Jab1/MPN (Nguyen et al., 2015). RecQ (*H. sapiens,* Werner Syndrome): Exo, 3′ to 5′ exonuclease domain; helicase and HRDC, RNase D C-terminal domain involved in DNA and protein interactions; NLS, nuclear localization signal (Oshima, 2000). RNase H1 (*H. sapiens*): RHBD, RNA/DNA hybrid binding domain (Nowotny et al., 2005).

structural functions and mediating protein–protein interactions as well as target specificity (Crow et al., 2006). This enzyme excises ribonucleotides misincorporated in DNA molecules that can otherwise induce DNA instability and mutations. Cellular RNases H are essential in mammals and higher eukaryotes but not in lower eukaryotes and prokaryotes.

Mice deficient in RNase H1 that localizes to mitochondria die during embryogenesis, probably due to the defective processing of R-loops (Cerritelli and Crouch, 2009). R-loops are formed when an RNA strand intercalates into dsDNA, resulting in RNA-DNA hybrids and single-stranded DNA loops. R-loops affect promoter activities, with a role in gene expression (e.g., of the c-Myc proto-oncogene), genome stability, CRISPR-Cas immunity, DNA repair, and cancer formation (Sollier and Cimprich, 2015). RNases H can remove the RNA moiety and prevent deleterious DNA breaks. RNase H2 knockout mice are also not viable, and mutations in either of the human genes can cause Aicardi-Goutières Syndrome, a severe inheritable neurodevelopmental disorder (Crow et al., 2006). In this disease, uncleaved RNA-DNA hybrids accumulate within cells that possibly upregulate interferon via the nucleic acid sensor cyclic GMP-AMP synthase (cGAS) and its adaptor protein STING (Mackenzie et al., 2016). Are there retrotransposon involved? RT inhibitors are under investigation and will show!

Synthetic DNA oligonucleotides can form local hybrids and direct the RNase H to cleave the RNA moiety, which was defined as silencing by siDNA in analogy to interference by siRNA (Moelling et al., 2006; Swarts et al., 2014).

RNase H STRUCTURE, FUNCTION AND SPECIFICITY

RNase HI of *Escherichia coli* was the first one of which the three-dimensional structure was solved, revealing a conserved protein architecture, the RNase H fold (Katayanagi et al., 1990; Yang et al., 1990). RNase H folds occur in a diverse number of enzymes involved in replication, recombination, DNA repair, splicing, (retro)transposition of TEs, RNA interference (RNAi) and CRISPR-Cas immunity. Enzymes with an RNase H fold have been designated as RNase H-like superfamily (Majorek et al., 2014). RNase H folds usually contain five β-sheets (numbered 1–5) with the second being antiparallel to the other four (Ma et al., 2008) (**Figure 1B**). The catalytic core is flanked by a varying number of α-helices (A, B, C, etc.; a total of four in HIV-1 RNase H). Three to four acidic amino acid residues (aspartic acid D, or glutamic acid E) that coordinate divalent cations are required for catalysis (D443 E478 D498 D549 in HIV-1). The DEDD residues are highly conserved among type I RNase H

proteins, whereas an additional C-terminal histidine (H) that is conserved in fungal, metazoan and retroviral RNases H, is replaced with an arginine (R) in archaea (Ustyantsev et al., 2015). The conserved amino acids in *E. coli* RNase H were the basis for loss-of-function mutagenesis of the HIV-1 RNase H to validate the necessary role of this enzyme for viral replication and as a drug target (Tisdale et al., 1991). RNases H act as dimers, with two Mg^{2+} or other divalent cations being essential for correct protein structure, stability and enzyme activity. Replacement of Mg^{2+} by Mn^{2+} or Co^{2+} is preferred in certain species, while Ca^{2+} ions generally inactivate catalysis (Ohtani et al., 2004).

RNase H-like proteins are directed to their nucleic acid substrates via additional factors. The best-studied example is fusion to the RT domain in retroviruses. The fusion leads to a distance of 18 nucleotides between the two active centers. An unusual structural bend is located within the extended polypurine tract (PPT) that comprises 18 nucleotides, interrupted by a CU dinucleotide. Cleavage by the RNase H domain occurs exactly at this site, generating a truncated PPT RNA primer for second strand DNA synthesis. A high degree of precision is required for this cut, leaving a terminal dinucleotide that is essential for integration and LTR formation (Sarafianos et al., 2001; Moelling, 2012).

In the case of HIV, the RNase H appears to be a processive exonuclease due to its fusion to the RT domain that pulls the RNase H during DNA synthesis. Thereby, processivity of the RNase H is pretended, even though this is an effect mediated by the RT. RNases H do not normally release mononucleotides as exonucleases would do. The catalytic activity of HIV RNase H is lower than that of the RT, and studdering of the RT enhances cleavage by the RNase H (Moelling, 2012). Exonuclease activity is controversial; however, it was recently described that RNase H-like enzymes can act as exonucleases depending on the orientation of a C-terminal alpha helix (Majorek et al., 2014). Furthermore, some RNases H can also cleave dsRNA, an activity that involves aspartic acid residues not required for cleavage of hybrid nucleic acids, as described for archaea (Ohtani et al., 2004). The cleavage of the phosphodiester bond is somewhat unusual, leading to 3′-OH, which can again be used as start for DNA synthesis, and 5′-phosphates.

RNase H FAMILY MEMBERS

RNases H mainly recognize structure, not sequence, and cleave one strand of double-stranded nucleic acid molecules such as RNA in DNA-RNA hybrids (RNases H), DNA in hybrids (Cas9), dsRNA (Ago/PIWI proteins of pro- and eukaryotes), and dsDNA (integrases or transposases). The importance of nucleic acid structure may be a remnant of a function in the ancient RNA world, where RNA structure was an important feature. Specificities of RNase H-like enzymes are governed by structural properties of the nucleic acid substrate, fused protein domains, protein–protein interacting factors, ion cofactors, and guide nucleic acids to find substrates in a sequence-specific manner. This, the RNases H are team players, which is supported by partnering or fused proteins that enable specific functions (**Figure 1C** and **Table 1**).

Retroviral RNase H is discussed above. The retroviral integrase has an N-terminal zinc finger (ZNF) and a C-terminal domain (CTD), both of which mediate DNA binding to guide the catalytic core to the dsDNA substrate (Hindmarsh and Leis, 1999).

Transposases of TEs, as well as the mammalian recombination-activating gene 1 (RAG1) enzyme, also contain ZNF domains that govern dsDNA binding (Yurchenko et al., 2003). ZNF domains with desired sequence specificities can be artificially fused to an RNase H enzyme to achieve cleavage at specific target DNA sequences with potential use in gene therapy (Sulej et al., 2012). Transposases, the most abundant RNase H-like proteins (Majorek et al., 2014), mediate cut-and-paste of DNA transposons, a phenomenon originally described by McClintock (1951). Transposases cleave DNA to excise and reinsert the transposon into the host chromosome. Examples include transposases of prokaryotic transposons Tn3, Tn5 and Mu and those of the mariner/Tc3, sleeping beauty and hAT (Hobo/Activator/Tam3) families of eukaryotic transposons. Their core domains adopt an RNase H fold with catalytic DD[E/D] motif.

Terminases of phages contain an RNase H domain that cleaves dsDNA concatemers of phages into single genomes and are crucial for packaging into the virion and assembly (Feiss and Rao, 2012).

An RNase H-like fold is found in the Cas9 protein in the CRISPR-Cas9 defense system (discussed in detail below).

Argonaute (Ago) proteins mediate silencing of foreign RNAs or mRNAs via small interfering RNA (siRNA) or micro RNA (miRNA) within the RNA-induced silencing complex (RISC). Ago contain PAZ and PIWI domains that are fused and are related to the retroviral RT-RNase H fusion protein including the linker region (Song et al., 2004; Moelling et al., 2006). It is surprising, that the viral "tool kits" are similar to those of antiviral defense. Both proteins resemble each other in structure, domain organization, RNA binding and cleavage properties. The catalytic PIWI domain requires a guide RNA to localize its target nucleic acid for sequence-specific cleavage. This guide RNA is supplied by the action of Dicer, the enzyme that trims dsRNA molecules to small RNAs. Then PAZ binds the 3′-hydroxyl end of the guide RNA and directs it to the target strand (Song et al., 2004; Jinek et al., 2014). The PAZ and PIWI domains ensure correct positioning of the small RNAs to the complementary target RNA for cleavage. Ago binds to guide RNA via its PAZ pocket structural motif, originally described as primer grip for the retroviral RT-RNase H that mediates binding to the DNA opposite of the scissile RNA phosphodiester (Song et al., 2004). Depending on whether the original RNA is a double strand or hairpin-looped, the silencing is designated as siRNA using Dicer or miRNA using Drosha to process the primary pri-miRNA transcript that is further cleaved by Dicer.

Cas9 mediates the best described prokaryotic CRISPR-Cas defense system against phages and plasmids. The enzyme contains two nuclease domains and is directed to its target dsDNA by a guide RNA, a transcript originating from a previous invader that matches the DNA sequence of a new invader. It

thereby creates an RNA-DNA hybrid and cleaves the invader DNA with the RuvC domain. The other one, HNH endonuclease, cleaves single-stranded DNA on the other side of the open loop. The HNH domain of the hybrid-specific enzyme contains two conserved histidines (H) and a central asparagine (N), which create a ZNF domain. The RuvC domain cleaves ssDNA, harboring an RNase H-like structure, and is split into three subdomains in the primary sequence of Cas9 (Jinek et al., 2014; Nishimasu et al., 2014) (**Figure 1C** and **Table 1**). The RuvC domain was originally identified in the RuvC resolvase that cleaves Holliday junctions, four-stranded DNA intermediates that form during recombination processes.

Remarkably, the RNase H fold is flexible enough to cleave the DNA moiety of a hybrid if a DNA phage or plasmid needs to be inactivated, whereas in the case of an RNA containing retrovirus the RNase H specifically destroys the RNA moiety of the hybrid. Thus, RNases H can specialize according to the required needs.

Poxviruses also encode a Holliday junction RNase H-like resolvase. Poxviruses are small giant viruses/*Megavirales* (described in more detail below).

Mammalian RAG1 contains a RING finger domain that is a type of ZNF domain mediating DNA binding. RAG1 is involved in V(D)J recombination during rearrangement of immunoglobulin genes and cleaves dsDNA at recombination signals via an RNase H domain with help of the RAG2 protein (Majorek et al., 2014; Moelling and Broecker, 2015). Then, V, D and J gene fragments are recombined. RAG1/RAG2 generate the diversity of T cell and B cell antigen receptors for antibodies. Both enzymes originate from a transposon and may have entered the vertebrate genomes from invertebrates such as sea urchin, sea star, and Aplysia more than 500 Mio years ago (Kapitonov and Koonin, 2015; Moelling and Broecker, 2015).

Prp8, the core protein of the eukaryotic spliceosome, has both an RNase H and an RT domain, suggesting an evolutionary origin from an ancient retroelement. Both domains have retained their three-dimensional structure and ability to bind to RNA but became enzymatically inactive (Dlakić and Mushegian, 2011). A conserved RNA recognition motif (RRM) in the RT domain is likely involved in pre-mRNA and/or small nuclear snRNA binding (Grainger and Beggs, 2005). The inactive RNase H domain may be essential for balancing accuracy and efficiency during the splicing process (Will and Lührmann, 2011; Mayerle et al., 2017). Interestingly, the RT-like domain of Prp8 likely originates from a prokaryotic group II intron (Dlakić and

TABLE 1 | RNase H-like family members and their functions.

RNHL family member	Species	Functions	Cleavage specificity
RNase H	(Para)retroviruses, ERVs, non-LTR and LTR retrotransposable elements of eukaryotes	During viral replication and retrotransposition, RNase H removes the RNA template for DNA synthesis by the RT	Endonuclease, RNA
Integrase	(Para)retroviruses, LTR retrotransposable elements	Removal of two or three nucleotides from the 3′ ends of the dsDNA copy for integration into host DNA	Endonuclease, DNA
Transposase	DNA transposons of pro- and eukaryotes	Excision of the transposon DNA from the host chromosome, for integration at a new site	Endonuclease, DNA
Terminases	DNA phages of prokaryotes	Terminases cleave the dsDNA concatemer genomes for DNA packaging and phage assembly	Endonuclease, DNA
Argonaute und Piwi-like proteins (PIWI domains)	Pro- and eukaryotes	Essential component of the RNAi immune system (siRNA, miRNA and piRNA pathways); cleavage of foreign nucleic acids, epigenetic or paragenetic silencing of TEs	Endonuclease, DNA or RNA
Cas proteins (RuvC domain)	Prokaryotes	Cas proteins for CRISPR-Cas immune system to inactivate foreign DNA of plasmids or phages	Endonuclease, DNA or RNA
RAG1	Mammals	RAG1 is a transposase, RAG1/2, are essential for V(D)J recombination, generate diversity of T and B cell receptors for antibodies	Endonuclease, DNA
Prp8	Eukaryotes	Prp8 is the "master regulator" of the spliceosome, its RNase H domain balances accuracy and efficiency of splicing	No enzymatic activity
RNase H (cellular)	Pro- and eukaryotes	Maintenance of genome stability; non-essential in prokaryotes and lower eukaryotes, but essential in higher eukaryotes	Endonuclease, RNA
RNase T	Prokaryotes	Processing of small RNAs and 23S rRNA, tRNA turnover	Exonuclease, DNA or RNA
DNA polymerases (3′ to 5′ exonuclease domains)	Pro- and eukaryotes, viruses	Proofreading during DNA replication	Exonuclease, DNA
Werner syndrome helicase (WSH) and related helicases of the RecQ family	Pro- and eukaryotes (WSH is a human protein)	Unwinding of dsDNA during repair of double strand breaks, single nucleotide damage; the RNase H domain degrades the recessed 3′ end	Exonuclease, DNA

Mushegian, 2011), and pre-mRNA splicing probably evolved from a group II intron ribozyme (Abelson, 2013). The catalysis originally performed by a single ribozyme has evolved into the spliceosome – a large multi-component RNA-protein complex involving dozens of proteins, with Prp8 in its center.

RecQ helicases also adopt an RNase H-like fold and exert 3′-5′ exonuclease activity. RecQ domains have been identified in various DNA polymerases of prokaryotes, eukaryotes, and viruses such as phage T7 (Majorek et al., 2014). RecQ is an ATP-dependent DNA helicase implicated in diseases such as Werner syndrome (Zuo and Deutscher, 2001). The RNase H-like domain of Werner syndrome RecQ helicase has 3′-5′ exonuclease activity and degrades the recessed 3′ ends during DNA repair processes (Oshima, 2000). Mutations can lead to Werner syndrome, which is associated with accelerated aging.

Eukaryotic RNases H1/H2 require protein–protein interactions to function. The active mammalian RNase H2 enzyme is a trimer of three subunits A, B, and C, with A being the catalytically active subunit and B and C stabilizing the proper conformation of A (Crow et al., 2006). Moreover, the B subunit interacts with the proliferating cell nuclear antigen (PCNA) to bind DNA polymerase for DNA replication and repair (Chon et al., 2009). PCNA guides the RNase H2 trimer to RNA primers or misincorporated ribonucleotides to exert maintenance and repair functions (Bubeck et al., 2011).

ABUNDANCE OF RNases H

Studying evolutionarily ancient proteins and their abundance is difficult since no protein fossils are available. They were first identified by protein architecture, conserved domains, or by conserved amino acids in their catalytic centers (e.g., aspartic acid and glutamic acid in RNases H). Among the most frequent proteins in the biosphere is the RNase H – more abundant than enzymes involved in nucleotide metabolism, polymerases or kinases (Ma et al., 2008), whereby structure undergoes more stringent evolutionary constraints than does sequence (Caetano-Anollés et al., 2009).

Transposases were described as the most ubiquitous (the majority of sequenced genomes encode at least one transposase) and abundant (present in highest copy number per genome) genes in nature (Aziz et al., 2010). Ubiquitous genes are essential and indispensable in every genome, whereas abundant genes can be frequent in only a few ecosystems. Transposases are ubiquitous but also have the highest copy number per genome and may accelerate biological diversification and evolution. They carry RNase H folds, of which recently more than about 60,000 unique domains were identified based on comparative structural analyses. RNase H-like domains can be grouped by their evolutionary relationships into 152 families (Majorek et al., 2014).

PLANKTON

Recently, metagenomic sequencing was performed with large sample sizes from marine samples, including small eukaryotes (protists), prokaryotes, ranging from 0.2 μm to 2 mm in size, as well as phages and viruses, obtained by the *Tara* Oceans project. According to a recent study, RTs predominate in the metagenomes (more than in metatranscriptomes), reaching up to 13.5% of the total gene abundance (Lescot et al., 2016). The authors identified about 3,000 retrotransposon/retrovirus-like RT and about 186 RNase H genes, but also 988 integrases, 556 endonucleases and helicases that add up to about 1,200 RNase H-like genes. Their weak transcriptional activity may reflect the active proliferation of retroelements that may contribute to genome evolution or adaptive processes of plankton. The RTs/RNases H are mainly found within retroelements of prokaryotes and eukaryotes. Retroelements constitute about 42% of the human genome, about 80% in maize and bread wheat, and 55% in red seaweed and many unicellular eukaryotes. Prokaryotes also harbor retroelements and DNA transposons, but less frequently than eukaryotes. About 25% of prokaryotic genomes encode at least one RT of over 1,000 different types (Simon and Zimmerly, 2008).

Sequencing of numerous genomes in the most distant organisms revealed only recently that RNase H-like molecules are among the most abundant protein entities on our planet (Simon and Zimmerly, 2008; Caetano-Anollés et al., 2009; Majorek et al., 2014). This is likely due to the fact that transposons, retrotransposons, and other retroelements are extremely abundant on our planet.

The RT has previously been better characterized than the RNase H-like superfamily, hence, we were wondering about the total abundance of RNases H-like genes. For that we analyzed the RNase H gene superfamily distribution and abundance in prokaryote-enriched samples of the *Tara* Ocean (Sunagawa et al., 2015) across the global ocean at three depth layers shown globally (left) and regionally (right) (**Figure 2**). RNase H-like genes with homology to a set of 151 RNase H-like gene families were identified in all regions. On average about 10 to 15 RNase H-like gene copies per cell were detectable at all three levels. These numbers are comparable to previous findings that an average of about 13 transposase and integrase genes, the two most abundant RNase H-like genes, are present per genome, including viral, prokaryotic and eukaryotic species (Aziz et al., 2010). Thus, our data provide evidence that RNase H-like genes are probably among the most abundant gene superfamilies found in plankton organisms throughout the global oceans.

Interestingly, certain plankton populations and gene functions as judged by taxonomic marker gene sequences and gene family abundances, were dominant and shared among different regions of the oceans designated core taxa and core gene families, respectively (Sunagawa et al., 2015). Dispersal mechanisms by currents are thought to distribute these species and their genes. The less abundant species were not easily detected (Sunagawa et al., 2013). A similar phenomenon about core sequences we observed in the human gut microbiota when we analyzed the virome in comparison to the bacterial and fungal communities of a patient, who underwent a fecal microbiota stool transfer, where core sequences also dominated (Broecker et al., 2017). Thus, in the oceans, it appears that RNase H-like proteins represent

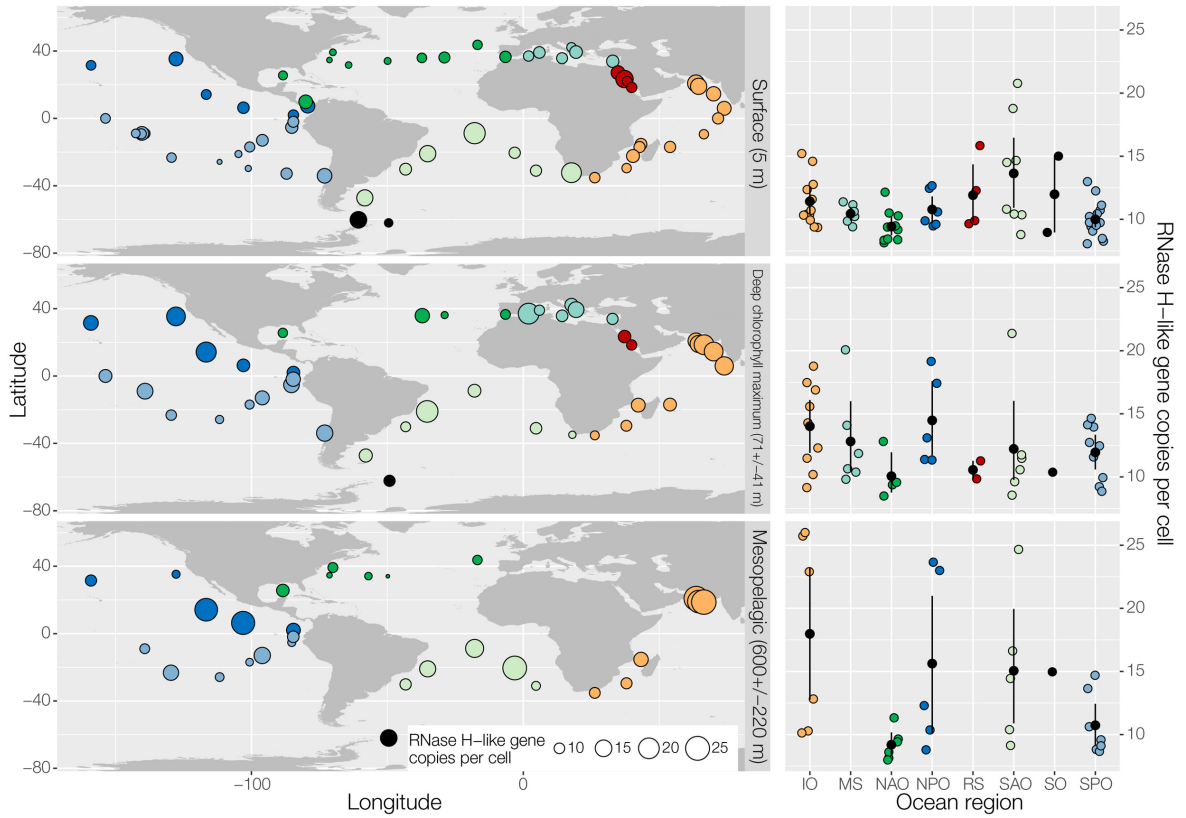


FIGURE 2 | Effective abundance of RNase H-like gene family members across the global ocean. The geographic distribution of effective abundances (per cell) of all genes homologous to a set of 151 RNase H-like gene families (Majorek et al., 2014) in the ocean is shown globally (left) and regionally (right) for three different depth layers (top: surface; middle: deep chlorophyll maximum; bottom: mesopelagic). Samples correspond to prokaryote-enriched size fractions collected in the context of the TARA Oceans project (Sunagawa et al., 2015). Effective gene abundances were calculated based on annotating the Ocean Microbial Reference Gene Catalog by homology using HMM profiles of RNase H-like superfamily members (Majorek et al., 2014) with a bit score cutoff of 20. For each sample, the sum of RNAse H-like gene abundances was normalized by the median abundance of 10 universal single copy phylogenetic marker genes (Sunagawa et al., 2013) to calculate effective abundances. MS, Mediterranean Sea (light blue); RS, Red Sea (red); IO, Indian Ocean (orange); SAO, South Atlantic Ocean (light green); SO, Southern Ocean (black); SPO, South Pacific Ocean (blue); NPO, North Pacific Ocean (dark blue); NAO, North Atlantic Ocean (dark green).

important core sequences, as they were identified in all samples tested here (**Figure 2**).

PROKARYOTES

The abundance of RTs without RNase H domains in bacteria is surprising (Simon and Zimmerly, 2008), many of them with unknown functions. It seems that host factors, precursors of RNases H or nucleases are often involved in removing RNAs instead of an attached RNase H. A number of 1021 RTs were identified in bacteria with the majority being those of group II introns (742, 73%) that encode the RT with its seven typical domains including palm and finger domains, as well as an additional endonuclease (Simon and Zimmerly, 2008). Group II self-splicing introns, a large class of mobile ribozymes, are found in all domains of life, eukaryotes, bacteria, archaea, plants, and marine plankton. An RNA loop designated as lariat RNA, and two molecules of intron-encoded protein X form an RNA-protein (RNP) complex that mediates mobility. It is site-specific and recognizes intron sequences transcribed into a DNA, targeted for target-primed RT. Group II introns may have evolved by fusion of a ribozyme with the DNA coding for the RT (Zimmerly and Semper, 2015). Group II introns are the only ones in bacteria, which are mobile (Simon and Zimmerly, 2008). Since group II introns encode an RT that lacks an RNase H domain, they require host RNases H to degrade the intron RNA template (Lambowitz and Belfort, 2015). Thus, an RNase H was required for the retromobility of the evolutionarily most ancient retroelements (**Figure 3A**).

Bacteria also harbor retrons that are related to retroelements with an unusual branched multicopy single-stranded DNA/RNA (msDNA/RNA) structure, in addition to a gene for a chromosomally encoded RT element (Inouye et al., 2011). RT partially reverse transcribes the RNA to form the branched RNA-DNA molecule by a 2′-5′ bond. This msDNA accumulates to high levels within the cell, yet its function remains elusive. The msDNA is about 3.5 billion years old. It is not mobile independently, consisting of an RT and overlapping multicopy

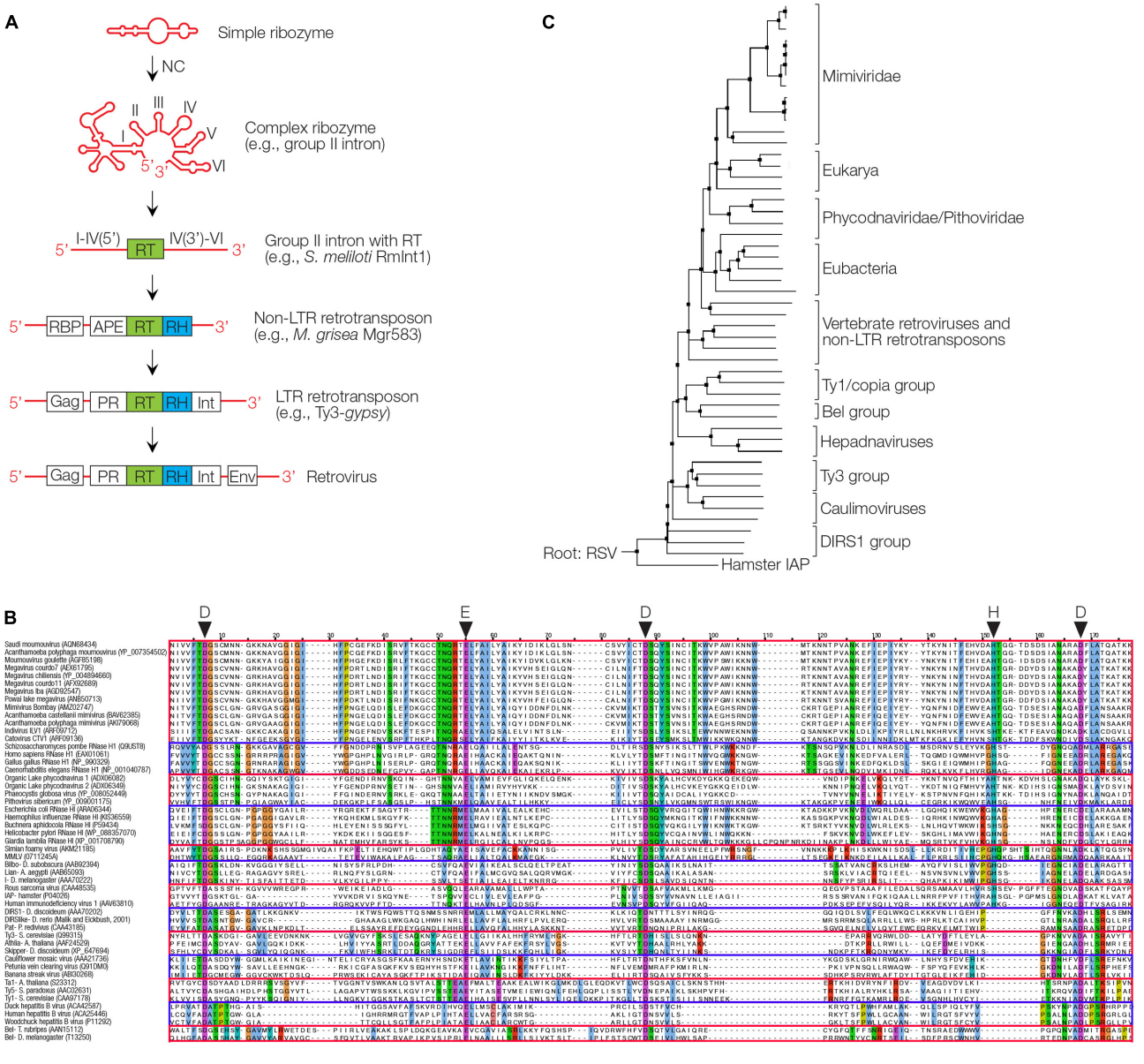


FIGURE 3 | Proposed evolutionary history of the RT/RNase H and phylogeny of RNases H. **(A)** Proposed evolution from a ribozyme or viroid to group II intron, with RT, RNase H (RH) via intermediates to a retrovirus. NC is a nucleoprotein that enhances ribozyme activity and reverse transcription. APE, apurinic endonuclease; Gag, group-specific protein; Env, envelope; Int, integrase; PR, protease; RBP, RNA-binding protein. **(B)** Sequence alignment of RNases H are indicated, based on the highly conserved aminoacids D, D, E, D and partially conserved H as indicated. **(C)** Phylogenetic tree of RNase H domains found in giant viruses (*Mimiviridae*, *Phycodnaviridae*, and *Pithoviridae*), eukaryotes, various retrotransposons, retroviruses and pararetroviruses (Hepadnaviruses and Caulimoviruses). The tree is rooted to the RNase H of chicken Rous Sarcoma virus (RSV). DIRS retrotransposon found in *Dictyostelium* is the origin of many siRNAs (Boesler et al., 2014), hamster intracisternal A type particles (IAP) are transposable elements closely related to retroviruses (Qin et al., 2010).

ssDNA/RNA genes that resemble tRNA-like primers used by the retroviral RT. These structures are reminiscent of a transition from the RNA to the DNA world (Moelling and Broecker, 2015; Moelling, 2017).

Diversity-generating retroelements (DGRs) are a class of mobile elements found in phages that infect whooping cough causing *Bordetella* bacteria. *Bordetella* bacteriophage BPP-1 integrates into the bacterial genome as temperate phage and encodes an error-prone RT. The RT generates variants of the phage protein at the tip of the tail fibers of the phages. This allows for tropism changes by switching binding specificity to bacterial receptors. The number of variants is extraordinary and reminescent of V(D)J antibody diversity. DGRs are widely distributed in bacterial chromosomes, phage and plasmid genomes (Miller et al., 2008; Guo et al., 2014).

EUKARYOTES

The abundance of RT and RNase H genes in prokaryotes and unicellular eukaryotes leads to the question about their abundance in eukaryotic genomes. The human genome contains about 40,000 complete or partially truncated HERVs. HERVs are proviruses of ancient retroviral infections that have mostly degenerated to solitary LTRs (about 400,000 per genome) via homologous recombination. In addition, the human genome contains 850,000 LINEs, the long interspersed nuclear elements, including the most recent and active LINE-1 (L1) subclass, and 1,500,000 SINEs (Lander et al., 2001). Human L1 elements, of which about 100 can actively retrotranspose, encode an RT and endonuclease but no RNase H. They do not utilize RNA primers that would require an RNase H, but a target site primed mechanism for reverse transcription. Instead of degrading the RNA template for dsDNA synthesis, the LINE RT is able to displace the RNA strand during synthesis (Kurzynska-Kokorniak et al., 2007). SINEs, about 10,000 of which are active in the human genome, rely for retrotransposition on gene products of LINEs supplied *in trans*, such as the RT. Movements require endonucleases, no RNases H. Overall, almost 50% of the human genome consists of TEs, including the LINEs and SINEs and a small fraction of DNA transposons. These sequences can be regarded as a graveyard of previous infections by retroviruses and retroviral-like elements, or as a viral archive (Lander et al., 2001). Most TEs have accumulated mutations and deletions over time, rendering them inactive. Interesting, DNA transposons are rare in humans. Of all TEs, DNA transposons constitute less than 10% (over 90% being retroelements), but they are relatively more frequent in other organisms such as *Drosophila melanogaster* (over 20%). Many plants have more than 50%, *Caenorhabditis elegans* close to 90% and *Trichomonas vaginalis* almost 100% (Feschotte and Pritham, 2007).

Human ERVs and ERVs are the results of genuine retroviral infections in the past, as demonstrated by reconstitution of an infectious retrovirus named 'Phoenix,' obtained from alignment of near-intact HERV-K(HML-2) proviruses that first invaded the human genome about 35 Mio years ago (Dewannieux et al., 2006; Lee and Bieniasz, 2007). The Alu elements (a subclass of SINEs), which are non-coding RNAs, are somewhat reminiscent of ribozymes or viroids. The total number of retroviruses and their genes calculated from the number of 400,000 solitary LTRs would exceed the total size of the present human genome that is about 3.3 gigabases. Could it have been larger in the past as is the case in some plants? TEs are even more abundant in some plant genomes, e.g., the maize and wheat genomes contain about 80% TEs, mostly LTR retrotransposons, enlarging the total genome size to 17 gigabases in the case of wheat (Baucom et al., 2009; Brenchley et al., 2012). The large genomes may be a consequence of breeding for improving the yields for food supply. Similarly, prokaryotic genomes can contain phage sequences within CRISPR arrays, albeit in smaller quantities per genome (see below). The promoters of retroviruses or retroelements such as the LTRs can influence host gene expression, for instance, by supplying transcription factor binding sites, altering chromatin structure, or promoting regulatory non-coding RNAs both *in trans* and *in cis* (Broecker et al., 2016a). The transcription of many REs, including HERVs, is upregulated in disease states including cancer. However, whether RE activity is causal for cancer or is simply a bystander effect remains unknown. Retroviruses and REs can mobilize flanking sequences, and thereby cause gene duplication events, one of the most impactful mechanisms for creating genes with novel functions (Feschotte and Pritham, 2007).

Despite the accumulation of often deleterious mutations, TEs frequently contain functional promoters and contribute to a large fraction of the human transcriptome (Faulkner et al., 2009; Broecker et al., 2016a). TEs can influence cellular genome architecture and function by means of gene duplication or mutagenic events caused by integration (Chuong et al., 2017). While one copy can continue fulfilling acquired necessary functions, the other copy can change substantially. An interesting example of a gene duplication event is the RNase H of retroviruses, whereby the RNase H linked to the RT has deteriorated to an inactive enzyme with only linker function to the neighboring active RNase H (Ustyantsev et al., 2015).

Every retrovirus infection supplies about 10 novel genes to the cell, including those encoding the RT, RNase H and integrase. Viral integration events can lead to horizontal gene transfer (HGT) or recombination events. For example, there are about 100 retroviral oncogenes known, whereby some of them are relevant targets for human cancer therapies today such as Raf, ErbB, etc. (Flint et al., 2015).

In summary, the global abundance of RTs and RNases H can be attributed to retroelements, retrotransposons, retro- and pararetroviruses, and endogenous viruses in eukaryotic genomes. RTs are more prevalent and occur also without RNases H. There are RT groups such as prokaryotic RTs, encoded by group II introns, retrons, retrogenes, and DGRs. There are also solo RNases H without RTs, which are not the exceptions but the rule.

EVOLUTION OF RNase H

In the early RNA world, before proteins and DNA arose, simple self-replicating RNAs with ribozyme activity formed as primary biological entities that were non-coding (nc) but relied on structural information based on robust hairpin-looped structures. These ribozymes are capable of cleaving, joining, and evolving, as demonstrated experimentally (Lincoln and Joyce, 2009). Replication must have occurred in a prebiotic environment at hydrothermal vents down in the dark oceans with energy supply from chemical reactions without light. Ribozymes lack coding information but rely on structural information, and today are still important for the biological function of the vast majority of ncRNAs in eukaryotic cells and RNA viruses. Ribozyme-like elements act as regulatory circular or circRNAs, as so-called 'sponges' for small RNAs (Hansen et al., 2013). Thus, circRNAs regulate other regulatory RNAs, which may be defined as back-up or chief regulation. circRNAs may have survived until today as living fossils because of their exquisite stability (Moelling, 2012, 2017). These ancient mobile genetic

elements are present until today, not only in eukaryotes but also in prokaryotes (Lambowitz and Zimmerly, 2004).

Remarkably, the ribozymes are highly related to today's viroids, catalytically active circular RNAs. They are naked viruses, free of proteins until today. Since such elements were not considered as viruses, they were designated as viroids. Just like some other viruses the viroids can be pathogenic and inflict significant damage to many plants (Moelling, 2017). They contain a core region, which is active as siRNA for gene regulation.

Their enzymatic activity has been lost in some viroid species today – presumably in the rich cellular environment of host cells, which today harbor viroids. Gene loss as a consequence of a rich milieu is a known principle. Even in the nutrient-rich environment of the guts of an obese patient the complexity is reduced (Moelling, 2017).

Improvement of the catalytic activity of ribozymes early during evolution can be easily imagined coming from RNA binding proteins (RBPs) or small peptides with positive charge based on *in vitro* studies. Ribozymes can be stimulated enormously by the addition of RBPs, as shown for the HIV nucleocapsid (NC), leading up to a 1000-fold stimulation of ribozyme activity, which we discovered during studies to improve ribozymes for gene therapy (Müller et al., 1994).

Nucleocapsids can be detected in every RNA virus today as nucleocapsids or ribonucleoproteins (Flint et al., 2015). They are surprisingly multifunctional proteins serving many purposes. In addition to significantly enhancing the catalytic activity of ribozymes, they protect the RNA from degradation and also enhance catalysis of the RT by acting as chaperones which is almost counterintuitive but is based on the unwinding effect and the cooperativity of the multimeric NC proteins (Müller et al., 1994). NCs are rich in basic peptides such as lysine and arginine, which may have formed as smaller precursors of the NCs, and are essential components in all RNA viruses today. The NC of HIV has three such basic stretches surrounding two zinc fingers, which provide high flexibility (Müller et al., 1994) (**Figure 3A**).

Peptides could have formed in the prebiotic environment at hydrothermal vents, before the translation machinery, codons or DNA arose. The protein translation machinery must have evolved later, since ribozymes themselves contributed to the protein synthesis machinery by supplying the enzymatically active component at the center of the protein synthesis apparatus. Ribosomes today consist of about one hundred scaffold proteins for maintenance of ribosomal structure and function, and in addition some ribosomal RNAs (that serve as the basis for determination of bacterial species in microbiomes). "The ribosome is a ribozyme" is the title of an article by the Nobel Prize awardee Thomas Cech (Cech, 2000), one of the discoverers of ribozymes. Viruses with only tRNA-like structures may have contributed to the evolution of the protein synthesis machinery. Such narnaviruses exist till today in fungal species (Moelling, 2017). Also, today's SINE or Alu sequences may fit into this concept of the predominant role of early non-coding RNAs.

Then there is a rare example of a retroviroid described in carnation plants. Carnations harbor a small viroid-like RNA, CarSV, whose homologous DNA is generated by an RT (Hegedus et al., 2001). Thus, this viroid exploits an RT, presumably provided *in trans* by a plant pararetrovirus, such as cauliflower mosaic virus.

Patel et al. (2015) recently succeeded in a "one-pot" reaction to synthesize nucleotides, fatty acids and amino acids from six elements (sulfur, nitrogen, oxygen, hydrogen, phosphorous, and carbon) in the test tube.

It is a frequent evolutionary progress and improvement that RNA leads to proteins, fulfilling similar functions but with significantly increased efficiencies. It can be easily envisaged that ribozymes became RNases H, probably by multistage processes. How the RT evolved, is still a matter of speculation. The transition from RNA to DNA must have occurred to conserve and stabilize beneficial achievements and may have led to an RT that is ubiquitously found in all coding group II introns, and whose appearance marks the beginning of the transition from the RNA to the DNA world (Lambowitz and Zimmerly, 2004; Koonin et al., 2006). Group II introns consist of highly structured RNA that developed coding capacity for an RT gene (**Figure 3A**).

An interesting intermediate between the RNA and DNA world was mentioned above, the msDNA/RNA, the retrons (Inouye et al., 2011). They appear like frozen intermediates and may be relics from early steps in evolution. They may be more ancient than the separation between prokaryotes and eukaryotes. They contain a very unusual branched rG residue covalently linking RNA and DNA. The open reading frames giving rise to msDNA/RNA are the shortest and simplest of the retroelements containing only RNase H activity in addition to the RT activity. Thus the retrons may point to the earliest possible roots of these elements and these two enzymes (Moelling, 2017).

RNase H and RT are involved in intron splicing by forming loops, the lariats. Today a eukaryotic spliceosome includes dozens of proteins. Surprisingly, Prp8 at the core of the spliceosome encodes an RT and RNase H domain, albeit both without enzymatic activities.

Group II introns have acquired an additional endonuclease activity (En) with an HNH fold. The En was likely acquired later during evolution than the RT, since no group II introns have been identified with an En in the absence of an RT (Lambowitz and Zimmerly, 2004). Of note, retrotransposition of group II introns requires host RNases H to remove the RNA template and to enable subsequent dsDNA synthesis by the RT (Smith et al., 2005). RT-encoding group II introns are likely the evolutionary precursors of non-LTR retrotransposons such as the human LINEs (Zimmerly and Semper, 2015); their RTs are highly related (Lambowitz and Belfort, 2015). LINEs no longer have ribozyme activity and encode a limited number of proteins in addition to the RT. An apurinic endonuclease (APE) cleaves DNA to initiate reverse transcription (Malik et al., 1999), replacing an RNA primer that would require digestion by an RNase H.

An evolutionary advancement of non-LTR retrotransposons, compared to group II introns, is their independence of a foreign RNase H for retrotransposition. Other non-LTR retrotransposons lacking an RNase H have evolved to an RT that displaces the RNA template during second-strand DNA synthesis

(Kurzynska-Kokorniak et al., 2007). An RNase H enzyme likely evolved later than the APE and RBP (Malik et al., 1999).

Non-LTR retrotransposons evolved into LTR retrotransposons (Malik and Eickbush, 2001), the known groups being Ty1-copia, Ty3-gypsy, and BEL-Pao-like, which all encode both an RT and an RNase H (Majorek et al., 2014). Interestingly, the RNase H domain, compared to that of non-LTR retrotransposons, has lost a subdomain perhaps resulting in weaker catalytic activity (Malik and Eickbush, 2001).

An important evolutionary event was the duplication of the RNase H domain with one component leading to a tether or connection region, an inactive RNase H. This duplication is surprising for minimalistic viral genomes and may serve to fine-tune the cleavage activity of the functional RNase H by the formation of p66/p51 heterodimers found in some retroviruses, which is associated with increased rates of DNA strand transfer during reverse transcription (Ustyantsev et al., 2015). A similar 'dual' RNase H, active and inactive, is present in some Ty3-gypsy LTR retrotransposons (Ustyantsev et al., 2015).

The integrase exhibits an RNase H fold likely derived from the DDE transposase of a bacterial DNA transposon (Malik and Eickbush, 2001; Rice and Baker, 2001; Majorek et al., 2014). This enzyme processes the reverse transcribed dsDNA copy of the element by cleaving two to three nucleotides from the 3′ ends to expose the invariant terminal dinucleotides for insertion into host DNA (Flint et al., 2015).

The archaeal RNase H2 has been suggested to be derived from retrovirus elements (Ohtani et al., 2004).

Compared to LTR retrotransposons, retroviruses additionally gained an Envelope (Env) protein that is required for cell-to-cell transmission. Env proteins of different retrovirus lineages may have been acquired independently from different viral sources. For instance, Env of gypsy/metaviruses is likely derived from baculoviruses, dsDNA viruses of insects. The Env-derived cellular protein Syncytin contributes to syncytia formation by cell-cell fusion and originates from an endogenous retrovirus ERV-W of about 35 Mio years ago (Dewannieux et al., 2006). Due to the immunosuppressive properties of ERV-W Env, the derived syncytin prevents immune rejection of the embryo by the mother in humans and other species.

It is rather unknown that there are not only retroviruses but even retrophages in bacteria. Assuming that the RNA world preceded the DNA world one may ask whether the abundant DNA phages or any other DNA viruses had RNA or retro-precedents in ancient times. Only a few such intermediate-type viruses are known, such as the BPP-1 retrophage hosted by *B. pertussis* bacteria. This temperate phage expresses an RT whose infidelity exerts a mutagenic effect on the phage receptor gene, which can alter phage tropism. At least 36 types of such retrophages exist (Guo et al., 2014). The infidelity of the RT leads to about one mutation per thousand nucleotides and round of replication. This is a major force for change of bacterial tropism and evolution in general. Thus, these diversity generating retroelements, DGRs, demonstrate the contribution of a phage retroelement with mutagenic RT to genetic diversity and genomic variation of surface proteins of phage particles, but also of bacterial cells themselves, such as *Legionella pneumophila* (Guo et al., 2014).

One may speculate that the fast replication rate of phages and high numbers of generations may have allowed them to progress away from the RNA world and the retrophages, resulting in predominantly dsDNA phages in today's biosphere (Moelling, 2017). The potential evolution from a non-coding ribozyme or its close relative, the viroids, to a coding one (group II intron), with the support of a NC, then to non-LTR and LTR REs and, finally, to retroviruses, is depicted in **Figure 3A**. If viroids are allowed to be defined as naked "viruses," just lacking proteins, then the most ancient form of life was a virus, a naked viroid, a ribozyme. Then viruses would be our "oldest ancestors" (Moelling, 2012, 2013, 2017)!

RNase H IN GIANT VIRUSES

Viruses or virus-like elements as the beginning of an RNA world have built up to bigger and more complex entities. Recently, intermediates between viruses and bacteria have attracted attention, the giant viruses or *Megavirales*. They are the biggest viruses known, surpassing the size of many bacteria and some encode genes involved in the protein translation machinery, an indicator of independent life. Are they half-finished bacteria or regressed from bacteria?

Interestingly, giant virus genomes can harbor retrotransposons (Maumus et al., 2014). The gvSAG AB-566-014 virus was found to encode an RT and transposase, a nuclease with an RNase H fold (Wilson et al., 2017). The virus is related to the *Cafeteria roenbergensis* (Cro) virus, a giant marine virus widespread in protists with more than 500 genes and a dsDNA genome of 730,000 nucleotides. We mined the NCBI protein and nucleotide databases for RNase H genes in the genomes of giant viruses. We thus identified a total of 17 unique RNase H proteins sequences (**Figures 3B,C**). The sequence alignment of RNase H-like enzyme sequences in comparison to known sequences with the highly conserved aminoacids D, E, D, D and the partially conserved H as indicated. The origin of the RNases H and a comparison will be subject of further analysis (Russo et al., unpublished observation).

The conserved amino acids (DEDD) are hallmarks of RNases H, and were identified in all of them, indicative of enzymatically active proteins. Giant virus RNases H stratified into two distinct clades, one containing all RNases H identified in Mimivirus-like genomes, and the other cluster containing RNases H of *Phycodnaviridae* and a *Pithoviridae*. Both clades seem to be related to eukaryotic RNases H and likely share a common ancestor, while the *Phycodnaviridae*/*Pithoviridae* RNases H are more related to eubacterial ones.

A prominent *phycodnavirus* is a green algae virus that infects *Emiliania huxleyi* coccoliths and leads to algae bloom. It also generated millions of years ago the white cliffs of Dover. We identified *Chlorella* virus sequences in the intestine of a patient after fecal transfer because of a *C. difficile* infection (Broecker et al., 2016b). No disease is known to be associated with intestinal *phycodnaviruses*.

VIRUSES PROTECT AGAINST VIRUSES

Invading viruses trigger cellular antiviral responses, whereby the first virus protects the host against a second virus, at least for some time. This allows the first virus to replicate and produce progeny without competition, since resources within the cell are limited. This phenomenon is called superinfection exclusion, first described in bacteria. The viral gene products themselves once integrated into host cells can directly interfere with *de novo* infections of related viruses (Moelling et al., 2006; Moelling, 2017). Viruses can also induce cellular antiviral responses indirectly. Superinfection exclusion is found in representatives of many viral lineages, such as the positive strand ssRNA virus hepatitis C virus (Schaller et al., 2007), retroviruses including HIV (Nethe et al., 2005), small DNA viruses and the phage phiX174 (Hutchison and Sinsheimer, 1971), *Caudovirales* phages like T4 (Lu and Henning, 1994) and large DNA viruses such as *Poxviridae* (Laliberte and Moss, 2014). It appears likely that an analogous viroid/ribozyme-based superinfection exclusion system existed before the evolution of more complex viruses or cellular immune systems such as RNAi. There are different ways how viruses achieved a monopoly after entering a host cell. The strategies of viruses and antiviral responses will be discussed in below (**Figure 4**).

RIBOZYMES MEDIATE IMMUNITY

Ribozymes have likely been among the first biomolecules, perhaps resembling present-day ribozymes and viroids. A ribozyme/viroid may have prevented invasion by other viroids through RNA cleavage *in trans* (**Figure 4A**). This could have happened at sufficiently high concentrations even in the absence of cells, perhaps in Darwin's famous 'warm little pond' or next to "Black smokers," where chemical energy was available. Ribozymes/viroids exhibit circular hairpin loop structures, have no coding capacity but structural information only. Many of them are catalytically active until today. With the advent of cells, catalytic activity may have been lost with host cells increasingly taking over enzymatic functions. The concept that a first viroid prevents a cell from infection by a second invader, in most cases a related one, has been first described in the 1930s in *Solanaceae* plants infected with the Potato-X-Virus (Salaman, 1933). Many more examples have been demonstrated in virus-host systems of various prokaryotes, animals, plants, and humans (Folimonova, 2012; Moelling, 2017).

The cellular organism benefits from superinfection exclusion when the infecting virus is mildly pathogenic and protects against a more virulent virus. This phenomenon can be considered an early form of immune system, immunization by an 'attenuated' virus. Superinfection exclusion has been exploited in agricultural practice, whereby crops are infected on purpose with mild viral isolates to induce 'cross-protection' (Niblett et al., 1978; Folimonova, 2012). Today, viroids are exclusively found in plants, the only known exception being the viroid-like hepatitis delta virus (HDV) that infects humans. HDV is a catalytically active naked RNA virus that requires the pararetrovirus HBV to supply proteins necessary for cell to cell transmission (Taylor, 2015). Possible mechanisms of cross-protection by viroids include post-transcriptional gene silencing such as RNAi (Kovalskaya and Hammond, 2014). Although the known natural ribozymes or viroids are self-cleaving, they can be modified with relative ease to give *trans*-cleaving derivatives (Jimenez et al., 2015). Therefore, *trans*-cleaving ribozymes might also have existed or may still exist in nature.

Cleaving foreign nucleic acid molecules may be characteristic of the viroid-based immune systems while later in the DNA/protein world, molecular scissors such as RNase H-like catalytic proteins with much higher efficiencies replaced ribozymes.

siRNA SILENCING/RNA INTERFERENCE

An RNAi-like defense mechanism to silence invading nucleic acids has likely evolved early during evolution, as variants of RNAi are present in all three cellular domains of cellular life (Koonin, 2017). The dsRNA of the invader is fragmented and processed to siRNA. Cell-expressed miRNA with mismatches to the target RNA is processed analogously and is used for miRNA-mediated gene silencing. The absence of RNAi in few organisms, including the yeast *S. cerevisiae*, is likely the result of gene loss (Cerutti and Casas-Mollano, 2006). RNAi, however, requires proteins, giving rise to the question how a primordial, pre-protein immune system might have looked like. siRNAs serve for antiviral defense in plants, *C. elegans,* and in many other species (Sarkies and Miska, 2014). Silencing involves RNase H-like activities found in Ago proteins for RNA cleavage and degradation. This terminates viral replication and mediates or suppresses gene expression (**Figure 4B**).

Most striking is the analogy between the antiviral siRNA system and retroviral components. We noted that PAZ-PIWI domains of Ago proteins closely resemble RT-RNase H of retroviruses. The antiviral defense system shares surprising similarities with the invading virus, including almost a dozen components of toolboxes for invasion and defense. The RT primer-grip corresponds to the PAZ pocket, DNA unwinding activity of the RT is found in the helicase domain of Dicer, the integrase generates dinucleotide overhangs with 3′-hydroxyl groups similar to Dicer, the nucleocapsid is a melting protein that protects viral RNA and may be the equivalent of the TAR RNA-binding protein, TRBP, and Fragile X Mental Retardation protein, FMRP, and the retroviral protease may be analogous to a cellular caspase (Moelling et al., 2006). Thus, a virus infection can supply the host with genes that directly become antiviral defense molecules. Retroviral infections can supply the siRNA tools for antiviral immunity.

Surprisingly enough, siRNA is not inhibiting retroviruses. A newly invading retrovirus encounters interference by several other mechanisms such as receptor blockade (see below), but not by the siRNA-based defense system. Possibly, RNAi evolved into a more efficient protein defense system against retroviruses in higher organisms. This question prompted us to analyze if antiviral siRNA activity is present in mammalian cells.

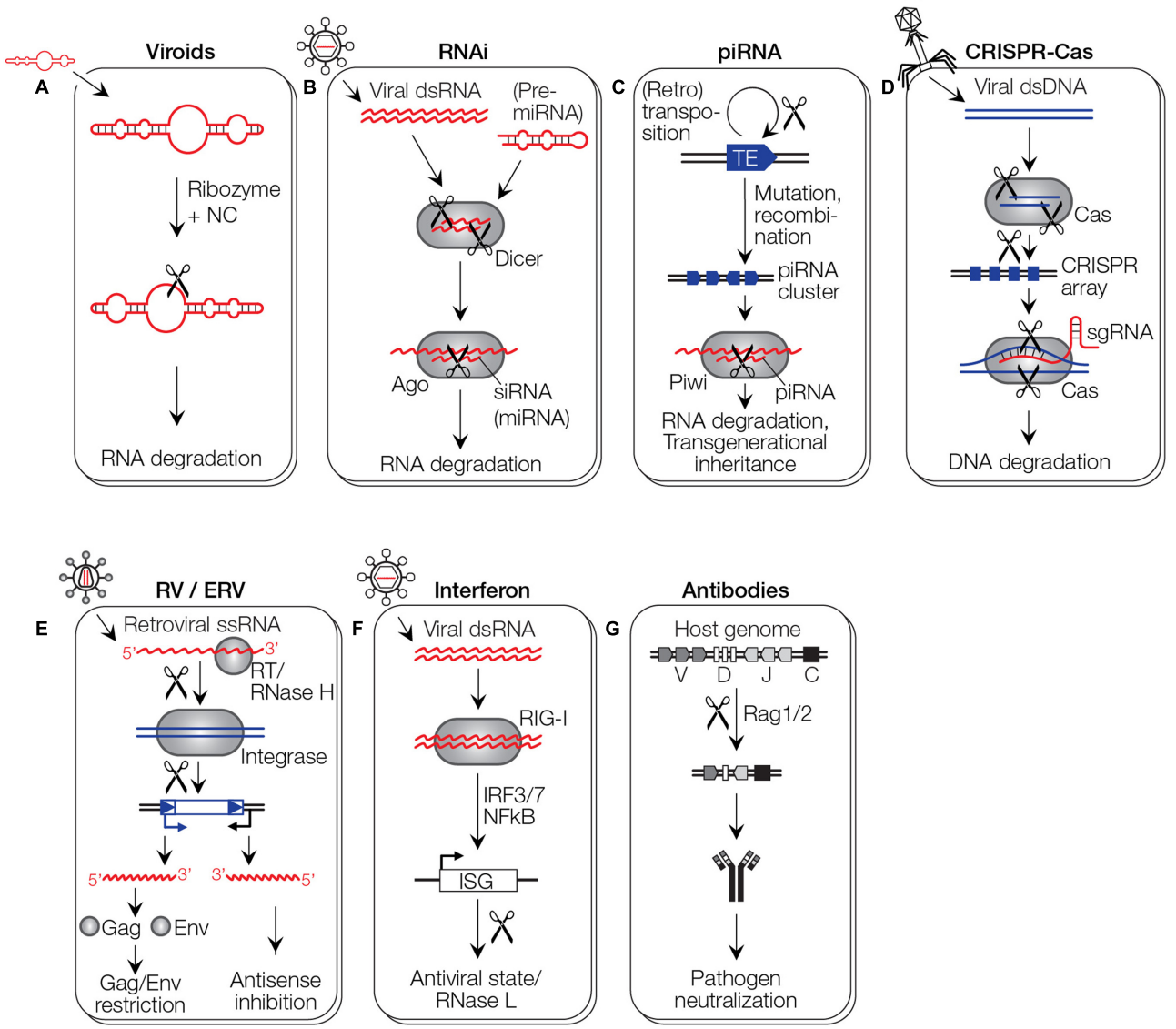


FIGURE 4 | Immune system pathways. **(A)** Proposed model for an RNA-mediated immune system. A viroid enters and acts on a related viroid *in trans* by ribozyme-mediated cleavage (scissors). **(B)** RNAi requires dsRNA or pre-miRNA transcripts that are fragmented by Dicer. The resulting siRNA/miRNA acts as guide to target Ago to incoming viral RNA, whereby the nuclease domain PIWI of Ago has an RNase H fold. Whether miRNAs mediate antiviral defense is still unclear (see text for details). **(C)** In the piRNA system of animals, a clustered archive of transposable elements (TE) is generated, likely involving chromosomal rearrangement events. The piRNA cluster is transcribed in both orientations, and the resulting dsRNA is fragmented into piRNAs that target TE RNAs for cleavage by the PIWI RNase H or epigenetic repression. piRNAs are implicated in transgenerational inheritance. **(D)** In CRISPR-Cas immunity, invading DNA is fragmented by the Cas1/2 complex and fragments are integrated into CRISPR arrays. The array expresses guide RNAs that target an effector Cas9 protein to invading DNA. Two cuts are made to induce a double-strand break, one involving an RNase H-like RuvC domain, the other an HNH endonuclease domain. **(E)** Infecting retroviruses (RV) or ERVs can act against invading viruses with similar sequences when their provirus is expressed in antisense orientation by a host promoter, or a bidirectional 3′ LTR. Proteins expressed from proviruses can interfere with the infection of similar retroviruses by preventing infection through Gag-like proteins, or receptor interference through Env-like proteins. **(F)** The interferon system recognizes dsRNA molecules through RIG-I like receptors, inducing a signaling cascade that leads to the expression of interferon stimulated genes (ISGs) that induce an antiviral state in the cell. **(G)** V(D)J recombination involves the RAG1 protein with an RNase H-like domain that generates, in cooperation with the RAG2 protein, the diversity of antibodies (and T cell receptors, not shown).

Indeed, we demonstrated a weak Dicer-dependent reduction of influenza virus production by about fivefold, which was only detectable in the absence of the interferon system that dominates in mammalian cells (Matskevich and Moelling, 2007). It is surprising that the weak siRNA defense system has been preserved in mammalian cells throughout evolution (Moelling, 2013; Moelling and Broecker, 2015). It is weak in mammals, suggesting that it may be a left-over that is now overshadowed

by interferons or other defense systems. RNAi mediated antiviral defense in mammals is debated by others (tenOever, 2017).

In contrast, RNAi in *C. elegans* is strong enough to efficiently prevent viral infections. siRNA is even secreted to alert neighboring worms, almost as in a larger multicellular organism. Only one virus is known to infect *C. elegans*, the Orsay virus isolated from orchards near Paris (Félix et al., 2011). Orsay virus can only be studied in worms without functional RNAi. In addition, the *C. elegans* genome harbors relatively few TEs, which may also indicate the presence of strong antiviral response. TEs and REs are among the first invading (retro)virus-like elements. In mammals, other vertebrates, and plants the RNAs of TEs are at least partially cleaved or 'domesticated' by the cellular antiviral defense, counteracting potentially dangerous abundant transposition by TEs, REs, SINEs, LINEs, even Alu RNA and ERVs (Roberts et al., 2014; Qin et al., 2015). This suppresses retrotransposition and supports genomic integrity (Heras et al., 2013).

piRNAs IN TRANSGENERATIONAL INHERITANCE

A variation of the siRNA silencing system is the PIWI-interacting RNA (piRNA) system. It was detected by the Canadian biologist Royal Alexander Brink when he studied maize genetics, similar to Barbara McClintock about 70 years ago. They were both puzzled by the genetics of the colors of maize kernels, which did not follow the Mendelian laws of inheritance. This way McClintock discovered epigenetics and Brink paragenetics or transgenerational inheritance. Both phenomena are based on environmental influences that induce transient modifications of the genomic DNA, not stable mutations (Moelling, 2017).

As a principle in nature all epigenetic modifications of the genome accumulated by the parents are erased from the DNA in their germ cells for their offspring, so that they start with pluripotent cells with unmodified genomes. Germ line cells and stem cells have a dedicated safety mechanism that is mediated by piRNAs. They guide Piwi proteins to transcripts of REs, with the PIWI RNase H domain silencing their activity (Malone and Hannon, 2009). This is essential for fertility of the sperm. About 26–31 nucleotides in length, piRNAs are slightly larger than siRNAs or miRNAs (21–24 nucleotides) and can influence gene and TE expression by inducing epigenetic modifications such as CpG methylation in promoter regions and changes in chromatin structure (Wilson and Doudna, 2013). Especially chromatin changes can last for many generations. piRNAs exist in insects, zebrafish, and rodents, more than 50,000 unique piRNAs were detected in mice. piRNAs are transcribed from piRNA clusters and are in antisense orientation to the TEs that are targeted by the PIWI protein for degradation (**Figure 4C**). Processing of piRNAs involves the so-called ping pong amplification loop. piRNAs clusters are 'transposon graveyards' and therefore reminiscent of CRISPR arrays, the archives of previous infections in prokaryote genomes (Girard and Hannon, 2008; Assis and Kondrashov, 2009; Weick and Miska, 2014). A subclass of piRNAs are rasiRNAs, repeat-associated small interfering RNAs.

piRNA-mediated epigenetic modifications can be passed on to the next generation, a process called transgenerational inheritance that has been experimentally demonstrated in *C. elegans*, *D. melanogaster* and mice (Weick and Miska, 2014). It can last for up to 60 generations in *C. elegans*. The offspring worms can remember the location of a pheromone for generations even if the pheromone is not present anymore (Sarkies and Miska, 2014; Moelling, 2017). In mice, epigenetic modulation of expression of the *Kit* gene, responsible for white stripes in the tail, can be serially transmitted by small RNAs into newborn mice (Rassoulzadegan and Cuzin, 2015).

Besides restricting TEs, it has also been shown that piRNAs can exert antiviral activity and inhibit HIV and possibly other pathogens in the germ cells, preventing their vertical transmission. Silencing of TEs seems to go all the way to silencing of the complex retrovirus HIV. This is also the case in teratocarcinomas that harbor large amounts of piRNAs. Transgenerational inheritance is reminiscent of the impact of environmental influence on genetics proposed by Lamarck, a concept that has been abandoned but now experiences a revival in the form of paramutations (Moelling, 2017).

CRISPR-Cas9

Bacteria can use the CRISPR-Cas9 defense system that is derived from previously invading phages and protects against a superinfecting new phage. This is an antiviral immune system based on the invader, where an RNase H-like molecular scissors are involved in the Cas9 protein (**Figure 4D**).

For antiviral immunity, a fragment of DNA of the first infecting phage is stored in the bacterial genome and is transcribed into messenger mRNA during a new infection. An intermediate hybrid structure is formed by the mRNA and the DNA of the newly invading phage. The DNA is cleaved by the Cas9 hybrid-specific RNase H-like activity, destroying the DNA of the new phage. The vast majority of phages have DNA genomes. In the bacterial genomes, all invading phage DNAs are stored as fragments, as an archive of the history of previous infections, the CRISPR arrays.

Interestingly, in the case of RNA phages such as MS2, a relative of Cas enzymes of a different class (Cas13a, formerly C2c2) exhibits CRISPR-Cas-like RNA-guided RNase H-like activity, as required to inactivate an RNA phage genome (Abudayyeh et al., 2016). Other Cas-related systems exist that act not only for spacer acquisition for adaptation, but also expression, target cleavage by interference, and regulatory functions (Nuñez et al., 2015; Silas et al., 2016). In addition to the numerous Cas systems a Cas10 cooperating with a ribonuclease has been described which not only destroys the DNA of the invader but also the RNA (Kazlauskiene et al., 2017), possibly improving the defense efficiency by a "back-up."

The defense strategies of CRISPR and siRNA are related, one difference lies in the storage of the genetic information of the invader. Double-stranded RNA of siRNA cannot be integrated into the host DNA as dsDNA but is either degraded or transiently stored within the proteins of RISC, but not for a new generation.

The CRISPR-Cas system most likely originated from a class of DNA transposons called casposons that rely on Cas-like activity to spread throughout prokaryotic genomes (Krupovic et al., 2017). Casposons may have contributed to the distribution of CRISPR-Cas defense system in archaea and bacteria.

It is often stressed that the CRISPR-Cas9 system is the only inheritable immune system, which does not exist in any other organism or host. One may contradict if the endogenous retroviruses are considered. They were once exogenous viruses and were endogenized and passed on for generations. Some of them can protect against superinfection by the same or related viruses. This is an inherited defense system, also reflecting the history of previous infections just like the archive described for CRISPR inserts. Archives of previous infections, fossil records, are represented by endogenous retroviruses. This has most surprisingly been demonstrated by the resurrected ERVs, which after 35 Mio years could be "repaired" from defective retrovirus inserts to an infectious one, designated as Phoenix (Dewannieux et al., 2006; Lee and Bieniasz, 2007).

What amount of phage DNA can be accumulated inside a bacterial genome? Can they become 35 Mio old or even older, as HERVs do? Can the archives be deleted, how and when? Can we learn a lesson from the ERVs? Could there be something similar to the removal of ERVs or HERVs, where shorter versions and finally solitary LTRs are left as minimal footprints? Is there something similar for phage genomes? What would minimal phage footprints look like? Are the direct repeats (DRs) with or without spacer sequences candidates resembling LTRs?

SUPERINFECTION EXCLUSION BY RETROVIRUSES AND ERVs

Retroviral infections in a mammalian or other cell types can lead to antiviral resistance due to receptor interference, which is based on the expression of a retroviral gene product, which binds to or downregulates the cell receptor for virus uptake and prevents *de novo* infection. This mechanism of superinfection exclusion resembles the effect of interference, the interfering consequence of the first viral infection against the next one. A basic viral principle can be assumed to be involved in this shut-off, the limited resources inside a cell for several simultaneous virus replications.

Retroviral endogenization and protection from superinfection was recently shown in koalas. Koalas were transferred as endangered species (by car accidents!) to an island next to the Australian mainland to be protected from going extinct, where they attracted Gibbon Ape Leukemia virus, GALV, infections and died of leukemia (Tarlinton et al., 2006). Within 100 years an antiviral resistance and survival developed as a consequence of endogenization of the retrovirus. The establishment of endogenization of a virus as the cause of resistance against infections of the same type is a more general mechanism, also known in the formation of resistance in bats against a variety of viruses (Wang et al., 2011). The gene products of the first virus in chimpanzees protects them against a novel infection (Tarlinton et al., 2006; Denner and Young, 2013) (**Figure 4E**).

Could HIV be able to endogenize and then prevent exogenous infections? For endogenization, the virus needs to infect germline cells. A recent report describes this possibility. But what required 100 years in koalas in about 20 generations would possibly require 300 to 500 years in humans (Moelling, 2017).

Also in honey bees, an endogenous Israeli Acute Paralysis Virus is known to protect against related viruses (Aswad and Katzourakis, 2012). Furthermore, Borna viruses that are replicating as well as being in the process of endogenization, may cause resistance. Also, Ebola-, Bunya- and Hantavirus-related sequences have been identified in vertebrate genomes and may protect the organisms against infection of the same virus. These viruses are single-stranded RNA viruses that would normally not integrate. Yet they were determined as endogenous viruses, indicating illegitimate reverse transcription and DNA integration. They must have been reverse transcribed by a foreign RT of LINE or other retroviral elements (Belyi et al., 2010a). In total, 10 types of incoming non-retroviral RNAs must have been illegitimately reverse transcribed and integrated into host genomes (Belyi et al., 2010a,b; Aswad and Katzourakis, 2012). Endogenous Bornavirus sequences express not only viral NC proteins in human cells but also polymerase and glycoproteins, which may protect humans from infection, conferring immunity against related viruses. In contrast, horses that lack endogenous Borna viruses, more frequently suffer from Borna disease that includes symptoms of depression. DNA viruses such as circoviruses may show similar modes of protection (Belyi et al., 2010a). Such viral archives also exist in prophages, where fragments of phage and plasmid DNA are integrated and inherited as spacers present in CRISPR arrays.

As mentioned, syncytins show sequence homology to the transmembrane protein gp41 of extant retroviruses as well as HIV. The transmembrane protein causes immune suppression in the mother to prevent an immunological rejection of her embryo. Syncytins are related to the retroviral Env proteins causing immune suppression there also. This is one of the most surprising examples of how a retrovirus shaped the human genome. The effect is also observed in other mammalian species such as cows and the syncytin genes have been acquired by independent retroviral endogenization events in different mammalian lineages (Imakawa et al., 2015). There are many host restriction factors against retroviruses developed by the hosts, many of them are not of direct viral origin but are cellular antiviral factors.

We have analyzed a HERV family member belonging to the mouse mammary tumor virus related family HERV-K (HML-10) (Broecker et al., 2016a). It has integrated into the human genome about 35 Mio years ago. We demonstrated that one of the HERVs expressed a transcript in antisense orientation to a transcript of a cellular pro-apoptotic gene, resulting in antisense inhibition of an apoptotic gene leading to cell survival and a malignant phenotype (**Figure 4E**). It is surprising that suppression of apoptosis was detected with HERV, because cell survival guarantees higher viral progeny, yet there is no progeny, as the open reading frames of this HERV have been inactivated by mutations. The anti-apoptotic effect may therefore be a relic from former days of replication competence 35 Mio years ago. The LTR promoters of this HERV family are cytokine-regulated and highly variable with

respect to orientation and expression levels, making it difficult to predict something about their general role in human cancer or other diseases. Recently, the Env protein of HERV-K was shown to inhibit HIV infection *in vitro*, which suggests that superinfection exclusion does not even affect the identical species only but also others if related enough (Terry et al., 2017).

Superinfection exclusion mediated by the expression of ERVs may constitute a simple form of inheritable immune system in eukaryotes. An antisense transcript of the ERV can be generated, for instance, if the ERV integrates in opposite orientation into the intron of a host gene. Indeed, the opposite orientation is usually favored for HERVs that integrate into introns. We and others have shown that HERV-originating transcripts that are opposite to intron sequences, can downregulate the expression of host genes *in cis*, and suppression in *trans* may also occur (Gosenca et al., 2012; Broecker et al., 2016a). Similarly, transcription of cellular genes that contain intronic ERVs in opposite orientation will generate retroviral antisense transcripts that might protect against exogenous infections (Mack et al., 2004). Such an inhibitory mechanism by long non-coding RNAs may not require the RNAi machinery, since in the RNAi-deficient *S. cerevisiae*, Ty1 LTR retrotransposons are suppressed *in trans* by lncRNAs originating from antisense promoters within Ty1 elements (Harrison et al., 2009).

Endogenous retrovirus-mediated superinfection exclusion can also be achieved through expression of retroviral proteins. In mice, a genetic factor puzzled retrovirologists for decades, the Fv1 in mice, which leads to genetic resistance against retrovirus infections and is associated with expression of a Gag-like protein of a mouse retrovirus. Similarly, Fv4 expresses an Env-like protein (Aswad and Katzourakis, 2012). Fv1 likely interacts with the pre-integration complex of MuLV, preventing genomic integration, while Fv4 acts via receptor interference to inhibit viral entry (**Figure 4E**). Receptor internalization is another mode of defense to prevent entry by a competitor, which is sometimes only transient until the cell has recovered.

Thus, retroviruses protect a host cell from another retrovirus by viral gene products, not the components of the siRNA system.

PROTEIN-BASED DEFENSE: SECRETED INTERFERON

The interferon (IFN) system is a form of protein-based immune defense with an orthologous signaling pathway as the siRNA system (**Figure 4F**). It is well accepted that the RNA world preceded the protein world and the IFN system may have been a later achievement during evolution. The most striking similarity is the secretory mechanism of IFN reminiscent of siRNA, both of which are secreted from an infected cell and warn the uninfected neighboring cells by stimulating their defense system, at the expense of the primary cell that dies. Three responses can be distinguished in either the siRNA or IFN system: silencing, mRNA degradation, and inhibition of translation (Moelling and Broecker, 2015). First, gene deamination in the IFN system correlates with methylation of chromatin by repetitive associated silencing rasiRNA. Secondly, dsRNA is detected by oligoadenylate synthetase (OAS), causing synthesis of 2′-5′ oligoadenylates that in turn activate RNase L for viral mRNA degradation by the IFN system, equivalent to mRNA degradation by siRNA or miRNA, whereby the miRNA system also leads to inhibition of translation. Thirdly, the protein kinase R (PKR) is activated through double-stranded RNA binding, which induces autophosphorylation, leading to activation of its kinase activity PKR. This then inactivates translation initiation factor eIF2a by phosphorylation, leading to inhibition of translation in the interferon system. More details have been published previously (Moelling and Broecker, 2015).

The immune systems are related, one based on nucleic acids, the other one on proteins. The innate IFN system in eukaryotes is based on proteins and is sequence independent. However, its mechanism closely resembles the sequence specific siRNA system in most steps. It appears that the RNA has evolved toward a protein-based mechanism in an orthologous fashion with similar steps to fulfill similar functions.

The IFN system recognizes dsRNA molecules through RIG-I like receptors, inducing a signaling cascade that leads to the expression of interferon-stimulated genes (ISGs) that induce an antiviral state in the cell. RIG-I is structurally related to Dicer, and similar to small RNAs in plants and *C. elegans*, IFNs are secreted. Although no RNase H molecule is involved in IFN signaling, there are other RNases and striking similarities between RNAi and IFN signaling, most strikingly the secretion and warning of neighboring cells (Moelling and Broecker, 2015).

ANTIBODY DIVERSITY GENERATED BY RAG1

Another protein-based immune system is constituted by antibodies. The diversity of populations of immunoglobulins and T cell receptors is generated in many species by V(D)J recombination, by combinatorial joining of segments of coding sequences. V(D)J recombinations can lead to millions of different functional immunoglobulin and T cell receptor genes. This recombination is mediated by the RAG1 recombinase protein with an RNase H-like domain with a zinc-binding catalytic center with the conserved D, D, E, similar to transposases. The catalytic activity of RAG1 is supported by complexing to the RAG2 protein, which may also be derived from transposases but is enzymatically inactive (**Figure 4G**). The high degree of sequence diversity required for antibody-based defense is generated by the transposon-type cut-and-paste mechanisms. The RNase H-like cleavage activity cleaves and performs an additional step by closing the DNA to hairpins and releasing excised circles. V(D)J recombinations evolved from transposons and possibly exhibited RNase H-like enzyme activities in the 900 Mio years old immune system. This transposon-like diversification system of our immune system was also described in the house fly, which, however, has no adaptive immune system. Its name is Hermes and it is very mobile for innovation in the fly genome (Moelling and Broecker, 2015). It is a remarkable 'altruism' that siRNAs in plants, *C. elegans*, IFNs, and the antibodies are secreted from an infected endangered cell to protect other cells.

One can classify antiviral systems by stating that invading RNA is counteracted by siRNA while invading DNA is defended by the CRISPR-Cas systems in bacteria. But other mechanisms: there are many more defense mechanisms in virus-infected cells, not only viral-coded gene products against viruses but also numerous cellular restriction "factors."

CONCLUSION

We are describing the importance and wide-spread distribution of the RNase H-like family members. They are among the most abundant molecules on our planet and present in all forms of life. This study contributes the identification of highly conserved RNase H-like proteins in a variety of marine samples. Indeed, also the RT and related other nucleic acid synthesizing enzymes are similarly abundant. Whenever nucleic acids are synthesized, also removal mechanisms must exist. Thus, RTs and RNases H may have cooperated throughout evolution. With RNA as the primary molecule in evolution it is not surprising, that RNA-degrading molecules arose for defense, ranging from ribozymes/viroids to RNases H. The role of RNases H-like molecules in antiviral defense was stressed here, among others by the surprising evidences in an earlier study, that the components of retroviruses and siRNA-mediated antiviral defense are orthologs, similar in structure and function.

The origin of the protein enzymes is not known, yet one might conceive that catalytic RNAs such as ribozymes and the viroids may be the RNA precursors to RNase H-related proteins. They are of universal usefulness due to their lack of specificity. This is compensated for by being a team-player with other factors, such as RNAs or proteins, to gain specificity. The RT may have evolved from simpler polymerizing structures, which may have some evolutionary connection to the rather unexplored msRNA/DNA elements. Retroelements are among the oldest ones as drivers of evolution and genome diversity. The relationship between so diverse species as bacteria and mammals was stressed here by comparing the CRISPR-spacers and the HERV sequences in the genomes as archives from earlier phage and viral infections.

Understanding the role of PIWI/RNase H in transgenerational inheritance will be fascinating.

AUTHOR CONTRIBUTIONS

KM has conducted research on the RNase H for 50 years, discovered the retroviral RNase H, drafted the concept of this article and wrote the majority of the text. FB has 10 years of experience in the research field, designed most of the figures and made significant contributions to the concept, bioinformatics and text. GR performed multiple sequence alignments and generated the phylogenetic tree of RNase H proteins in **Figure 3** (unpublished). SS is part of the TARA Oceans project and generated the data and figure presented in **Figure 2** (unpublished). All authors read and approved the final version of the manuscript.

FUNDING

This work was partially funded by the German National Academy of Sciences Leopoldina to FB. SS is supported by the Helmut Horten Foundation.

ACKNOWLEDGMENTS

KM and FB thank Prof. Peter Palese (Icahn School of Medicine at Mount Sinai) for his generous support. FB acknowledges financial support by the German National Academy of Sciences Leopoldina through a postdoctoral stipend. We would like to acknowledge Dr. Miguel Cuenca for his assistance in producing **Figure 2**, and the work of the *Tara* Oceans consortium.

REFERENCES

Abelson, J. (2013). Toggling in the spliceosome. *Nat. Struct. Mol. Biol.* 20, 645–647. doi: 10.1038/nsmb.2603

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353:aaf5573. doi: 10.1126/science.aaf5573

Artymiuk, P. J., Grindley, H. M., Kumar, K., Rice, D. W., and Willett, P. (1993). Three-dimensional structural resemblance between the ribonuclease H and connection domains of HIV reverse transcriptase and the ATPase fold revealed using graph theoretical techniques. *FEBS Lett.* 324, 15–21. doi: 10.1016/0014-5793(93)81523-3

Assis, R., and Kondrashov, A. S. (2009). Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7079–7082. doi: 10.1073/pnas.0900523106

Aswad, A., and Katzourakis, A. (2012). Paleovirology and virally derived immunity. *Trends Ecol. Evol.* 27, 627–636. doi: 10.1016/j.tree.2012.07.007

Aziz, R. K., Breitbart, M., and Edwards, R. A. (2010). Transposases are the most ubiquitous genes in nature. *Nucleic Acids Res.* 38, 4207–4217. doi: 10.1093/nar/gkq140

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226, 1209–1211. doi: 10.1038/2261209a0

Bassing, C. H., Swat, W., and Alt, F. W. (2002). The mechanism and regulation of chromosomal V(D)J recombination. *Cell* 109(Suppl.), S45–S55. doi: 10.1016/S0092-8674(02)00675-X

Baucom, R. S., Estill, J. C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J. M., et al. (2009). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLOS Genet.* 5:e1000732. doi: 10.1371/journal.pgen.1000732

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010a). Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLOS Pathog.* 6:e1001030. doi: 10.1371/journal.ppat.1001030

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010b). Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J. Virol.* 84, 12458–12462. doi: 10.1128/JVI.01789-10

Boesler, B., Meier, D., Förster, K. U., Freidrich, M., Hammann, C., Sharma, C. M., et al. (2014). Argonaute proteins affect siRNA levels and accumulation of a novel extrachromosoma DNA from the *Dictyostelium* retrotransposons DIRS-1. *J. Biol. Chem.* 289, 35124–35138. doi: 10.1074/jbc.M114.612663

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L., D'Amore, R., Allen, A. M., et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491, 705–710. doi: 10.1038/nature11650

Broecker, F., Horton, R., Heinrich, J., Franz, A., Schweiger, M. R., Lehrach, H., et al. (2016a). The intron-enriched HERV-K(HML-10) family suppresses apoptosis, an indicator of malignant transformation. *Mob. DNA* 7, 25.

Broecker, F., Klumpp, J., Schuppler, M., Russo, G., Biedermann, L., Hombach, M., et al. (2016b). Long-term changes of bacterial and viral compositions in the intestine of a recovered *Clostridium difficile* patient after fecal microbiota transplantation. *Cold Spring Harb. Mol. Case Stud.* 2:a000448. doi: 10.1101/mcs.a000448

Broecker, F., Russo, G., Klumpp, J., and Moelling, K. (2017). Stable core virome despite variable microbiome after fecal transfer. *Gut Microbes* 8, 214–220. doi: 10.1080/19490976.2016.1265196

Bubeck, D., Reijns, M. A., Graham, S. C., Astell, K. R., Jones, E. Y., and Jackson, A. P. (2011). PCNA directs type 2 RNase H activity on DNA replication and repair substrates. *Nucleic Acids Res.* 39, 3652–3666. doi: 10.1093/nar/gkq980

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D., and Mittenthal, J. E. (2009). The origin, evolution and structure of the protein world. *Biochem. J.* 417, 621–637. doi: 10.1042/BJ20082063

Cech, T. R. (2000). Structural biology. The ribosome is a ribozyme. *Science* 289, 878–879.

Cerritelli, S. M., and Crouch, R. J. (2009). Ribonuclease H: the enzymes in eukaryotes. *FEBS J.* 276, 1494–1505. doi: 10.1111/j.1742-4658.2009.06908.x

Cerutti, H., and Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.* 50, 81–99. doi: 10.1007/s00294-006-0078-x

Chon, H., Vassilev, A., DePamphilis, M. L., Zhao, Y., Zhang, J., Burgers, P. M., et al. (2009). Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex. *Nucleic Acids Res.* 37, 96–110. doi: 10.1093/nar/gkn913

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18, 71–86. doi: 10.1038/nrg.2016.139

Crow, Y. J., Leitch, A., Hayward, B. E., Garner, A., Parmar, R., Griffith, E., et al. (2006). Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection. *Nat. Genet.* 38, 910–916. doi: 10.1038/ng1842

Denner, J., and Young, P. R. (2013). Koala retroviruses: characterization and impact on the life of koalas. *Retrovirology* 10:108. doi: 10.1186/1742-4690-10-108

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., et al. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* 16, 1548–1556. doi: 10.1101/gr.5565706

Dlakić, M., and Mushegian, A. (2011). Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA* 17, 799–808. doi: 10.1261/rna.2396011

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* 41, 563–571. doi: 10.1038/ng.368

Feiss, M., and Rao, V. B. (2012). The bacteriophage DNA packaging machine. *Adv. Exp. Med. Biol.* 726, 489–509. doi: 10.1007/978-1-4614-0980-9_22

Félix, M. A., Ashe, A., Piffaretti, J., Wu, G., Nuez, I., Bélicard, T., et al. (2011). Natural and experimental infection of *Caenorhabditis* nematodes by novel viruses related to nodaviruses. *PLOS Biol.* 9:e1000586. doi: 10.1371/journal.pbio.1000586

Feschotte, C., and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368. doi: 10.1146/annurev.genet.40.110405.090448

Flint, J. S., Enquist, L. W., Racaniello, V. R., Rall, G. F., and Skalka, A. M. (2015). *Principles of Virology*, 4th Edn. Washington, DC: ASM Press.

Folimonova, S. Y. (2012). Superinfection exclusion is an active virus-controlled function that requires a specific viral protein. *J. Virol.* 86, 5554–5561. doi: 10.1128/JVI.00310-12

Girard, A., and Hannon, G. J. (2008). Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol.* 18, 136–148. doi: 10.1016/j.tcb.2008.01.004

Gosenca, D., Gabriel, U., Steidler, A., Mayer, J., Diem, O., Erben, P., et al. (2012). HERV-E-mediated modulation of PLA2G4A transcription in urothelial carcinoma. *PLOS ONE* 7:e49341. doi: 10.1371/journal.pone.0049341

Grainger, R. J., and Beggs, J. D. (2005). Prp8 protein: at the heart of the spliceosome. *RNA* 11, 533–557. doi: 10.1261/rna.2220705

Guo, H., Arambula, D., Ghosh, P., and Miller, J. F. (2014). Diversity-generating retroelements in phage and bacterial genomes. *Microbiol. Spectr.* 2:MDNA3-0029-2014. doi: 10.1128/microbiolspec

Hansen, J., Schulze, T., Mellert, W., and Moelling, K. (1988). Identification and characterization of HIV-specific RNase H by monoclonal antibody. *EMBO J.* 7, 239–243.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Harrison, B. R., Yazgan, O., and Krebs, J. E. (2009). Life without RNAi: noncoding RNAs and their functions in *Saccharomyces cerevisiae*. *Biochem. Cell Biol.* 87, 767–779. doi: 10.1139/O09-043

Hegedus, K., Palkovics, L., Tóth, E. K., Dallmann, G., and Balázs, E. (2001). The DNA form of a retroviroid-like element characterized in cultivated carnation species. *J. Gen. Virol.* 82(Pt 3), 687–691. doi: 10.1099/0022-1317-82-3-687

Heras, S. R., Macias, S., Plass, M., Fernandez, N., Cano, D., Eyras, E., et al. (2013). The Microprocessor controls the activity of mammalian retrotransposons. *Nat. Struct. Mol. Biol.* 20, 1173–1181. doi: 10.1038/nsmb.2658

Hindmarsh, P., and Leis, J. (1999). Retroviral DNA integration. *Microbiol. Mol. Biol. Rev.* 63, 836–843.

Hutchison, C. A. III, and Sinsheimer, R. L. (1971). Requirement of protein synthesis for bacteriophage phi X174 superinfection exclusion. *J. Virol.* 8, 121–124.

Imakawa, K., Nakagawa, S., and Miyazawa, T. (2015). Baton pass hypothesis: successive incorporation of unconserved endogenous retroviral genes for placentation during mammalian evolution. *Genes Cells* 20, 771–788. doi: 10.1111/gtc.12278

Inouye, K., Tanimoto, S., Kamimoto, M., Shimamoto, T., and Shimamoto, T. (2011). Two novel retron elements are replaced with retron-Vc95 in *Vibrio cholerae*. *Microbiol. Immunol.* 55, 510–513. doi: 10.1111/j.1348-0421.2011.00342.x

Jiang, X. Y., Hou, F., Shen, X. D., Du, X. D., Xu, H. L., and Zou, S. M. (2016). The N-terminal zinc finger domain of Tgf2 transposase contributes to DNA binding and to transposition activity. *Sci. Rep.* 6:27101. doi: 10.1038/srep27101

Jimenez, R. M., Polanco, J. A., and Lupták, A. (2015). Chemistry and biology of self-cleaving ribozymes. *Trends Biochem. Sci.* 40, 648–661. doi: 10.1016/j.tibs.2015.09.001

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343:1247997. doi: 10.1126/science.1247997

Kapitonov, V. V., and Koonin, E. V. (2015). Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol. Direct* 10, 20. doi: 10.1186/s13062-015-0055-8

Karwan, R., and Wintersberger, U. (1986). Yeast ribonuclease H(70) cleaves RNA-DNA junctions. *FEBS Lett.* 206, 189–192. doi: 10.1016/0014-5793(86)80978-4

Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Ikehara, M., et al. (1990). Three-dimensional structure of ribonuclease H from *E. coli*. *Nature* 347, 306–309. doi: 10.1038/347306a0

Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* 357, 605–609. doi: 10.1126/science.aao0100

Koonin, E. V. (2017). Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol. Direct* 12:5. doi: 10.1186/s13062-017-0177-2

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Direct* 1, 29. doi: 10.1186/1745-6150-1-29

Kovalskaya, N., and Hammond, R. W. (2014). Molecular biology of viroid-host interactions and disease control strategies. *Plant Sci.* 228, 48–60. doi: 10.1016/j.plantsci.2014.05.006

Krupovic, M., Béguin, P., and Koonin, E. V. (2017). Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr. Opin. Microbiol.* 38, 36–43. doi: 10.1016/j.mib.2017.04.004

Kurzynska-Kokorniak, A., Jamburuthugoda, V. K., Bibillo, A., and Eickbush, T. H. (2007). DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. *J. Mol. Biol.* 374, 322–333. doi: 10.1016/j.jmb.2007.09.047

Laliberte, J. P., and Moss, B. (2014). A novel mode of poxvirus superinfection exclusion that prevents fusion of the lipid bilayers of viral and cellular membranes. *J. Virol.* 88, 9751–9768. doi: 10.1128/JVI.00816-14

Lambowitz, A. M., and Belfort, M. (2015). Mobile bacterial group II introns at the crux of eukaryotic evolution. *Microbiol. Spectr.* 3:MDNA3-0050-2014. doi: 10.1128/microbiolspec

Lambowitz, A. M., and Zimmerly, S. (2004). Mobile group II introns. *Annu. Rev. Genet.* 38, 1–35. doi: 10.1146/annurev.genet.38.072902.091600

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 412:565.

Lee, Y. N., and Bieniasz, P. D. (2007). Reconstitution of an infectious human endogenous retrovirus. *PLOS Pathog.* 3:e10. doi: 10.1371/journal.ppat.0030010

Lescot, M., Hingamp, P., Kojima, K. K., Villar, E., Romac, S., Veluchamy, A., et al. (2016). Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *ISME J.* 10, 1134–1146. doi: 10.1038/ismej.2015.192

Lincoln, T. A., and Joyce, G. F. (2009). Self-sustained replication of an RNA enzyme. *Science* 323, 1229–1232. doi: 10.1126/science.1167856

Lu, M. J., and Henning, U. (1994). Superinfection exclusion by T-even-type coliphages. *Trends Microbiol.* 2, 137–139. doi: 10.1016/0966-842X(94)90601-7

Ma, B. G., Chen, L., Ji, H. F., Chen, Z. H., Yang, F. R., Wang, L., et al. (2008). Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* 366, 607–611. doi: 10.1016/j.bbrc.2007.12.014

Mack, M., Bender, K., and Schneider, P. M. (2004). Detection of retroviral antisense transcripts and promoter activity of the HERV-K(C4) insertion in the MHC class III region. *Immunogenetics* 56, 321–332. doi: 10.1007/s00251-004-0705-y

Mackenzie, K. J., Carroll, P., Lettice, L., Tarnauskaitė, Ž., Reddy, K., Dix, F., et al. (2016). Ribonuclease H2 mutations induce a cGAS/STING-dependent innate immune response. *EMBO J.* 35, 831–844. doi: 10.15252/embj.201593339

Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., et al. (2014). The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* 42, 4160–4179. doi: 10.1093/nar/gkt1414

Malik, H. S., Burke, W. D., and Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* 16, 793–805. doi: 10.1093/oxfordjournals.molbev.a026164

Malik, H. S., and Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.* 11, 1187–1197. doi: 10.1101/gr.185101

Malone, C. D., and Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell* 136, 656–668. doi: 10.1016/j.cell.2009.01.045

Matskevich, A. A., and Moelling, K. (2007). Dicer is involved in protection against influenza A virus infection. *J. Gen. Virol.* 88(Pt 10), 2627–2635. doi: 10.1099/vir.0.83103-0

Maumus, F., Epert, A., Nogué, F., and Blanc, G. (2014). Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* 5:4268. doi: 10.1038/ncomms5268

Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., et al. (2017). Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. *Proc. Natl. Acad. Sci. U.S.A.* 114, 4739–4744. doi: 10.1073/pnas.1701462114

McClintock, B. (1951). Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* 16, 13–47. doi: 10.1101/SQB.1951.016.01.004

Miller, J. L., Le Coq, J., Hodes, A., Barbalat, R., Miller, J. F., and Ghosh, P. (2008). Selective ligand recognition by a diversity-generating retroelement variable protein. *PLOS Biol.* 6:e131. doi: 10.1371/journal.pbio.0060131

Moelling, K. (2012). Are viruses our oldest ancestors? *EMBO Rep.* 13, 1033. doi: 10.1038/embor.2012.173

Moelling, K. (2013). What contemporary viruses tell us about evolution: a personal view. *Arch. Virol.* 158, 1833–1848. doi: 10.1007/s00705-013-1679-6

Moelling, K. (2017). *Viruses More Friends than Foes*. Singapore: World Scientific Press.

Moelling, K., and Broecker, F. (2015). The reverse transcriptase-RNase H: from viruses to antiviral defense. *Ann. N. Y. Acad. Sci.* 1341, 126–135. doi: 10.1111/nyas.12668

Moelling, K., Matskevich, A., and Jung, J. S. (2006). Relationship between retroviral replication and RNA interference machineries. *Cold Spring Harb. Symp. Quant. Biol.* 71, 365–368. doi: 10.1101/sqb.2006.71.010

Mölling, K., Bolognesi, D. P., Bauer, H., Büsen, W., Plassmann, H. W., and Hausen, P. (1971). Association of viral reverse transcriptase with an enzyme degrading the RNA moiety of RNA-DNA hybrids. *Nat. New Biol.* 234, 240–243. doi: 10.1038/newbio234240a0

Müller, G., Strack, B., Dannull, J., Sproat, B. S., Surovoy, A., Jung, G., et al. (1994). Amino acid requirements of the nucleocapsid protein of HIV-1 for increasing catalytic activity of a Ki-ras ribozyme in vitro. *J. Mol. Biol.* 242, 422–429. doi: 10.1006/jmbi.1994.1592

Nethe, M., Berkhout, B., and van der Kuyl, A. C. (2005). Retroviral superinfection resistance. *Retrovirology* 2:52. doi: 10.1186/1742-4690-2-52

Nguyen, T. H., Galej, W. P., Bai, X. C., Savva, C. G., Newman, A. J., Scheres, S. H., et al. (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* 523, 47–52. doi: 10.1038/nature14548

Niblett, C. L., Dickson, E., Fernow, K. H., Horst, R. K., and Zaitlin, M. (1978). Cross protection among four viroids. *Virology* 91, 198–203. doi: 10.1016/0042-6822(78)90368-9

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., et al. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156, 935–949. doi: 10.1016/j.cell.2014.02.001

Nowotny, M., Gaidamakov, S. A., Crouch, R. J., and Yang, W. (2005). Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* 121, 1005–1016. doi: 10.1016/j.cell.2005.04.024

Nuñez, J. K., Lee, A. S., Engelman, A., and Doudna, J. A. (2015). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519, 193–198. doi: 10.1038/nature14237

Ohtani, N., Yanagawa, H., Tomita, M., and Itaya, M. (2004). Cleavage of double-stranded RNA by RNase HI from a thermoacidophilic archaeon, *Sulfolobus tokodaii* 7. *Nucleic Acids Res.* 32, 5809–5819. doi: 10.1093/nar/gkh917

Oshima, J. (2000). The Werner syndrome protein: an update. *Bioessays* 22, 894–901. doi: 10.1002/1521-1878(200010)22:10<894::AID-BIES4>3.0.CO;2-B

Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D., and Sutherland, J. D. (2015). Common origin of RNA, protein and lipid precursors in a cyanosulfidic protometabolis. *Nat. Chem.* 7, 301–307. doi: 10.1038/nchem.2202

Qin, C., Wang, Z., Shang, J., Bekkari, K., Liu, R., Pacchione, S., et al. (2010). Intracisternal A particle genes: distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. *Mol. Carcinog.* 49, 54–67. doi: 10.1002/mc.20576

Qin, S., Jin, P., Zhou, X., Chen, L., and Ma, F. (2015). The role of transposable elements in the origin and evolution of MicroRNAs in human. *PLOS ONE* 10:e0131365. doi: 10.1371/journal.pone.0131365

Rassoulzadegan, M., and Cuzin, F. (2015). Epigenetic heredity: RNA-mediated modes of phenotypic variation. *Ann. N. Y. Acad. Sci.* 1341, 172–175. doi: 10.1111/nyas.12694

Rice, P. A., and Baker, T. A. (2001). Comparative architecture of transposase and integrase complexes. *Nat. Struct. Biol.* 8, 302–307.

Roberts, J. T., Cardin, S. E., and Borchert, G. M. (2014). Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mob. Genet. Elements* 4:e29255. doi: 10.4161/mge.29255

Salaman, R. N. (1933). Protective inoculation against a plant virus. *Nature* 131, 468. doi: 10.1038/131468a0

Sarafianos, S. G., Das, K., Tantillo, C., Clark, A. D. Jr., Ding, J., Whitcomb, J. M., et al. (2001). Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J.* 20, 1449–1461. doi: 10.1093/emboj/20.6.1449

Sarkies, P., and Miska, E. A. (2014). Small RNAs break out: the molecular cell biology of mobile small RNAs. *Nat. Rev. Mol. Cell Biol.* 15, 525–535. doi: 10.1038/nrm3840

Schaller, T., Appel, N., Koutsoudakis, G., Kallis, S., Lohmann, V., Pietschmann, T., et al. (2007). Analysis of hepatitis C virus superinfection exclusion by using novel fluorochrome gene-tagged viral genomes. *J. Virol.* 81, 4591–4603. doi: 10.1128/JVI.02144-06

Silas, S., Mohr, G., Sidote, D. J., Markham, L. M., Sanchez-Amat, A., Bhaya, D., et al. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* 351:aad4234. doi: 10.1126/science.aad4234

Simon, D. M., and Zimmerly, S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res.* 36, 7219–7229. doi: 10.1093/nar/gkn867

Smith, D., Zhong, J., Matsuura, M., Lambowitz, A. M., and Belfort, M. (2005). Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes Dev.* 19, 2477–2487. doi: 10.1101/gad.1345105

Sollier, J., and Cimprich, K. A. (2015). R-loops breaking bad. *Trends Cell Biol.* 25, 514–522. doi: 10.1016/j.tcb.2015.05.003

Song, J. J., Smith, S. K., Hannon, G. J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* 305, 1434–1437. doi: 10.1126/science.1102514

Stein, H., and Hausen, P. (1969). Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase. *Science* 166, 393–395. doi: 10.1126/science.166.3903.393

Sulej, A. A., Tuszynska, I., Skowronek, K. J., Nowotny, M., and Bujnicki, J. M. (2012). Sequence-specific cleavage of the RNA strand in DNA-RNA hybrids by the fusion of ribonuclease H with a zinc finger. *Nucleic Acids Res.* 40, 11563–11570. doi: 10.1093/nar/gks885

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348:1261359. doi: 10.1126/science.1261359

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693

Swarts, D. C., Jore, M. M., Westra, E. R., Zhu, Y., Janssen, J. H., Snijders, A. P., et al. (2014). DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* 507, 258–261. doi: 10.1038/nature12971

Tadokoro, T., and Kanaya, S. (2009). Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes. *FEBS J.* 276, 1482–1493. doi: 10.1111/j.1742-4658.2009.06907.x

Tarlinton, R. E., Meers, J., and Young, P. R. (2006). Retroviral invasion of the koala genome. *Nature* 442, 79–81. doi: 10.1038/nature04841

Taylor, J. M. (2015). Hepatitis D virus replication. *Cold Spring Harb. Perspect. Med.* 5:a021568. doi: 10.1101/cshperspect.a021568

Temin, H. M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of *Rous sarcoma virus*. *Nature* 226, 1211–1213. doi: 10.1038/2261211a0

tenOever, B. R. (2017). Questioning antiviral RNAi in mammals. *Nat. Microbiol.* 2:17052. doi: 10.1038/nmicrobiol.2017.52

Terry, S. N., Manganaro, L., Cuesta-Dominguez, A., Brinzevich, D., Simon, V., and Mulder, L. C. F. (2017). Expression of HERV-K108 envelope interferes with HIV production. *Virology* 509, 52–59. doi: 10.1016/j.virol.2017.06.004

Tisdale, M., Schulze, T., Larder, B. A., and Moelling, K. (1991). Mutations within the RNase H domain of human immunodeficiency virus type 1 reverse transcriptase abolish virus infectivity. *J. Gen. Virol.* 72, 59–66. doi: 10.1099/0022-1317-72-1-59

Ustyantsev, K., Novikova, O., Blinov, A., and Smyshlyaev, G. (2015). Convergent evolution of ribonuclease h in LTR retrotransposons and retroviruses. *Mol. Biol. Evol.* 32, 1197–1207. doi: 10.1093/molbev/msv008

Wang, L. F., Walker, P. J., and Loon, L. L. (2011). Mass extinctions, biodiversity and mitochondrial function: are bats "special" as reservoirs for emerging viruses? *Curr. Opin. Virol.* 1, 649–657. doi: 10.1016/j.coviro.2011.10.013

Weick, E. M., and Miska, E. A. (2014). piRNAs: from biogenesis to function. *Development* 141, 3458–3471. doi: 10.1242/dev.094037

Will, C. L., and Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* 3:a003707. doi: 10.1101/cshperspect.a003707

Wilson, R. C., and Doudna, J. A. (2013). Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.* 42, 217–239. doi: 10.1146/annurev-biophys-083012-130404

Wilson, W. H., Gilg, I. C., Moniruzzaman, M., Field, E. K., Koren, S., LeCleir, G. R., et al. (2017). Genomic exploration of individual giant ocean viruses. *ISME J.* 11, 1736–1745. doi: 10.1038/ismej.2017.61

Yang, W., Hendrickson, W. A., Crouch, R. J., and Satow, Y. (1990). Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein. *Science* 249, 1398–1405. doi: 10.1126/science.2169648

Yurchenko, V., Xue, Z., and Sadofsky, M. (2003). The RAG1 N-terminal domain is an E3 ubiquitin ligase. *Genes Dev.* 17, 581–585. doi: 10.1101/gad.1058103

Zimmerly, S., and Semper, C. (2015). Evolution of group II introns. *Mob. DNA* 6, 7. doi: 10.1186/s13100-015-0037-5

Zuo, Y., and Deutscher, M. P. (2001). Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. *Nucleic Acids Res.* 29, 1017–1026. doi: 10.1093/nar/29.5.1017

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 01 February 2018
doi: 10.3389/fmicb.2018.00101

Stem-Loop RNA Hairpins in Giant Viruses: Invading rRNA-Like Repeats and a Template Free RNA

Hervé Seligmann and Didier Raoult*

Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UMR MEPHI, Aix-Marseille Université, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée-Infection, Marseille, France

We examine the hypothesis that *de novo* template-free RNAs still form spontaneously, as they did at the origins of life, invade modern genomes, contribute new genetic material. Previously, analyses of RNA secondary structures suggested that some RNAs resembling ancestral (t)RNAs formed recently *de novo*, other parasitic sequences cluster with rRNAs. Here positive control analyses of additional RNA secondary structures confirm ancestral and *de novo* statuses of RNA grouped according to secondary structure. Viroids with branched stems resemble *de novo* RNAs, rod-shaped viroids resemble rRNA secondary structures, independently of GC contents. 5′ UTR leading regions of West Nile and Dengue flavivirid viruses resemble *de novo* and rRNA structures, respectively. An RNA homologous with Megavirus, Dengue and West Nile genomes, copperhead snake microsatellites and levant cotton repeats, not templated by Mimivirus' genome, persists throughout Mimivirus' infection. Its secondary structure clusters with candidate *de novo* RNAs. The saltatory phyletic distribution and secondary structure of Mimivirus' peculiar RNA suggest occasional template-free polymerization of this sequence, rather than noncanonical transcriptions (swinger polymerization, posttranscriptional editing).

Keywords: systematic nucleotide exchange, swinger DNA polymerization, invertase, 3′-to-5′ polymerization, transcription, *Acanthamoeba castellanii*

OPEN ACCESS

Edited by:
Guenther Witzany, Independent Researcher, Salzburg, Austria

Reviewed by:
Cristina Romero-López, Institute of Parasitology and Biomedicine "López-Neyra" (CSIC), Spain
Wenbing Zhang, Wuhan University, China

****Correspondence:***
Hervé Seligmann
podarcissicula@gmail.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *18 August 2017*
Accepted: *16 January 2018*
Published: *01 February 2018*

Citation:
Seligmann H and Raoult D (2018) Stem-Loop RNA Hairpins in Giant Viruses: Invading rRNA-Like Repeats and a Template Free RNA. Front. Microbiol. 9:101. doi: 10.3389/fmicb.2018.00101

INTRODUCTION

Diverse numbers of simple organic compounds spontaneously self-organized at life's origins. This system crystallized around the ribonucleic-protein system forming the main organizational building blocks of known organisms (Szostak, 2009; Ruiz-Mirazo et al., 2014). Then presumably the tRNA-rRNA information-storage/translation apparatus developed (Fox, 2010; Root-Bernstein and Root-Bernstein, 2015, 2016) through segment accretion (Di Giulio, 1992, 1994, 1995, 1999, 2008, 2009, 2012, 2013; Widmann et al., 2005; Branciamore and Di Giulio, 2011, 2012; Seligmann, 2014; Petrov et al., 2015). Presumably, self-replicating systems evolved, producing/parasitized by molecules lacking self-replication capacities (Bansho et al., 2012). The system potentially stabilized by evolving molecular cooperation between molecules with replicating capacities (Penny, 2015) and others lacking this capacity but contributing otherwise to the system's persistence (Higgs and Lehman, 2015). This process would have produced the modern translation/replication system(s).

Other molecules (short parasitic repeats, frequently forming stem-loop hairpins, viroids, viruses, etc.) presumably subsisted mainly as parasites and occasionally contributing new, sometimes functional parts. Persistence of the cooperative system implies integrating new molecules with new functions, while channeling most resources to critical components such as ribosomes.

Nowadays, ribosomes compete with parasitic elements that hijack the cell's integrated cooperative system (Xie and Scully, 2017). Viruses frequently mimic cellular processes (Hiscox, 2007), including the cell's replication/transcription compartments (Chaikeeratisak et al., 2017). Hence when life began, ribosomal RNAs had virus-like properties. This arms race might explain why >95% of the cell's transcriptome consists of ribosomal RNA (Peano et al., 2013). Here we hypothesize that new RNAs still spontaneously emerge and integrate molecular cooperative systems of organisms.

We use several analyses, including comparisons among RNA secondary structures, to detect and test for *de novo* RNA emergence.

Structural Homology

Classical sequence homology between linear sequences is inefficient at reconstructing ancient evolution because sequences evolve relatively fast. Structures, rather than sequences, are conserved for longer periods. For example, analyses considering structural homology among proteins suggest a common cellular ancestry for modern cells and viruses (Nasir and Caetano-Anollés, 2015). Similarly, analyses using simple properties of secondary structures formed by diverse RNAs detect two main clusters (A1 and A2, small vs. complex RNAs), each subdivided into two main groups (A1:B1-B2; A2:D1-D2), schematized in **Figure 1**.

Cluster B1 is the functionally most diverse group and includes several presumably ancestral RNAs, such as replication origin, tRNA and some ribozymes. Its "sister" cluster B2 includes diverse, probably more derived/recent molecules (i.e., the only known protein-encoding viroid, AbouHaidar et al., 2014). Clusters D1 and D2 are characterized by rRNA subunits, associated with parasitic RNAs: D1 includes all six 23S rRNA subunits and retroviruses; and D2 groups all 16S rRNA subunits and most families of RPEs, rickettsial palindromic elements, which infest specifically *Rickettsia* genomes (Amiri et al., 2002; Gillespie et al., 2012). Secondary structure similarities between parasitic and ribosomal RNAs underscore virus-like rRNA properties (**Figure 1**), presumably due to the assumed arms race between rRNAs and parasitic RNAs.

Cluster D2 includes half of the secondary structures formed by tRNA sequences. Hence B1–D2 would represent organic life's main tRNA-rRNA axis of molecular evolution, where simple ancestral tRNA-like RNAs complexified into rRNA-like RNAs (Bloch et al., 1983, 1984, 1989). This interpretation is in line with detections of candidate tRNA genes within mitochondrial 16S rRNA of chaetognath mitogenomes that otherwise would lack tRNAs (Barthélémy and Seligmann, 2016).

Here we analyze additional types of RNAs and explore their similarities with clusters B1-2/D1-2. Additional viroids, are analyzed to explore the possibility that some viroids date from the precellular world (Bussière et al., 1995; Diener, 1996) and others emerged *de novo* recently (Koonin and Dolja, 2013; Seligmann and Raoult, 2016), potentially solving the conundrum about primordial/recent *de novo* viroid origins (Diener, 2016). Two RNA types function as controls to confirm the statuses of

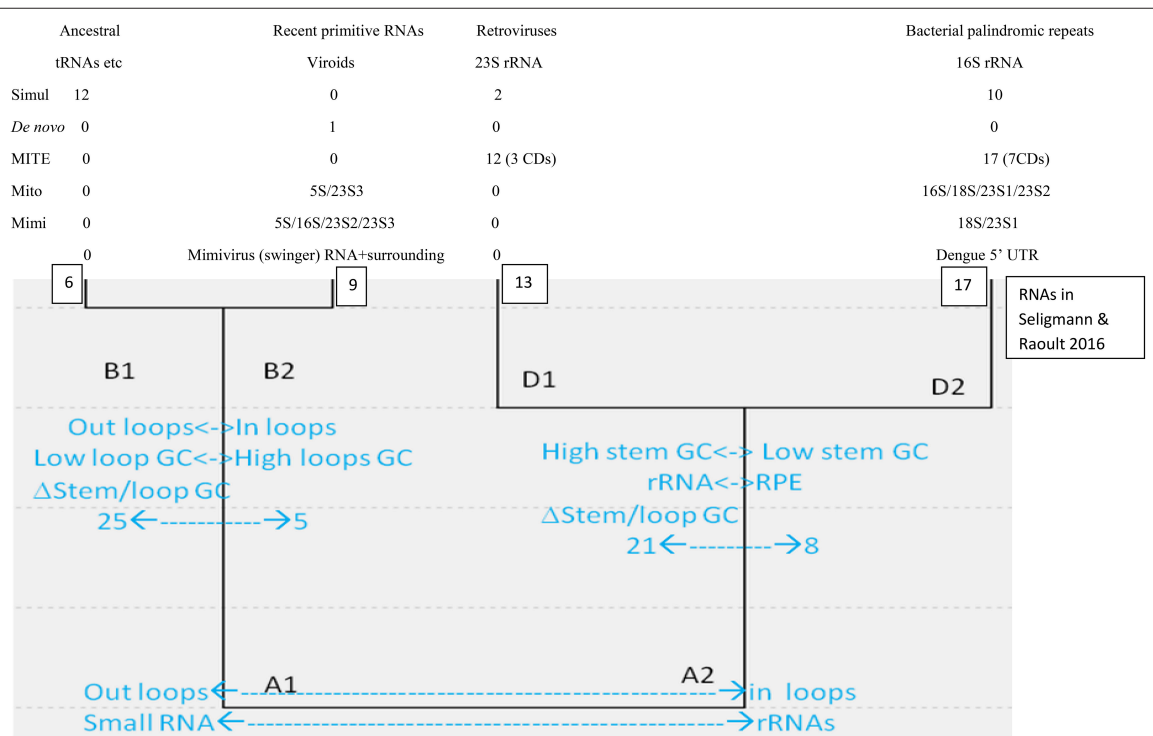


FIGURE 1 | Hierarchical cluster of secondary structures formed by diverse RNA molecules, adapted from Seligmann and Raoult (2016).

putative ancestral/*de novo* clusters B1 and B2. The remaining RNAs originate from giant viruses and test putative evolutionary links between giant viruses and rRNAs.

MATERIALS AND METHODS

Secondary Structure Predictions, Properties, and Comparisons

Secondary structure predictions follow previous analyses (Seligmann and Raoult, 2016). Four variables are estimated from the optimal secondary structure predicted by Mfold (Zuker, 2003): 1. overall nucleotide percentage not involved in self-hybridization (loops); 2. percentage of nucleotides in closed loops at stem extremity among all nucleotides in loops (external loops); and 3. GC contents in loops, and 4. in stems. RNA size is not included among these variables.

Table 1 presents these variables for sequences analyzed here. They are compared with corresponding data from a previous collection of RNA sequences (Seligmann and Raoult, 2016, therein **Table 1** and **Figure 2**) that defined clusters in **Figure 1**. Comparisons use Pearson's correlation coefficient r as a similarity estimate, setting statistically significant similarity at $r = 0.95$ (one tailed $P = 0.05$). Correlation analyses plot each of the four variables 1–4 of Y for one of the new RNAs analyzed here as a function of corresponding variables 1–4 of X for each of previously analyzed RNAs (Seligmann and Raoult, 2016, therein Table 1). Correlation coefficients estimate similarities between X and Y according to variables 1–4 (example in **Figure 2**). Results of statistical analyses presented here are validated also by the Benjamini-Hochberg correction method for false discovery rates which accounts for multiple tests (Benjamini and Hochberg, 1995; methodology detailed for unrelated analyses in Seligmann and Warthi, 2017). This test, unlike the classical Bonferroni approach that minimizes false positive detection rates and misses numerous positive results (Perneger, 1998) optimizes between false negative and false positive detection rates (Käll et al., 2007). The method is not affected by lack of independence between observations, and accounts for multiple testing. These additional analyses confirm classical statistics and therefore are not detailed in Results.

Mimivirus' RNA

Mimivirus' transcriptome (public data available at http://sra.dnanexus.com/studies/SRP001690/experiments; Legendre et al., 2010) are analyzed using CLCgenomicswb7. Reads are mapped on Mimivirus' reference genome NC_014649, according to the following criteria: at least half of the read maps to the reference genome, with at least 80% identity.

RESULTS

Simulation-Generated Palindromes

Previous analyses produced the classification scheme of secondary structures formed by RNAs in **Figure 1**. Secondary structures formed by selected RNAs are analyzed to test this scheme. We classified the secondary structures of 24 sequences produced *in silico* by Demongeot and Moreira (2007). These

TABLE 1 | Secondary structure properties of sequences analyzed here.

	N	Loop	eLoop	Stem GC	Loop GC	Cluster
Circ1	22	36.36	62.50	36.36	27.27	B1
Circ2	22	27.27	83.33	27.27	27.27	B1
Circ3	22	59.09	38.46	59.09	27.27	D2
Circ4	22	54.55	41.67	54.55	27.27	D2
Circ5	22	63.64	35.71	63.64	27.27	D2
Circ6	22	54.55	35.71	63.64	27.27	D2
Circ7	22	72.73	25.00	54.55	13.64	D2
Circ8	22	54.55	31.25	72.73	27.27	D1
Circ9	22	45.46	41.67	54.55	27.27	D2
Circ10	22	18.18	30.00	45.46	13.64	D2
Circ11	22	18.18	75.00	18.18	13.64	B1
Circ12	22	63.64	75.00	18.18	13.64	B1
Circ13	22	54.55	35.71	63.64	27.27	D2
Circ14	22	27.27	25.00	54.55	13.64	D2
Circ15	22	18.18	50.00	27.27	13.64	B1
Circ16	22	18.18	75.00	18.18	13.64	B1
Circ17	22	54.55	75.00	18.18	13.64	B1
Circ18	22	63.64	25.00	54.55	13.64	D2
Circ19	22	18.18	21.43	63.64	13.64	D1
Circ20	22	27.27	75.00	18.18	13.64	B1
Circ21	22	27.27	50.00	27.27	13.64	B1
Circ22	22	27.27	50.00	27.27	13.64	B1
Circ23	22	27.27	83.33	27.27	27.27	B1
Circ24	22	27.27	83.33	27.27	27.27	B1
De novo	51	16.67	40.00	56.00	60.00	B2
MITE16*	48	70.83	23.53	78.57	38.24	D1
MITE20*	66	48.48	28.13	58.82	28.13	D1
MITE4*	27	85.19	17.39	50.00	43.48	D2
MITE21*	60	85.00	54.90	90.00	35.29	D2
MITE14*	60	66.67	32.50	70.00	32.50	D2
MITE8*	57	68.42	53.85	100.00	30.77	D2
MITE2*	43	67.44	17.24	64.29	34.48	D2
MITE25*	62	54.84	38.24	64.29	20.59	D2
MITE9*	66	54.55	36.11	80.77	27.78	D1
MITE7*	57	68.42	30.77	50.00	33.33	D2
MITE12	68	64.71	22.73	87.50	31.82	D1
MITE29	62	70.97	34.09	94.44	27.27	D1
MITE28	71	61.97	20.45	84.62	31.82	D1
MITE24	64	59.38	23.68	50.00	28.95	D2
MITE19	67	71.64	22.92	80.00	35.42	D1
MITE17	65	73.85	29.17	87.50	37.50	D1
MITE10	72	52.78	47.37	82.35	36.84	D1
MITE26	62	67.74	26.19	70.00	26.19	D2
MITE13	63	68.25	76.74	95.00	27.91	D2
MITE22	62	64.52	32.50	75.00	35.00	D1
MITE5	45	60.00	37.04	68.75	33.33	D2
MITE11	41	56.10	26.09	88.89	21.74	D1
MITE23	62	70.97	34.09	83.33	34.09	D1
MITE1	61	60.66	37.84	58.33	37.84	D2
MITE15	61	67.21	31.71	75.00	29.27	D2
MITE6	36	61.11	36.36	57.14	40.91	D2

(Continued)

TABLE 1 | Continued

	N	Loop	eLoop	Stem GC	Loop GC	Cluster
MITE3	57	63.16	22.22	54.55	33.33	D2
MITE18	62	61.29	42.11	75.00	26.32	D2
MITE27	62	64.52	25.00	60.00	32.50	D2
Mito 5S	44	36.36	37.50	14.29	0.00	B2
Mito 16S	46	52.17	25.00	4.46	29.17	D2
Mito 18S	59	66.10	28.21	45.00	20.51	D2
Mito 23S1	375	46.93	16.48	14.82	20.46	D2
Mito 23S2	276	43.12	19.33	29.30	23.53	D2
Mito 23S3	209	37.32	24.36	23.19	19.23	B2
Mimi 5S	41	75.61	25.81	10.00	16.13	B2
Mimi 16S	39	45.76	33.33	16.17	14.82	B2
Mimi 18S	59	52.54	12.90	17.86	12.90	D2
Mimi 23S1	380	48.95	19.36	11.60	19.36	D2
Mimi 23S2	349	39.26	23.36	21.91	25.55	B2
Mimi 23S3	207	34.78	22.22	18.06	12.50	B2
WNV 3′-5	105	20.00	19.05	54.02	33.33	D1
DENV 3′-5	105	27.62	24.14	50.00	27.59	D1
JEV 3′-5	111	23.13	13.79	51.22	27.59	D1
YFV 3′-5	107	23.37	20.00	57.32	36.00	D1
5′-UTR Dengue	123	49.59	34.43	26.02	45.90	D2
Surrounding	92	52.17	43.75	30.68	31.25	B2

Variables are: N- sequence length (not used for classifying RNAs in further analyses); Loop-percentage of nucleotides not involved in self-hybridization; eLoop-percentage of nucleotides among those in loops, in closed loops at stem extremity; percentage of GC in stems, and loops. "Cluster" indicates the cluster in ***Figure 1*** *with the highest similarity in secondary structure properties. "Circ" indicates sequences generated by simulations attempting to mimic circular RNA genesis (Demongeot and Moreira, 2007); "De novo" is a template-free synthesized sequence (Béguin et al., 2015); MITE1-29 are Pandoravirus'miniature inverted-repeat transposable elements from Submariner family (Sun et al., 2015), * indicates MITE inserted in a protein coding gene; Mito and Mimi are amoeban rRNA sequences aligning with Mimivirus sequences (alignments described in* ***Table 3****);" 5′-UTR" leader sequence of Dengue and West Nile virus (Gale et al., 2000); "Swinger alone" is the part of the 5′-UTR leader detected in Mimivirus' transcriptome, not templated by its genome; "surrounding" integrates the former swinger sequence with its untransformed surrounding Mimivirus sequences.*

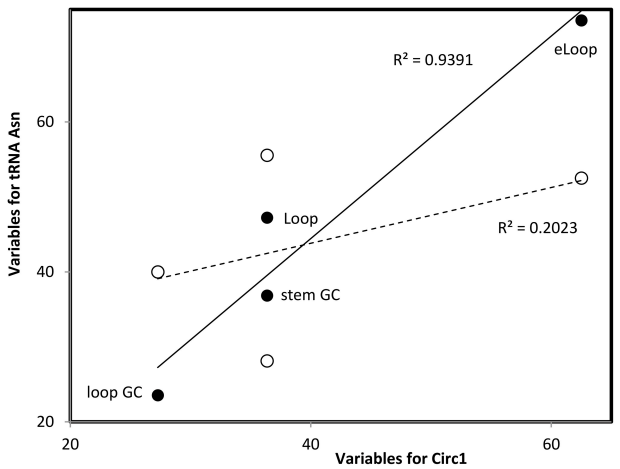


FIGURE 2 | Multivariate comparisons between estimates of secondary structure variables for theoretical sequence Circ1 (from **Table 1**, x-axis) with corresponding variables for the cloverleaf (filled symbols, continous line) and OL-like (empty symbols, interrupted line) structures formed by the human mitochondrial tRNA Asn. Pearson correlation coefficients $r = 0.969$, $P = 0.0$; and $r = 0.450$, $P < 0.05$, respectively (values in the figure indicate the square of r). Such comparisons are calculated for each pair of secondary structures that are compared, RNAs are classified into secondary structure clusters (**Figure 1**) according to highest r.

RNAs were generated by simulations designed to reconstruct likely short primordial genes. Simulations were constrained to produce circular RNAs with codons coding for all 20 amino acids and a stop codon, and to form a stem-loop hairpin. The resulting theoretical RNAs have consensual tRNA sequence properties (Demongeot and Moreira, 2007).

Optimal secondary structures of these 24 theoretical RNAs (Circ1-24, **Table 1**) are compared with optimal secondary structures formed by RNAs used previously (Seligmann and Raoult, 2016, therein Table 1) as shown for Circ1, in **Figure 2**. The secondary structure of Circ1 resembles the cloverleaf structure of tRNA Asn, and much less the OL-like structure formed by that tRNA sequence. Mitochondrial tRNAs form secondary structures that resemble mitochondrial light strand replication origins (OL), presumably function as OLs in OL absence (Desjardins and Morais, 1990; Seligmann and Labra, 2014).

Half of the 24 theoretical RNAs designed by simulations cluster with B1. Ten cluster with D2 (as shown in **Figure 2**), and the remaining two with D1. There are $24 \times 6 = 144$ comparisons between the 24 simulation-generated RNAs and the six RNAs in B1. Of these comparisons, 19 (13.2%) have Pearson correlation coefficients r with $P < 0.05$. Among the $24 \times 17 = 408$ comparisons with D2, 10 (2.45%) have $P < 0.05$. There are $24 \times 13 = 312$ comparisons with D1; only two (0.64%) have $P < 0.05$. No comparison between the 24 simulation-generated RNAs with D2 has $P < 0.05$. Hence our interpretation of B1 as representing secondary structures formed by ancestral RNAs agrees with the independent approach developed by Demongeot and Moreira (2007).

The fact that secondary structures formed by the 24 simulation-generated, presumably ancestral-like RNAs of Demongeot and Moreira (2007) preferentially cluster with B1 confirms the ancestral status of these theoretical RNAs and of cluster B1. The observation that translation/genetic code properties converge with tRNA sequence properties (Demongeot and Moreira, 2007) is also congruent with analyses of tRNAs along the principles of the natural circular code (Michel, 2012, 2013; El Soufi and Michel, 2014, 2015): a set of 20 codons overrepresented in coding vs. other frames of protein coding genes (Arquès and Michel, 1996), which enable coding frame retrieval (Ahmed et al., 2007, 2010; Michel and Seligmann, 2014). Results also fit a model of evolution of secondary structures into sequence signals such as codons (El Houmami and Seligmann, 2017).

Template-Free Synthesized Sequence

Simulation-generated sequences such as those in the previous section are suboptimal to confirm the ancestral/*de novo* status of clusters B1/B2. We analyze the predicted

optimal secondary structure formed by a short sequence that is synthesized template-free by a combination of three archaeal enzymes, including DNA polymerase PolB (Béguin et al., 2015). This sequence is by definition "*de novo*." Its secondary structure properties most resemble those of an exceptional, protein-encoding viroid in cluster B2 (one tailed $P = 0.034$). This result is statistically significant also after considering multiple testing, using the Benjamini-Hochberg correction for false discovery rates that accounts for the number of tests done (Benjamini and Hochberg, 1995). This strengthens the status of cluster B2 as recent spontaneously generated RNAs.

Ancient and *de novo* Viroids

Evidence for recent vs. precellular origins of viroids are equivocal. Potentially, results of previous classifications of viroid secondary structures might be biased by including in analyses viroids with high GC contents and forming complex branching secondary structures. Analyses here include ten viroids forming rod-like secondary structures with GC contents ranging from 35 to 61%. All cluster with D1 and D2 (seven and three viroids, respectively, **Table 2**). No comparison with B1 and B2 has $P < 0.05$. Among 130 and 170 comparisons of these 10 rod-shaped viroids with secondary structures belonging to D1 and D2, 47 and 8 (36.2 and 4.7%, respectively), have $P < 0.05$.

Hence rod-shaped viroids belong to the ancestral tRNA-rRNA axis of molecular evolution. Viroids with more complex branching patters clustered with B2 according to previous analyses (Seligmann and Raoult, 2016). This suggests that viroid 'survival' requires evolutionary secondary structure simplification, perhaps because endonucleases target secondary branching (Fujishima et al., 2011). Results are compatible with mixed evidence for ancient precellular origins and recent *de novo* emergence of viroids.

TABLE 2 | Secondary structure variables and classification of 10 viroids forming rod-shaped secondary structures.

Viroid	*N*	Loop	eLoop	Stem GC	Loop GC	Cluster
CVd IV[a]	284	28.87	9.76	64.36	36.37	D1
TPMV[b]	360	33.06	10.08	36.37	42.02	D2
TASV[b]	360	28.33	11.77	60.02	34.31	D1
ASBV[c]	359	29.81	7.48	63.49	42.06	D1
PTSV[c]	247	32.39	17.50	40.12	31.25	D2
CSV[c]	366	28.96	6.60	55.39	41.51	D2
PBCVd[d]	315	31.43	31.32	71.76	37.37	D1
ADFVd[e]	310	34.52	11.22	61.58	37.37	D1
CEVd[f]	371	30.46	7.97	66.67	46.90	D1
HSVd[f]	302	31.46	8.42	66.67	33.68	D1

Secondary structures are according to corresponding references 1–6. All rod-shaped viroids, independently of GC contents, group with 23S (D1) and 16S (D2) rRNAs, indicating ancient origins.

[a]Citrus viroid IV (Puchta et al., 1991); [b]Tomato planta macho viroid, tomato apical stunt viroid (Kiefer et al., 1983); [c]Avocado sunblotch viroid, Potato spindle tuper viroid, Chrysanthemum stunt viroid (Symons, 1981); [d]Pear blister canker viroid (Hernandez et al., 1992); [e]Apple dimple fruit viroid, (Chiumenti et al., 2014); [f]Citrus exocortis viroid, Hop stunt viroid (Lin et al., 2015).

TABLE 3 | Sequences aligned between *Acanthamoeba castellani*'s mitochondrial rRNA genes and *Acanthamoeba polyphaga* Mimivirus' genome.

rRNA	E value	Mito 5′-3′	Mimivirus 5′-3′
5S	0.001	58–101	768283–768243
18S	0.069	104–162	23259–23317
16S	0.009	1103–1148	759756–759718
23S1	0.0007	3–377	500064–500442
23S2	0.002	962–1291	175514–175166
23S3	0.0003	1970–2181	89598–89391

Pandoravirus' Miniature Inverted-Repeat Transposable Elements, Mite

Genomes of giant viruses include many inverted repeats forming stem-loop hairpins regulating transcription (Byrne et al., 2009; Claverie and Abergel, 2009), reminiscent of mitochondrial posttranscriptional tRNA punctuation (Ojala et al., 1981). These include miniature inverted-repeat transposable elements, MITEs (Fattash et al., 2013). The MITE family submarine in the giant virus *Pandoravirus salinus* presumably invaded that genome relatively recently (Sun et al., 2015).

We classified optimal secondary structures formed by these 29 MITE submarine sequences with clusters in **Figure 1**. No submarine MITE clusters within B1 or B2, but 12 cluster within D1 [11 among 377 comparisons (2.9%) with $P < 0.05$] and 17 cluster within D2 [30 among 493 comparisons (6.1%) with $P < 0.05$]. The 10 submarine MITE sequences integrated in Pandoravirus' protein coding genes cluster slightly more frequently with D2 than D1, as compared to the remaining submarine MITEs (difference not statistically significant). Hence most Pandoravirus submarine MITEs cluster with D2 (characterized by bacterial RPEs and 16S rRNA subunits), resembling RNAs from the main tRNA-rRNA axis. Under this scenario, similarities with D1 (characterized by retroviruses and 23S rRNA subunits) would result from chance or secondary convergences. However, this interpretation does not account for additional lateral transfers: many genes of giant viruses originate from lateral transfers, mainly from bacteria sharing their habitat in their amoeban host (Moliner et al., 2010; Georgiades and Raoult, 2012).

Ribosomal RNA-Like Mimivirus Sequences

We used blastn (Altschul et al., 1997) to explore putative links between ribosomal RNAs and giant virus genomes. For that purpose we extracted rRNA sequences from *Acanthamoeba castellanii*'s complete mitochondrial genome (NC_001637) and aligned these with *Acanthamoeba polyphaga* Mimivirus' complete genome (NC_014649). The amoeba's mitogenome includes four rRNA genes—5S, 16S, and 18S rRNAs—and a 23S-like sequence (Burger et al., 1995). Blastn alignment criteria were set at the shortest word size (7), the weakest match/mismatch scores (1/−1) and gap costs (existence 0; extension 2). For each rRNA we chose the alignment with the lowest (best) e value among the alignments detected by blastn between these mitochondrial rRNA genes and Mimivirus' genome (**Table 3**).

For 23S-like rRNA, three alignments are considered because they represent similarities with different Mimivirus sequences and the alignments have low e values. Secondary structures formed by four of the six rRNA sequences aligning with Mimivirus sequences cluster best with D2 and two sequences cluster best with B2 (**Figure 1**). Hence these rRNA sequences most resemble the cluster that includes bacterial 16S rRNA subunits.

These rRNA sequences aligned with sequences from the Mimivirus genome. These Mimivirus sequences also form secondary structures which cluster differently in **Figure 1** than their putative amoeban mitochondrial rRNA homologs. Optimal secondary structures formed by four of the six corresponding Mimivirus DNA sequences cluster best with B2 and the remaining two with D2. Very few of the r coefficients used for these classifications have $P < 0.05$. Hence these results must be considered as tentative.

Differences in clustering by secondary structures formed by mitochondrial rRNA sequences vs. Mimivirus' genome for aligning sequence pairs suggest that alignments are frequently due to convergences between rRNAs and viral sequences. A possible interpretation of these clustering results (**Table 1**) is that viruses tend to create *de novo* rRNA-like sequences, though some of the alignments might suggest regular homology due to common ancestry. Lateral transfers between the host and Mimivirus is also a reasonable explanation. Independently of lateral transfers, this putative mitochondrial rRNA-Mimivirus homology is in line with common ancestry between Megavirales and a cellular ancestor of mitochondria, as suggested by homologies between polymerases of these organisms (Kempken et al., 1992; Rohe et al., 1992; Kapitonov and Jurka, 2006; Yutin et al., 2013; Krupovic and Koonin, 2016; Koonin and Krupovic, 2017) and above noted similar regulations of posttranscriptional processing (vertebrate mitochondria, Ojala et al., 1981; Claverie and Abergel, 2009; Mimivirus, Byrne et al., 2009). Hence secondary structure analyses apparently strengthen the hypothesis that mitochondria share a common ancestor with Megavirales.

Flavivirid Virus Leading Regions

Analysis of conserved secondary structures formed by 3′ and 5′ RNA structures in four flavivirid viruses [Dengue (DENV), West Nile (WNV), Japanese encephalitis (JEV) and tick-borne encephalitis (YFV) viruses, Brinton and Basu, 2015; therein Figure 1B, secondary structure variables here in **Table 1**] show that these structures cluster with D1 (characterized by 23S rRNA subunits). Hence these Flavivirus sequences crucial to replication form structures that resemble 23S rRNA, as observed for most rod-shaped viroids (section Ancient and *de novo* Viroids). Most Pandoravirus MITE sequences resemble D2, characterized by 16S rRNA (section Pandoravirus' Miniature Inverted-Repeat Transposable Elements, Mite) and some Mimivirus sequences resembling mitochondrial rRNAs cluster within D2 (section Ribosomal RNA-Like Mimivirus Sequences). Overall results indicate the hypothesized link between viral RNAs and rRNAs.

A Template-Free RNA in Mimivirus' Transcriptome Persists during Infection

Mimivirus' transcriptome [data from (Legendre et al., 2010), available at: http://sra.dnanexus.com/studies/SRP001690/experiments] includes numerous short RNAs that are not homologous to Mimivirus' genome. Here we focus on a 42-nucleotide-long sequence (5′-GAGACACGCAACAGGGG ATAGGCAAGGCACACAGGGGATAGG-3′) because this sequence also occurs according to Blastn in diverse taxa: Megavirus, Dengue and West Nile genomes, copperhead snake microsatellites and levant cotton repeats. This Mimivirus RNA matches the giant virus Megavirus terra1 (KF527229, positions 903759-903800) and some (not all) genomes of Dengue and West Nile viruses (both are Flaviviridae). In these Flaviviridae, the sequence is inserted in (or close to) the 5′ UTR leading region. The e value of alignments between Mimivirus' RNA and the flavivirid 5′ UTR sequences is 2×10^{-11}. This RNA is detected by Blastn (word size 7; Match/Mismatch scores 1,−1; Gap costs, existence 1, extension 1) in Mimivirus' transcriptome throughout amoeban infection: −15, 0, 90, 180, 360a,b, 540, and 720 min after infection. This RNA does not map on any region of Mimivirus' genome.

Potential Origins of Mimivirus' Template-Free RNA

This RNA not templated by Mimivirus' genome might originate from accidental contamination. However, three arguments suggest less conventional hypotheses. Firstly, this RNA is repeatedly detected throughout the virus' infection cycle, hence in different sequencing events. Secondly, the exact same RNA is detected each time in terms of sequence and length. Thirdly, this exact sequence occurs in the genome of another giant virus, Megavirus terra1. These arguments suggest that the occurrence of this RNA in Mimivirus' transcriptome is not circumstantial.

This RNA could originate from (a) *de novo* creation, as suggested for some short stem-loop hairpin RNAs (Seligmann and Raoult, 2016) and analyses in previous sections, (b) pools of vertically transmitted RNAs originating from horizontal transfers (Stedman, 2013) forming quasi species groups (Villarreal, 2015, 2016), or (c) noncanonical transcriptions of genomic DNA, such as RDDs [RNA-DNA differences, which result from posttranscriptional nucleotide substitutions (Li et al., 2011) or indels (insertions/deletions, Chen and Bundschuh, 2012)]. In addition, this RNA might result from a peculiar type of transcription, which produces transcripts matching genomes only if one assumes that transcription systematically exchanges nucleotides over the whole length of the transcript, which is called swinger transcription. There are 23 possible nucleotide exchange rules, 9 are symmetric exchanges of type X<->Y and 14 are asymmetric exchanges of the type X->Y->Z->X. These 23 transformations are each separately applied *in silico* to Mimivirus' genome to produce 23 swinger-transformed versions of that genome which are used for further analyses. Current information on systematic nucleotide exchanges is reviewed by Seligmann (2017), with further references therein (see also Seligmann, 2013; Michel and Seligmann, 2014).

Swinger Transcript or Template Free RNA Polymerization?

Our preliminary explorations of the kinetic data of the transcriptome of the giant virus Mimivirus (Raoult et al., 2004; Legendre et al., 2010, 2011) detected putative swinger RNAs among RNAs that do not match the regular genome sequence, after excluding regular RNAs by mapping on the regular genome (**Figures 3–5**). **Table 4** compares abundances and mean lengths of detected putative swinger RNAs among transcriptomic data produced by 454 and SOLID massive sequencing techniques (SOLID data unpublished, available in our laboratory). Results by both techniques are comparable. Abundances estimated from 454 sequencing correlate positively with those produced by SOLID (Spearman rank nonparametric correlation rs = 0.323, one tailed $P = 0.066$). Similarly, mean lengths of swinger reads correlate positively for data produced by 454 and SOLID (rs = 0.394, one tailed $P = 0.082$). Combining P-values from these two tests using Fisher's method for combining P values (Fisher, 1950), which sums $-2 \times \ln(\text{Pi})$ where i ranges from 1 to k tests (here $k = 2$) yields a chisquare statistic with $2 \times$ k degrees of freedoms with a combined $P = 0.034$. Hence results from both sequencing methods are overall congruent,

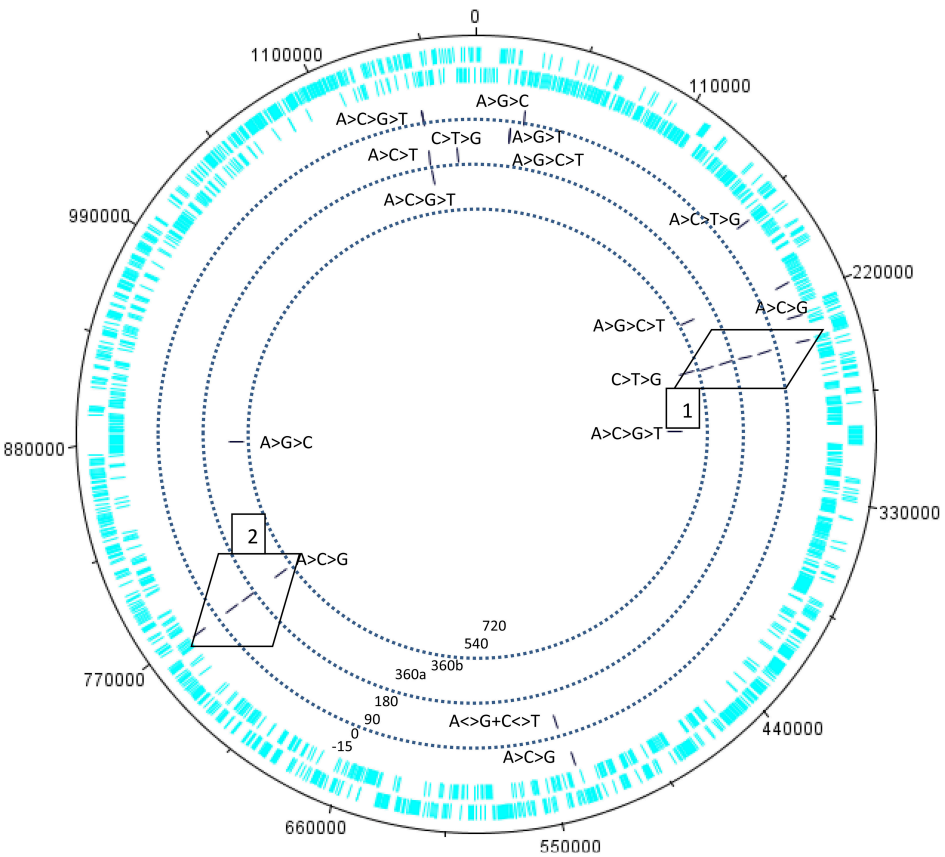


FIGURE 3 | Kinetics of the swinger transcriptome (asymmetric nucleotide exchanges) of *Mimivirus*. Concentric circles indicate different periods in the cycle of *Mimivirus*. Bars indicate swinger RNAs detected for a given genomic region, the swinger transformation rule is indicated next to the bar. Swinger RNAs in boxes 1 and 2 are the exact inverse complements of each other, and correspond to the 5′ UTR leading region of Dengue and West Nile viruses.

Box1

Mimivirus DNA 243480 caattagtttcgtcatcag---AAGGGAGGTTA----ACAAGGGGATAGGCAAGGCAGGGAGTAGT----gataatcagaaaattcaccgtac
Swinger C>T/U>G RNAs AAGCAGTGGTATCAACGCAGCAGGGGATAGGCAAGGCACACAGGGGATAGG
Box2

Mimivirus DNA 768534 ctgcgtcttttgtcc----CTCTATTGGCGACGTCCCTATCCCCTGCTGCCCCGCTCGCCCTTGCTT----cttcttcccctcctg
Swinger A>C>G RNAs CCTATCCCCTGTGTGCCTTGCCTATCCCCTGCTGCGTTGATACC-ACTGCTTTAGG

FIGURE 4 | Alignment between sequence in **Boxes 1** and **2** (**Figure 3**) and homologous sequences: *Mimivirus* reference genome sequence.Underlined are the detected swinger RNA sequences. Small caps indicate neighboring sequences that are not transformed by nucleotide exchange.

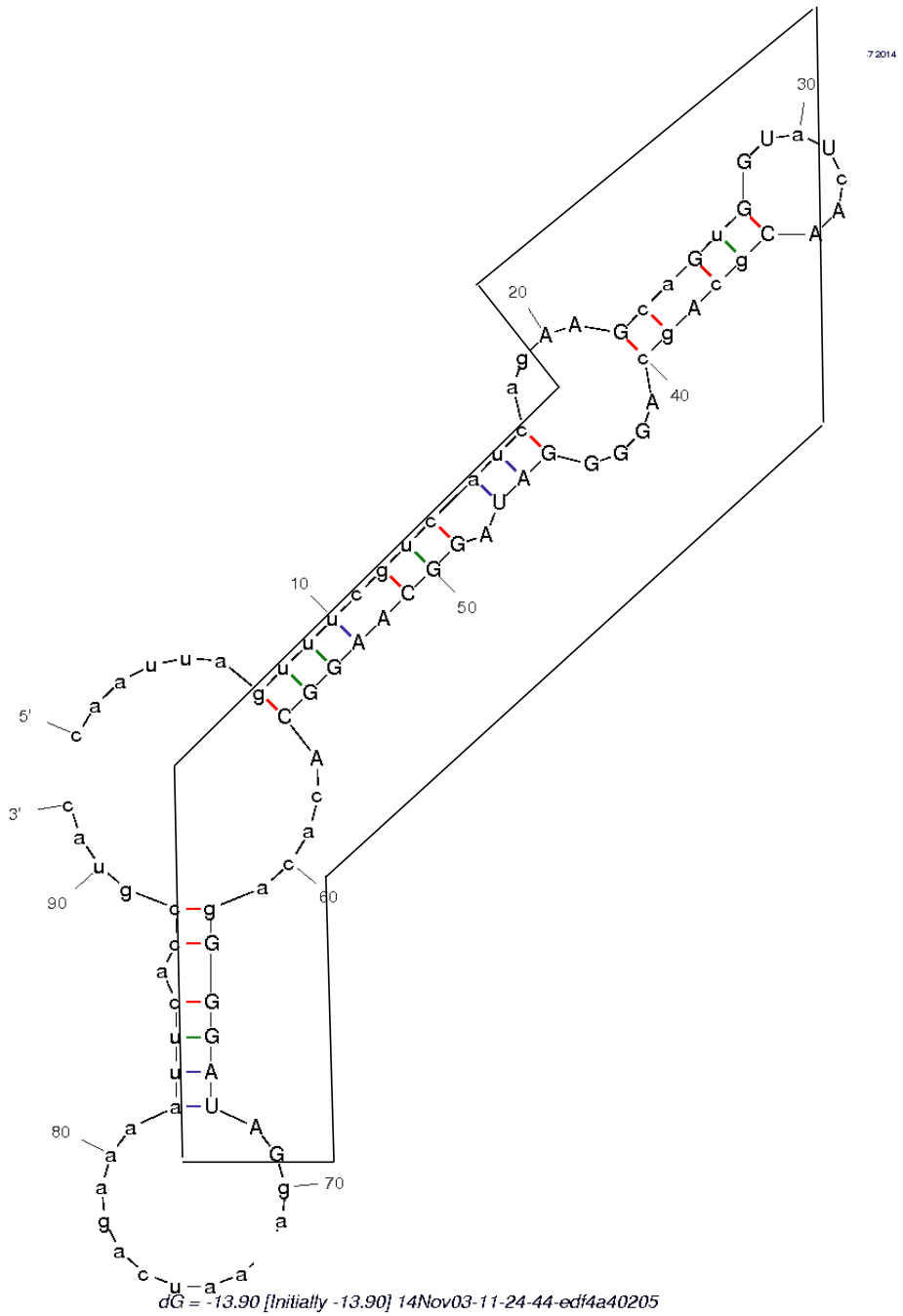


FIGURE 5 | Secondary structure formed by the swinger C>T/U>G RNA (in box) and its untransformed neighboring regions.

swinger RNAs are not artifacts resulting from massive sequencing technologies.

Within 454 data, two swinger RNAs (boxes in **Figure 4**, primary and secondary structures in **Figures 4**, **5**) were detected throughout most of the viral cycle, corresponding to genomic sequences at positions 243499-243537 (C->T->G->C swinger rule, box 1 in **Figure 3**) and 768549-768596 (A->C->G->A swinger rule, box 2 in **Figure 4**). These two genomic positions are each other's inverse complement. This is also the case for the corresponding swinger RNA reads. These reads correspond to the aforementioned template-free RNA and are homologous with Megavirus terra1 (KF527229: positions 903764-903800). This candidate swinger RNA aligns only with swinger transformed versions of Mimivirus' genomic sequence, but its similarity with the swinger transformed genome is also low.

TABLE 4 | Swinger transformations of the genome of Mimivirus, and swinger transcripts.

	454		Solid	
Swinger	***N***	**Mean**	***N***	**Mean**
A<->C	5	51.40	7261	35.52
A<->G	8	211.63	5294	36.47
A<->T	0		7234	35.99
C<->G	173	196.74	7314	35.92
C<->T	12	199.58	5800	36.29
G<->T	1	56.00	5553	35.63
A<->C-G<->T	0		2634	36.60
A<->G-C<->T	1	68.00	2672	36.37
A<->T-C<->G	0		1805	37.14
A->C->G->A	9	62.11	6238	35.73
A->C->T->A	1	64.00	4963	36.64
A->G->C->A	2	70.50	5013	36.68
A->G->T->A	1	79.00	6127	35.70
A->T->C->A	0		6325	35.97
A->T->G->A	0		4826	36.72
C->G->T->C	0		4856	36.68
C->T->G->C	9	52.67	5962	35.70
A->C->G->T->A	4	84.50	6007	36.28
A->C->T->G->A	1	77.00	8421	35.83
A->G->C->T->A	2	74.50	7675	35.45
A->G->T->C->A	0		8444	35.85
A->T->C->G->A	0		5289	35.71
A->T->G->C->A	0		5429	36.41

Columns are: 1–2. Number and mean length (nucleotides) of swinger RNAs matching the Mimivirus' genome swinger version over>half of its length with >80% identity, 454 sequencing, all times since infection pooled (Legendre et al., 2011); 3–4. Number and mean length (nucleotides) of swinger RNAs matching the Mimivirus' genome swinger version over >half of its length with>80% identity, SOLID sequencing, at 16 h post infection.

Putatively, some genomic sequences in *Mimivirus* could originate from horizontal transfers from other viruses, suggesting a potential implication in the horizontal transfer of the Sputnik virophage that infests *Mimivirus* (La Scola et al., 2008). This is in line with the chimeric origin of most *Mimivirus* genes (originating from eukaryotic hosts and bacterial co-parasites of the eukaryotic host, Moreira and Brochier-Armanet, 2008; Jeudy et al., 2012), and confirms horizontal transfer of other viral sequences (Yutin et al., 2013), via swinger transformations.

GENERAL COMMENTS

Analyses that integrate molecular structure information are surprisingly successful at resolving various biological problems, notably of ancient evolution (for example the presumed monophyletic cellular origin of viruses, Nasir and Caetano-Anollés, 2015). However, techniques enabling this remain relatively inaccessible, notably for RNAs. This situation is particularly true if analyzes are supposed to integrate nonequilibrium dynamics of folding during the synthesis of the molecule [proteins: cotranslational folding, (Holtkamp et al., 2015; Seligmann and Warthi, 2017); RNAs: cotranscriptional folding, Gong et al., 2017]. Indeed, most RNAs fold during their syntheses, the new fold appearing after a new stretch of nucleotides is added to the elongating RNA (Schroeder et al., 2002). Each new fold is partially constrained by the previous one, which renders prediction algorithms more complex (Zhao et al., 2010; Frieda and Block, 2012).

The approach used here does not require any special computational skills. It could be adapted to dynamical contexts, and to include information on groups of closely related, suboptimal secondary structures, when these are only slightly more unstable than the optimal structure. Previous analyses (Seligmann and Raoult, 2016) included such special cases, for mitochondrial tRNAs in their classical cloverleaf fold, and in their replication origin (OL)-like fold. Analyzes separating different folds with similar stabilities for structures formed by the same sequence, such as cloverleaf vs. OL-like tRNA structures, yield sometimes different results in classifications such as that in **Figure 1** (see Seligmann and Raoult, 2016, therein Figure 2).

In addition, molecules with very similar estimates for all variables may form very different structures, or similar structures of very different sizes (the variables do not include sequence length). Hence the simple approach used here could be applied to study closely related clouds of secondary structures formed by a given RNA. It could also be adapted to discriminate further between similar secondary structures by including additional variables, such as GC contents in internal vs. external loops, and angular rotation between stem branches.

Overall our results indicate that the versatility of RNA structures enable for functional novelties. Their physicochemical properties are also compatible with this role in primitive protolife organic conditions.

CONCLUSIONS

Interpretations of a previous classification of RNA secondary structures formed by a variety of RNAs (Seligmann and Raoult, 2016) were tested by classifying specific RNAs of special interest. For example, circular RNAs generated by simulations presumably mimicking ancestral RNAs cluster mainly, as expected, with cluster B1, a presumed group of ancestral RNAs (**Figure 1**).

Cluster B2 was previously interpreted as representing *de novo* emerged RNAs because it grouped short simple RNAs from very different functional types (tRNAs, ribozymes, viroids, etc.). Secondary structures formed by a sequence synthesized template free (hence *de novo*) cluster with B2.

Rodshaped viroids from a wide range of GC contents belong independently of GC contents to D1 and D2, two clusters characterized by rRNAs and parasitic sequences. Viroids forming secondary structures characterized by more complex branching patterns belong to cluster B2 (putative *de novo* RNAs). Hence results suggest that some viroids are recent, rodshaped ones are presumably ancient RNAs.

Presumed parasitic palindromic sequences from Pandoravirus (MITE submarine family) resemble cluster D2. D2 is characterized by16S rRNA subunits and Rickettsial palindromic elements that parasitize *Rickettsia* genomes. This fits previous grouping of secondary structures formed by parasitic sequences with 23S and 16S rRNAs (clusters D1 and D2).

Mimivirus sequences that align with amoeaban mitochondrial rRNA genes also strengthen suspected evolutionary links between rRNA and viral sequences. Mitochondrial rRNA sequences aligning with Mimivirus sequences cluster as expected with D2, characterized by bacterial 16S rRNA. Interestingly, secondary structures formed by Mimivirus sequences with which the mitochondrial rRNAs align, cluster mainly with B2, suggesting recent *de novo* origins for viral sequences resembling rRNAs. These results are based on relatively weak similarities and hence can only be considered as preliminary. Nevertheless, they suggest that viruses produce rRNA-like sequences, in line with the prediction that the study of giant viruses will 'change current conceptions of life, diversity and evolution' (Abrahao et al., 2014).

The secondary structure formed by the 5′ UTR leading region of flavivirid viruses clusters also with D2, hence it is a further viral, rRNA-like sequence. An RNA persisting throughout Mimivirus' infection cycle and lacking homology with Mimivirus' genome occurs also in some flavivirus genomes, Megavirus, in copperhead snakes and levant cotton. This saltatory phylogenetic distribution is compatible with repeated spontaneous, template free synthesis by polymerases deterministically producing specific sequences, as observed in Archaea (Béguin et al., 2015). Indeed, this RNA, embedded in the surrounding regular 5′ and 3′ sequences (**Table 1** and **Figure 5**), forms a secondary structure that clusters with B2. This would suggest *de novo* emergence of this unusual RNA.

Analyses assuming different scenarios based on noncanonical transcriptions do not reach clear-cut conclusions on the origin of that RNA that does not map to the Mimivirus genome. Though contamination cannot be totally excluded, other hypotheses seem plausible. Indeed, this RNA maps imperfectly on Mimivirus' swinger-transformed genome (Mimivirus' transcriptome includes numerous swinger RNAs, results from each 454 and SOLID massive sequencing methods are congruent). Overall, results hint that parasitic RNAs form rRNA-like secondary structures, and template free polymerizations apparently enrich genomes with new RNA/DNA sequences.

AUTHOR CONTRIBUTIONS

HS and DR designed the study and analyses.

FUNDING

This study was supported by Méditerranée Infection and the National Research Agency under the program "Investissements d'avenir," reference ANR-10-IAHU-03 and the A∗MIDEX project (no ANR-11-IDEX-0001-02). We thank Thi Tien Nguyen for technical help.

REFERENCES

AbouHaidar, M. G., Venkataraman, S., Golshani, A., Liu, B., and Ahmad, T. (2014). Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14542–14547. doi: 10.1073/pnas.1402814111

Abrahao, J. S., Dornas, F. P., Silva, L. C., Almeida, G. M., Boratto, P. V., Colson, P., et al. (2014). *Acanthamoeba polyphaga* mimivirus and other giant viruses: an open field to outstanding discoveries. *Virol. J.* 11:120. doi: 10.1186/1743-422X-11-120

Ahmed, A., Frey, G., and Michel, C. J. (2007). Frameshift signals in genes associated with the circular code. *In Silico Biol.* 7, 155–168.

Ahmed, A., Frey, G., and Michel, C. J. (2010). Essential molecular functions associated with the circular code evolution. *J. Theor. Biol.* 264, 613–622. doi: 10.1016/j.jtbi.2010.02.006

Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Amiri, H., Alsmark, C. M., and Anderson, S. G. (2002). Proliferation and deterioration of Rickettsia palindromic elements. *Mol. Evol. Biol.* 19, 1234–1243. doi: 10.1093/oxfordjournals.molbev.a004184

Arquès, D. G., and Michel, C. J. (1996). A complementary circular code in the protein coding genes. *J. Theor. Biol.* 182, 45–58. doi: 10.1006/jtbi.1996.0142

Bansho, Y., Ichihashi, N., Kazuta, Y., Matsuura, T., Suzuki, H., and Yomo, T. (2012). Importance of parasite RNA species repression for prolonged translation-coupled RNA self-replication. *Chemistry Biol.* 19, 478–487. doi: 10.1016/j.chembiol.2012.01.019

Barthélémy, R. M., and Seligmann, H. (2016). Cryptic tRNAs in chaetognath mitochondrial genomes. *Comput. Biol. Chem.* 62, 119–132. doi: 10.1016/j.compbiolchem.2016.04.007

Béguin, P., Gill, S., Charpin, N., and Forterre, P. (2015). Synergistic template-free synthesis of dsDNA by *Thermococcus nautili* primase PolpTN2, DNA polymerase PolB, and pTN2 helicase. *Extremophiles* 19, 69–76. doi: 10.1007/s00792-014-0706-1

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 27, 289–300.

Bloch, D. P., McArthur, B., Guimaraes, R. C., Smith, J., and Staves, M. P. (1989). tRNA-rRNA sequence matches inter- and intraspecies comparisons suggest common origins for the two RNAs. *Braz. J. Med. Biol.* 22, 931–944.

Bloch, D. P., McArthur, B., Widdowson, R., Spector, D., Guimaraes, R. C., and Smith, J. (1983). tRNA-rRNA sequence homologies: evidence for a common evolutionary origin? *J. Mol. Evol.* 19, 420–428. doi: 10.1007/BF02102317

Bloch, D. P., McArthur, B., Widdowson, R., Spector, D., Guimaraes, R. C., and Smith, J. (1984). tRNA-rRNA sequence homologies: a model for the origin of a common ancestral molecule, and prospects for its reconstruction. *Orig. Life* 14, 571–578. doi: 10.1007/BF00933706

Branciamore, S., and Di Giulio, M. (2011). The presence in tRNA molecule sequences of the double hairpin, an evolutionary stage through which the origin of this molecule is thought to have passed. *J. Mol. Evol.* 72, 352–363. doi: 10.1007/s00239-011-9440-9

Branciamore, S., and Di Giulio, M. (2012). The origin of the 5s ribosomal RNA molecule could have been caused by a single inverse duplication: strong evidence from its sequences. *J. Mol. Evol.* 74, 170–186. doi: 10.1007/s00239-012-9497-0

Brinton, M. A., and Basu, M. (2015). Functions of the 3′ and 5′ genome regions of members of the genus Flavivirus. *Virus Res.* 206, 108–119. doi: 10.1016/j.virusres.2015.02.006

Burger, G., Plante, I., Lonergan, K. M., and Gray, M. W. (1995). The mitochondrial DNA of the amoeboid protozoon, *Acanthamoeba castellanii*:

complete sequence, gene content and genome organization. *J. Mol. Biol.* 245, 522–537. doi: 10.1006/jmbi.1994.0043

Bussière, F., Lafontaine, D., Côté, F., Beaudry, D., and Perreault, J. P. (1995). Evidence for a model ancetsral viroid. *Nucleic Acids Symp. Ser.* 1995, 143–144.

Byrne, D., Grzela, R., Larigue, A., Audic, S., Chenivesse, S., Encinas, S., et al. (2009). The polyadelynaton site of Mimivirus transcripts obeys a stringent "hairpin rule". *Genome Res.* 19, 1233–1242. doi: 10.1101/gr.091561.109

Chaikeeratisak, V., Nguyen, K., Khanna, K., Brilot, A. F., Erb, M. L., Coker, J. K. C., et al. (2017). Assembly of a nucleus-like structure during viral replication in bacteria. *Science* 355, 194–197. doi: 10.1126/science.aal2130

Chen, C., and Bundschuh, R. (2012). Systematic investigation of insertional and deletional RNA–DNA differences in the human transcriptome. *BMC Genomics* 13:616. doi: 10.1186/1471-2164-13-616

Chiumenti, M., Torchetti, E. M., Di Serio, F., and Minafra, A. (2014). Identification and characterization of a viroid resembling apple dimple fruit viroid in fig (*Ficus carica* L.) by next generation sequencing of small RNAs. *Virus Res.* 188, 54–59. doi: 10.1016/j.virusres.2014.03.026

Claverie, J. M., and Abergel, C. (2009). Mimivirus and its virophage. *Annu. Rev. Genet.* 43, 49–66. doi: 10.1146/annurev-genet-102108-134255

Demongeot, J., and Moreira, A. (2007). A possible circular RNA at the origin of life. *J. Theor. Biol.* 249, 314–324. doi: 10.1016/j.jtbi.2007.07.010

Desjardins, P., and Morais, R. (1990). Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. *J. Mol. Biol.* 2012, 599–634. doi: 10.1016/0022-2836(90)90225-B

Di Giulio, M. (1992). The evolution of aminoacyl-tRNA synthetases, the biosynthetic pathways of amino acids and the genetic code. *Orig. Life Evol. Biosph.* 22, 309–319. doi: 10.1007/BF01810859

Di Giulio, M. (1994). The phylogeny of tRNA molecules and the origin of the genetic code. *Orig. Life Evol. Biosph.* 24, 425–434. doi: 10.1007/BF015 82018

Di Giulio, M. (1995). Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? *J. Theor. Biol.* 177, 95–101. doi: 10.1016/S0022-5193(05) 80007-4

Di Giulio, M. (1999). The non-monophyletic origin of the tRNA molecule. *J. Theor. Biol.* 197, 403–414. doi: 10.1006/jtbi.1998.0882

Di Giulio, M. (2008). The split genes of *Nanoarchaeum equitans* are an ancestral character. *Gene* 421, 20–26. doi: 10.1016/j.gene.2008.06.010

Di Giulio, M. (2009). Formal proof that the split genes of tRNAs of *Nanoarchaeum equitans* are an ancestral character. *J. Mol. Evol.* 69, 505–511. doi: 10.1007/s00239-009-9280-z

Di Giulio, M. (2012). The 'recently' split transfer RNA genes may be close to merging the two halves of the tRNA rather than having just separated them. *J. Theor. Biol.* 310, 1–2. doi: 10.1016/j.jtbi.2012.06.022

Di Giulio, M. (2013). A polyphyletic model for the origin of tRNAs has more support than a monophyletic model. *J. Theor. Biol.* 318, 124–128. doi: 10.1016/j.jtbi.2012.11.012

Diener, T. O. (1996). Origin and evolution of viroids and viroid-like satellite RNAs. *Virus Genes* 11, 119–131. doi: 10.1007/978-1-4613-1407-3_5

Diener, T. O. (2016). Viroids: "living fossils" of primordial RNAs? *Biol. Direct* 11, 15. doi: 10.1186/s13062-016-0116-7

El Houmami, N., and Seligmann, H. (2017). Evolution of nucleotide punctuation marks: from structural to linear signals. *Front. Genet.* 8:36. doi: 10.3389/fgene.2017.00036

El Soufi, K., and Michel, C. J. (2014). Circular code motifs in the ribosome decoding center. *Comput. Biol. Chem.* 52, 9–17. doi: 10.1016/j.compbiolchem.2014.08.001

El Soufi, K., and Michel, C. J. (2015). Circular code motifs near the ribosome decoding center. *Comput. Biol. Chem.* 59A, 158–176. doi: 10.1016/j.compbiolchem.2015.07.015

Fattash, I., Rooke, R., Wong, A., Hui, C., Luu, T., Bhardwaj, P., et al. (2013). Miniature inverted-repeat transposable elements: discovery, distribution, and activity. *Genome* 56, 475–486. doi: 10.1139/gen-2012-0174

Fisher, R. A. (1950). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Fox, G. E. (2010). Origin and evolution of the ribosome. *Cold Spring Harb. Perspect. Biol.* 2:a003483. doi: 10.1101/cshperspect.a003483

Frieda, K. L., and Block, S. M. (2012). Direct observation of cotranscriptional folding in an adenine riboswitch. *Science* 338, 397–400. doi: 10.1126/science.1225722

Fujishima, K., Sugihara, J., Miller, C. S., Baker, B. J., Di Giulio, M., Takesue, K., et al. (2011). A novel three-unit tRNA splicing endonuclease found in ultrasmall Archaea possesses broad substrate specificity. *Nucleic Acids Res.* 39, 9695–9704. doi: 10.1093/nar/gkr692

Gale, M., Tan, S. L., and Katze, M. G. (2000). Translational control of viral gene expression in eukaryotes. *Microbiol. Mol. Biol. Rev.* 64, 239–280. doi: 10.1128/MMBR.64.2.239-280.2000

Georgiades, K., and Raoult, D. (2012). How microbiology helps defined the rhizome of life. *Front. Cell Infect. Microbiol.* 2:60. doi: 10.3389/fcimb.2012.00060

Gillespie, J. J., Joardar, V., Williams, K. P., Driscoll, T., Hostetler, J. B., Nordberg, E., et al. (2012). A Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. *J. Bacteriol.* 194, 376–394. doi: 10.1128/JB.06244-11

Gong, S., Wang, Y., Wang, Z., and Zhang, W. (2017). Computational methods for modeling aptamers and designing riboswitches. *Int. J. Mol. Sci.* 18:2442. doi: 10.3390/ijms18112442

Hernandez, C., Elena, S. F., Moya, A., and Flores, R. (1992). Pear blister canker viroid is a member of the apple scar skin subgroup (apscaviroids) and also has sequence homology with viroids from other subgroups. *J. Gen. Virol.* 73, 2503–2507. doi: 10.1099/0022-1317-73-10-2503

Higgs, P. G., and Lehman, N. (2015). The RNA world: molecular cooperation at the origins of life. *Nat. Rev. Genet.* 16, 7–17. doi: 10.1038/nrg3841

Hiscox, J. A. (2007). RNA viruses: hijacking the dynamic nucleolus. *Nature Rev. Microbiol.* 5, 119–127. doi: 10.1038/nrmicro1597

Holtkamp, W., Kokic, G., Jäger, M., Mittelstaet, J., Komar, A. A., and Rodnina, M. V. (2015). Cotranslational protein folding on the ribosome monitored in real time. *Science* 350, 1104–1107. doi: 10.1126/science.aad0344

Jeudy, S., Abergel, C., Claverie, J. M., and Legendre, M. (2012). Translation in giant viruses: a unique mixture of bacterial and eukaryotic termination schemes. *PLoS Genet.* 8:e1003122. doi: 10.1371/journal.pgen.1003122

Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2007). Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Prot.* 7, 40–44. doi: 10.1021/pr700739d

Kapitonov, V. V., and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4540–4545. doi: 10.1073/pnas.0600833103

Kempken, F., Hermanns, J., and Osiewacz, H. D. (1992). Evolution of linear plasmids. *J. Mol. Evol.* 35, 502–513. doi: 10.1007/BF00160211

Kiefer, M. C., Owens, R. A., and Diener, T. O. (1983). Structural similarities between viroids and transposable genetic elements. *Proc. Natl. Acad. Sci. U.S.A.* 80, 6234–6238. doi: 10.1073/pnas.80.20.6234

Koonin, E. V., and Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* 3, 546–557. doi: 10.1016/j.coviro.2013.06.008

Koonin, E. V., and Krupovic, M. (2017). Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Curr. Opin. Virol.* 25, 7–15. doi: 10.1016/j.coviro.2017.06.008

Krupovic, M., and Koonin, E. V. (2016). Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Curr. Opin. Microbiol.* 31, 25–33. doi: 10.1016/j.mib.2016.01.006

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., et al. (2008). The virophage as a unique parasite of the giant Mimivirus. *Nature* 455, 100–104. doi: 10.1038/nature07218

Legendre, M., Audic, S., Poirot, O., Hingamp, P., Seltzer, V., Byrne, D., et al. (2010). mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res.* 20, 664–674. doi: 10.1101/gr.102582.109

Legendre, M., Santini, S., Rico, A., Abergel, C., and Claverie, J. M. (2011). Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing. *Virol. J.* 8:99. doi: 10.1186/1743-422X-8-99

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., and Tug, J. M. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 333, 53–58. doi: 10.1126/science.1207018

Lin, C. Y., Wu, M. L., Shen, T. L., Yeh, H. H., and Hung, T. H. (2015). Multiplex detection, distribution, and genetic diversity of Hop stunt viroid

and Citrus exocortis viroid infecting citrus in Taiwan. *Virol. J.* 12, 11. doi: 10.1186/s12985-015-0247-y

Michel, C. J. (2012). Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. *Comput. Biol. Chem.* 37, 24–37. doi: 10.1016/j.compbiolchem.2011.10.002

Michel, C. J. (2013). Circular code motifs in transfer RNAs. *Comput. Biol. Chem.* 45, 17–29. doi: 10.1016/j.compbiolchem.2013.02.004

Michel, C. J., and Seligmann, H. (2014). Bijective transformation circular codes and nucleotide exchanging RNA transcription. *Biosystems* 118, 39–50. doi: 10.1016/j.biosystems.2014.02.002

Moliner, C., Fournier, P. E., and Raoult, D. (2010). Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiol. Rev.* 34, 281–294. doi: 10.1111/j.1574-6976.2009.00209.x

Moreira, D., and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* 8:12. doi: 10.1186/1471-2148-8-12

Nasir, A., and Caetano-Anollés, G. (2015). A phylogenomic data-driven exploration of viral origins and evolution. *Science Adv.* 1:e1500527. doi: 10.1126/sciadv.1500527

Ojala, D., Montoya, J., and Attardi, G. (1981). tRNA punctuation model of RNA processing in human mitochondria. *Nature* 290, 470–474. doi: 10.1038/290470a0

Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., et al. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Inform. Exp.* 3:1. doi: 10.1186/2042-5783-3-1

Penny, D. (2015). Cooperation and selfishness both occur during molecular evolution. *Biol. Direct* 10:26. doi: 10.1186/s13062-014-0026-5

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments? *BMJ* 318:1236. doi: 10.1136/bmj.316.7139.1236

Petrov, A. S., Gulen, B., Norris, A. M., Kovaxs, N. A., Berbier, C. R., Lanier, K. A., et al. (2015). History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15396–15401. doi: 10.1073/pnas.1509761112

Puchta, H., Ramm, K., Luckinger, R., Hadas, R., Bar-Joseph, M., and Saenger, H. L. (1991). Primary and secondary structure of citrus viroid IV (CVd IV), a new chimeric viroid present in dwarfed grapefruit in Israel. *Nucleic Acids Res.* 19:6640. doi: 10.1093/nar/19.23.6640

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., et al. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science* 305, 1344–1350. doi: 10.1126/science.1101485

Rohe, M., Schründer, J., Tudzynski, P., and Meinhardt, F. (1992). Phylogenetic relationships of linear, protein-primed replicating genomes. *Curr. Genet.* 21, 173–176. doi: 10.1007/BF00318478

Root-Bernstein, M., and Root-Bernstein, R. (2015). The ribosome as a missing link in the evolution of life. *J. Theor. Biol.* 367, 130–158. doi: 10.1016/j.jtbi.2014.11.025

Root-Bernstein, R., and Root-Bernstein, M. (2016). The ribosome as a missing link in prebiotic evolution II: ribosomes encode ribosomal proteins that bind to common regions of their own mRNAs and rRNAs. *J. Theor. Biol.* 397, 115–127. doi: 10.1016/j.jtbi.2016.02.030

Ruiz-Mirazo, K., Briones, C., and de la Escosura, A. (2014). Prebiotic systems chemistry: new perspectives for origins of life. *Chem. Rev.* 114, 285–366. doi: 10.1021/cr2004844

Schroeder, R., Grossberger, R., Pichler, A., and Waldsich, C. (2002). RNA folding *in vivo*. *Curr. Opin. Struct. Biol.* 12, 296–300. doi: 10.1016/S0959-440X(02)00325-1

Seligmann, H. (2013). Polymerization of non-complementary RNA: systematic symmetric nucleotide exchanges mainly involving uracil produce mitochondrial RNA transcripts coding for cryptic overlapping genes. *Biosystems* 111, 156–174. doi: 10.1016/j.biosystems.2013.01.011

Seligmann, H. (2014). Putative anticodons in mitochondrial tRNA sidearm loops: pocketknife tRNAs? *J. Theor. Biol.* 340, 155–163. doi: 10.1016/j.jtbi.2013.08.030

Seligmann, H. (2017). Natural mitochondrial proteolysis confirms transcription systematically exchanging/deleting nucleotides, peptides coded by expanded codons. *J. Theor. Biol.* 414, 76–90. doi: 10.1016/j.jtbi.2016.11.021

Seligmann, H., and Labra, A. (2014). The relation between hairpin formation by mitochondrial WANCY tRNAs and the occurrence of the light strand replication origin in Lepidosauria. *Gene* 542, 248–257. doi: 10.1016/j.gene.2014.02.021

Seligmann, H., and Raoult, R. (2016). Unifying view of stem–loop hairpin RNA as origin of current and ancient parasitic and non-parasitic RNAs, including in giant viruses. *Curr. Opin. Microbiol.* 31, 1–8. doi: 10.1016/j.mib.2015.11.004

Seligmann, H., and Warthi, G. (2017). Genetic code optimization for cotranslational protein folding: codon directional asymmetry correlates with antiparallel betasheets, tRNA synthetase classes. *Comput. Struct. Biotechnol. J.* 15, 412–424. doi: 10.1016/j.csbj.2017.08.001

Stedman, K. (2013). Mechanisms for RNA capture by ssDNA viruses: grand theft RNA. *J. Mol. Evol.* 76, 359–364. doi: 10.1007/s00239-013-9569-9

Sun, C., Feschotte, C., Wu, Z., and Mueller, R. L. (2015). DNA transposons have colonized the genome of the giant virus *Pandoravirus salinus*. *BMC Biol.* 15:38. doi: 10.1186/s12915-015-0145-1

Symons, R. H. (1981). Avocado sunblotch viroid: primary sequence and proposed secondary structure. *Nucleic Acids Res.* 9, 6527–6537. doi: 10.1093/nar/9.23.6527

Szostak, J. W. (2009). Origins of life: systems chemistry on Earth. *Nature* 459, 171–172. doi: 10.1038/459171a

Villarreal, L. P. (2015). Force for ancient and recent life: viral and stem-loop RNA consortia promote life. *Ann. N.Y. Acad. Sci.* 1341, 25–34. doi: 10.1111/nyas.12565

Villarreal, L. P. (2016). Persistent virus and addiction modules: an engine of symbiosis. *Curr. Opin. Microbiol.* 31, 70–79. doi: 10.1016/j.mib.2016.03.005

Widmann, J., Di Giulio, M., Yarus, M., and Knight, R. (2005). tRNA creation by hairpin duplication. *J. Mol. Evol.* 61, 524–530. doi: 10.1007/s00239-004-0315-1

Xie, A., and Scully, R. (2017). Hijacking the DNA damage response to enhance viral replication: γ-herpesevirus 68 orf36 phosphorylates histone H2AX. *Mol. Cell* 27, 178–179. doi: 10.1016/j.molcel.2007.07.005

Yutin, N., Raoult, D., and Koonin, E. V. (2013). Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol. J.* 10, 158. doi: 10.1186/1743-422X-10-158

Zhao, P., Zhang, W.-B., and Chen, S.-J. (2010). Predicting secondary structural folding kinetics for nucleic ccids. *Biophys. J.* 98, 1617–1625. doi: 10.1016/j.bpj.2009.12.4319

Zuker, M. (2003). mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415 doi: 10.1093/nar/gkg595

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

PERSPECTIVE
published: 20 November 2017
doi: 10.3389/fmicb.2017.02292

Natural Antisense Transcripts at the Interface between Host Genome and Mobile Genetic Elements

*Hany S. Zinad, Inas Natasya and Andreas Werner**

RNA Interest Group, Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, United Kingdom

Non-coding RNAs are involved in epigenetic processes, playing a role in the regulation of gene expression at the transcriptional and post-transcriptional levels. A particular group of ncRNA are natural antisense transcripts (NATs); these are transcribed in the opposite direction to protein coding transcripts and are widespread in eukaryotes. Their abundance, evidence of phylogenetic conservation and an increasing number of well-characterized examples of antisense-mediated gene regulation are indicative of essential biological roles of NATs. There is evidence to suggest that they interfere with their corresponding sense transcript to elicit concordant and discordant regulation. The main mechanisms involved include transcriptional interference as well as dsRNA formation. Sense–antisense hybrid formation can trigger RNA interference, RNA editing or protein kinase R. However, the exact molecular mechanisms elicited by NATs in the context of these regulatory roles are currently poorly understood. Several examples confirm that ectopic expression of antisense transcripts trigger epigenetic silencing of the related sense transcript. Genomic approaches suggest that the antisense transcriptome carries a broader biological significance which goes beyond the physiological regulation of the directly related sense transcripts. Because NATs show evidence of conservation we speculate that they played a role in evolution, with early eukaryotes gaining selective advantage through the regulatory effects. With the surge of genome and transcriptome sequencing projects, there is promise of a more comprehensive understanding of the biological role of NATs and the regulatory mechanisms involved.

Keywords: natural antisense transcripts, gene expression regulation, double stranded RNA (dsRNA), non-coding RNA, RNA interference, DNA methylation, histone modifications

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria

Reviewed by:
Tara Patricia Hurst,
Abcam, United Kingdom
King-Hwa Ling,
Universiti Putra Malaysia, Malaysia

***Correspondence:**
Andreas Werner
andreas.werner@ncl.ac.uk

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *31 July 2017*
Accepted: *06 November 2017*
Published: *20 November 2017*

Citation:
Zinad HS, Natasya I and Werner A (2017) Natural Antisense Transcripts at the Interface between Host Genome and Mobile Genetic Elements. Front. Microbiol. 8:2292. doi: 10.3389/fmicb.2017.02292

INTRODUCTION

Natural antisense transcripts (NATs) are arguably the oldest group within the family of non-coding RNAs. The first examples of bi-directionally transcribed genes were detected as early as in the 1980s (Beiter et al., 2009). It then emerged that human and mouse imprinted gene clusters express antisense transcripts. Interestingly, antisense transcription is associated with allele-specific gene silencing, not only in imprinted gene clusters but also other bi-directionally transcribed loci (Verona et al., 2003; Carlile et al., 2009; Werner and Swan, 2010). The general and widespread expression of NATs emerged at the beginning of the genomic era with the computational analyses of human and mouse NATs (Lehner et al., 2002; Shendure and Church, 2002). These reports analyzed

the ever increasing repository of full-length sequences and sequence tags for complementary transcripts and identified over a 100 sense–antisense pairs. They set the stage for a series of seminal computational and large scale experiments to detect complementary transcripts and decipher their putative biological roles (Kiyosawa et al., 2003; Chen et al., 2004; Katayama et al., 2005; Werner et al., 2007). The initial efforts to characterize antisense RNAs preceded the development of RNA-seq platforms and, as a consequence, only include reasonably abundant, stable and mostly cloned transcripts. Antisense transcripts were then defined as long, non-coding RNAs that are complementary to a coding transcript from the opposite strand (**Figure 1**). Nowadays, NATs are widely recognized as versatile regulators of gene expression. Intriguingly, many of the associated regulatory pathways involve double-stranded RNA intermediates that are reminiscent of viral structures or transposon intermediates.

Natural antisense transcription has been detected in bacteria, yeast and all eukaryotes. Extensive research into various aspects of RNA biology and epigenetics have revealed a variety of species-specific mechanisms dealing with bi-directional transcription and complementary RNA molecules. As a result, in multicellular organisms such as plants, *Caenorhabditis elegans* or mammals, NATs will trigger different mechanisms and elicit drastically different cellular responses. In plants, best described in *Arabidopsis thaliana*, complementary RNA triggers a strong RNA interference response and the formation of siRNAs from the double-stranded sequence (Baulcombe, 2004). Moreover, DNA methylation can be induced as a consequence of double stranded RNA (dsRNA) formation and the action of an RNA-dependent RNA polymerase (RdRP) (Matzke and Birchler, 2005). The system is thought to protect plants from viral infections and genomic parasites. *C. elegans* also expresses an RdRP and has the potential to amplify a dsRNA response that leads to endo-siRNA production from endogenous dsRNA structures (Gu et al., 2009). These are thought to enable self-recognition and prevent the integration of foreign DNA into the genome. More complex animals lack an amplifying system for dsRNA and, as a consequence, at least mechanisms that involve sense–antisense transcript hybridization will differ significantly between various organisms. To what extent natural antisense related dsRNA formation triggers an antiviral response, RNA interference or helps to control transposon activity is intensely debated. It appears that in chordates, including human and mouse, the response to dsRNA varies fundamentally between germ cells, stem cells and differentiated somatic cells (Cullen et al., 2013).

In recent years the field has focused on the characterization of specific sense–antisense transcript pairs predominantly in a pathophysiological context. The scope of this article is to discuss a few prominent examples of gene regulation by NATs and set them in context with evolutionary considerations.

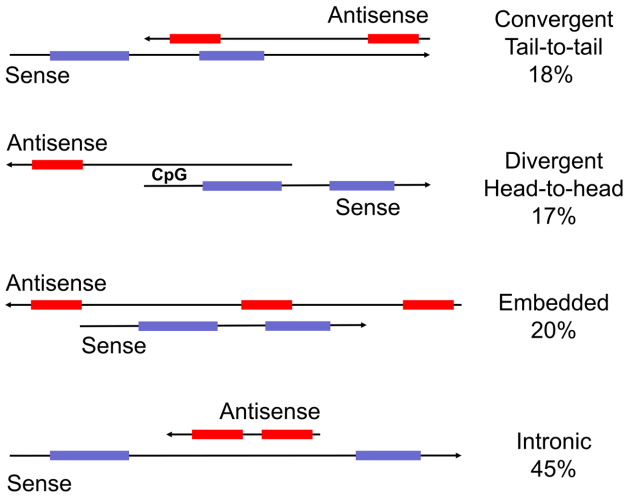


FIGURE 1 | Schematic representation of sense and antisense transcripts from bi-directionally transcribed loci. Commonly used descriptions are head-to-head, tail-to-tail, embedded and intronic (from top). The percent representation in the human genome is from Balbin et al. (2015). Importantly, transcripts in tail-to-tail and full overlap configuration potentially form double-stranded RNA whereas the other two categories do not. Other terms used are 'convergent' (tail-to-tail), divergent (head-to-head), or 'non-overlapping' (embedded).

MOBILE GENETIC ELEMENTS AND ANTISENSE TRANSCRIPTION

The significance of NATs in a genomic context cannot be appreciated without considering the impact of mobile genetic elements, including transposons and viruses. For example, antisense transcription is often initiated by the insertion of transposable elements with promoter activity downstream of protein coding genes (Conley et al., 2008). NATs have also been associated with controlling the activity of transposons and mitigating the consequences of their insertion into a complex genome (Stein et al., 2015). Importantly, NATs that are co-expressed with their cognate sense transcripts may form dsRNA intermediates reminiscent of viral structures that activate an immune response, which in turn induces significant expression changes in the antisense transcriptome (Ilott et al., 2014).

The expansion of mobile genetic elements has not only increased genetic plasticity but also introduced promoters and enhancers to initiate the transcription of novel loci. Since the number of protein coding genes has not increased significantly during the evolution of complex organisms the transposition of genetic elements has primarily resulted in enhanced transcription of non-coding, regulatory RNAs (Mattick, 2001). This also applies to NATs, demonstrated by a moderate but significant accumulation of antisense transcriptional start sites downstream of protein coding genes. These coincide with ancient MIR and L2 transposon sequences; the observation that ancient transposons drive more antisense transcripts than the younger L1 or Alu elements suggests phylogenetic functional conservation of antisense transcription (Conley et al., 2008). Interestingly, the mammalian X chromosome shows an inverse trend: NATs are significantly under-represented despite an accumulation of transposable elements and a proposed role for L1 elements in maternal X chromosome inactivation (Emerson et al., 2004; Abrusan et al., 2008). Considering the potential of NATs to

epigenetically silence the related protein-coding sense gene, it is conceivable that a bi-directional arrangement is detrimental in a monoallelic context whereas it may prove advantageous in a bi-allelic background (Werner et al., 2009).

Transposon mobilization, viral infection and sense/antisense expression can all form dsRNA intermediates that are potentially damaging to the cell. In vertebrates, two principal mechanisms have evolved to mitigate the deleterious consequences of transposon mobilization and protect cells from viral infections: the piRNA/endo-siRNA system and protein kinase R/interferon, respectively (**Figure 2A**). The two protective mechanisms show distinct expression patterns. In pluripotent stem cells, during early embryogenesis as well as in female and male germ cells the piRNA/endo-siRNA system restricts retro-transposition (Okamura and Lai, 2008). On the other hand, PKR/IFN are predominantly active in differentiated, somatic cells and provide protection against viral infections. To what extent RNA interference plays a role in somatic cells against viruses is a matter of intense debate (Cullen et al., 2013). Experimental attempts to demonstrate virus-derived siRNAs after infection of cultured cells are technically challenging and may not represent a physiologically relevant model (Jeffrey et al., 2017; tenOever, 2017). Accordingly, we found no evidence of abundant endo-siRNA expression in human cells, though the few loci that produced endo-siRNAs tended to be bi-directionally transcribed (Werner et al., 2014).

On the other hand, endo-siRNAs and piRNA are readily detectable in germ and zygotic cells. Interestingly, the short RNA pattern is qualitatively very similar between zygote and oocytes but distinctly different in spermatozoa where it includes piRNAs and endo-siRNAs from loci that potentially form dsRNA precursors (Garcia-Lopez et al., 2014). This may reflect pervasive transcription and the resulting highly complex transcriptome in these cells (Laiho et al., 2013; Soumillon et al., 2013). This assumption concurs with findings from Dicer knock-out mice that showed spermatogenic defects that coincided chronologically with transcriptional silencing and chromatin condensation (Korhonen et al., 2011). Moreover, abundant endo-siRNAs map to protein coding genes and potentially regulate the expression of their targets (Song et al., 2011). The finding that endo-siRNAs silence L1 retrotransposons through DNA methylation suggest that these short RNAs are in fact capable of establishing a widespread, cell specific genomic imprint in male germ cells (Chen et al., 2012). This observation has prompted speculations that spermatogenic endo-siRNAs and hence NATs could play an essential role in the evolution of complex organisms (Werner et al., 2015). The genome undergoes various changes during spermatogenesis such as demethylation and potential activation of transposable elements as well as genomic recombination that requires extensive DNA repair. We recently proposed a hypothesis how NATs help to detect genes that produce inadequate RNA output thus providing a genomic quality control (Werner et al., 2015). In various contexts RNA is being used to maintain genome integrity (Duharcourt et al., 2009) or distinguish self from novel genetic material (Gu et al., 2009). An RNA-based control mechanism to maintain integrity of the genome seems therefore conceivable.

Natural antisense transcripts are involved in regulating gene expression in an immune challenge; however, a protective reaction involving NATs seems unlikely. Conversely, recent evidence suggests that herpesviruses induce wide-spread host antisense transcription to interfere with the expression of pro-apoptotic genes (Wyler et al., 2017). Upon lipopolysaccharide exposure, human monocytes differentially express more than 200 long non-coding RNAs of which about half can be categorized as NATs (Ilott et al., 2014). Two of these were further characterized and shown to regulate the proinflammatory mediators IL1β (Interleukin 1β) and CXCL8 (C-X-C Motif Chemokine Ligand 8). Likewise, the expression of IL1α (Interleukin 1 α) is regulated by a NAT (Chan et al., 2015). These findings indicate that NATs in immune cells have specific, gene regulatory tasks whereas in germ cells a broader role in genome maintenance has been suggested.

FUNCTIONAL RELEVANCE OF NATs

The contribution of NATs to maintaining cellular homeostasis is a matter of intense scrutiny – though many questions remain. On the one hand NATs are abundant in every genome and numbers tend to increase in higher eukaryotes. On the other hand, the evidence of antisense transcripts contributing to *homeostatic* gene regulation – apart from parentally imprinted genes – is circumstantial.

Genome-wide studies have established phylogenetic conservation, expression pattern or co-regulation with other transcripts and support the biological relevance of non-coding transcripts, as do loss of function experiments (Diederichs, 2014; Goff and Rinn, 2015). The phylogenetic conservation of NATs has been scrutinized widely and the perception has changed over time. The observation in early computational experiments that antisense transcripts show splicing differences and often minimal sequence identity between closely related species argued against stringent conservation (Veeramachaneni et al., 2004; Wood et al., 2013). However, recent reports based on microarray or RNAseq data, taking conserved transcription or expression patterns into account, confirm phylogenetic conservation of antisense transcription (Ling et al., 2013; Hezroni et al., 2015; Ning et al., 2017).

The vast majority of NATs are expressed at low levels, one to three magnitudes lower than the corresponding sense transcripts, and the two RNAs tend to co-purify (Okazaki et al., 2002; Werner et al., 2007; Ling et al., 2013). In mammals, testis shows the highest level of antisense transcription, specifically in developing sperm cells (Carlile et al., 2009; Soumillon et al., 2013). However, the wide-spread antisense transcription in testis could be a mere consequence of the post-mitotic transcriptional burst during spermatogenesis (Lee et al., 2009; Laiho et al., 2013).

To what extent the sense and antisense transcripts are present in the same cell is often unclear. A recent report has demonstrated co-localization of Sox4 sense/antisense transcripts in mouse brain cells (Ling et al., 2016). Moreover, antisense transcripts in a head-to-head orientation tend to show concordant expression, possibly the result of bi-directional CpG-island containing promoters, which results in the co-expression of sense and antisense

transcripts (Balbin et al., 2015). On the other hand, we and others have found limited evidence for the presence of genic dsRNA (Werner et al., 2014). It is conceivable, however, that co-expression of head-to-head sense/antisense pairs is tolerated whereas tail-to-tail pairs tend to exclude each other.

MECHANISMS OF ANTISENSE REGULATION

There are three different levels at which bi-directional transcription and a putative NAT can affect the corresponding sense RNA. Firstly, transcription from one strand can interfere with the transcription on the opposite strand, thus influencing the production of the sense transcript, so-called 'transcriptional interference' (**Figure 2B**). This mechanism is often portrayed as two polymerase complexes crashing into each other which may happen under experimentally engineered circumstances but is unlikely to be relevant *in vivo* (Prescott and Proudfoot, 2002; Osato et al., 2007; Wang et al., 2014). A more likely cause of events would see transcription of one strand altering DNA structure and DNA–protein interactions of the specific locus on the opposite strand, thus affecting its transcription. Alternatively, two close transcription sites may compete for protein factors that enable initiation and elongation. Secondly, the complementary sense and antisense transcripts can hybridize and form a dsRNA intermediate (**Figure 2A**). This interaction potentially interferes with the processing of both RNAs, their splicing, nuclear export or even translation (Hastings et al., 2000; Ning et al., 2017). Alternatively, dsRNA is recognized by enzymes that resolve the double-strand structure and trigger various cellular responses (Wang and Carmichael, 2004). The best described dsRNA specific enzymes include ADARs (Adenosine Deaminases Acting on RNA) (Mannion et al., 2015), RNases type III (Dicer) (Svobodova et al., 2016) and protein kinase R (Munir and Berg, 2013). Thirdly, NATs may act independently of the cognate sense transcript and adopt the function of a long non-coding RNA. In fact, one of the best described lncRNA, HOTAIR, is transcribed antisense to HOXC11 and both HOTAIR and HOXC11 are concordantly upregulated in urothelial cancer (Heubach et al., 2015). Nevertheless, extensive research in the field has not investigated sense/antisense interactions but established HOTAIR's interaction with polycomb repressive complex 2 and histone modification complexes (Heubach et al., 2015). The function of lncRNAs in sequestering proteins and miRNAs to provide a scaffold for regulatory complexes is described in detail elsewhere and is beyond the scope of this article (Rinn and Chang, 2012).

TRANSCRIPTIONAL INTERFERENCE

It is well-established that expression of an antisense transcript leads to epigenetic repression of the related sense transcript. This has been extensively described in the context of X chromosome inactivation and parental imprinting. For example, suppression of Tsix (antisense to Xist) leads to ectopic expression of Xist and concomitant bi-allelic X inactivation in XX cells or silencing of the X chromosome in XY cells (Lee, 2000). Likewise, the imprinted gene clusters *Igf2r*/*Slc22a2*/*Slc22a3* and *Kcnq1* contain NATs (Airn and Kcnq1ot1, respectively) that are essential for parental imprinting (Sleutels et al., 2002; Thakur et al., 2004). Deletion of the antisense transcript interferes with the methylation status of the locus, alleviates silencing and leads to bi-allelic expression of the gene cluster (Sleutels et al., 2002; Mohammad et al., 2010). Thereby, the interference of the antisense transcript with either promoter or enhancer region triggers a gene-specific- or a broader response affecting the entire gene cluster, respectively (Kornienko et al., 2013). Bi-directional transcription is one of the key features of parentally imprinted gene clusters. The exact mechanistic consequences of the regulatory antisense transcripts, however, are not fully understood and distinct, cluster-specific differences occur (Kanduri, 2016).

In humans and mice, a few examples of transcriptional interference have been studied in detail and represent paradigms for the consequences of aberrant expression of NATs. All lead to specific pathological phenotypes that are related to the protein coding sense gene, predicting a strictly *cis*-acting mechanism of interaction. The first example relates to a rare form of α-thalassemia; in affected patients, the constitutively active LUC7L gene downstream of HBA2 is truncated, including the loss of the polyadenylation site. As a consequence, *LUC7L* transcription continues into *HBA2* and a NAT to *HBA2* is produced. Comparable effects were achieved when *LUC7L* was replaced with a different gene (*UBC*) confirming an essential role for transcription, independent of the nature of the gene (Tufarelli et al., 2003). The bi-directional tumor suppressor gene *p15*/*p15AS* shows a comparable arrangement, with a naturally occurring, lowly expressed antisense transcript (Yu et al., 2008). Enhanced expression of the antisense transcript and concomitant reduction of p15 cyclin-dependent kinase inhibitor was found in leukemia patient samples and also in two acute myeloid leukemia lines. The mechanism of silencing appeared to involve both altered histone modifications, increased H3K9me2 and decreased H3K4me2, as well as promoter DNA methylation depending on the cellular model system studied. Interestingly, a transfected construct that recapitulated the p15 genomic arrangement with inducible antisense transcription showed *cis*-silencing of the exogenous construct but also, with lesser penetrance, reduced endogenous p15. Both silencing mechanisms were shown to be Dicer-independent and to introduce stable epigenetic modifications (Yu et al., 2008). The reported *trans*-effect of the p15AS transcript suggests that the different mechanisms by which NATs interfere with sense transcript expression may depend on cell-specific features or sense/antisense transcript levels. Of note, transcriptional interference has also been reported between two consecutive genes on the same DNA strand. In a patient cohort with Lynch syndrome, the mismatch repair gene *MSH2* is epigenetically silenced by the truncated *TACSTD1 upstream* of *MSH2*. The resulting read-through transcript runs into *MSH2* and induces CpG methylation and silencing of the promoter (Ligtenberg et al., 2009). The underlying mechanism is yet unclear but could involve interactions of the read-through RNA

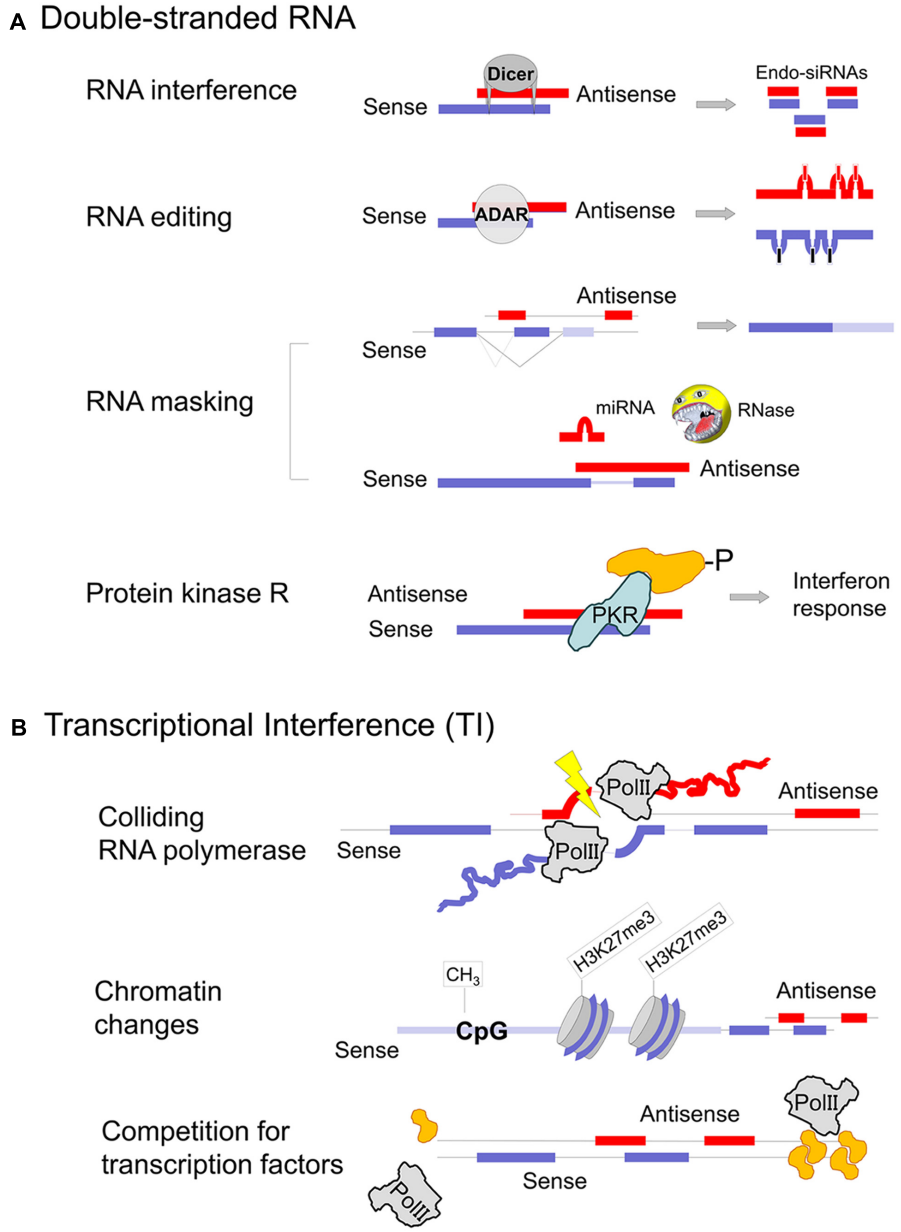


FIGURE 2 | Mechanisms of gene regulation involving natural antisense transcripts. **(A)** Co-expression of sense and antisense transcripts in the same cell may cause dsRNA formation. RNA masking can have inhibitory as well as stimulatory consequences on the protein-coding sense mRNA expression, depending on the motif that is obstructed. Due to potential PKR activation dsRNA formation must either occur in specific cellular compartments or in specific cell types that do not rise an interferon response (germ cells and stem cells). Parts of the figure are modified from Wight and Werner (2013). **(B)** Transcriptional interference, where the expression of one transcript affects transcription of the opposite strand, can occur at several levels. Sense–antisense expression shows a discordant pattern, the majority as a result of antisense transcription induced chromatin changes (Weinberg and Morris, 2016).

with antisense transcripts produced from the bidirectional *MSH* promoter (Grzechnik et al., 2014; Uesaka et al., 2014).

DOUBLE-STRANDED RNA FORMATION

With the scale of antisense transcription emerging, it became evident that sense and antisense transcripts tend to be found in the same RNA preparations (Kiyosawa et al., 2003; Werner et al., 2007). This suggested that dsRNA formation could be an important intermediate in gene regulatory mechanisms involving NATs. The cellular pathways triggered by dsRNA were well-established at that time and included processing by Dicer into endo-siRNAs (RNA interference), A to I conversion by adenosine deaminases (RNA editing) as well as the activation of PKR (Wang and Carmichael, 2004). These pathways result in characteristic

intermediates such as short RNAs, modified RNA or increased levels of IFN-α/β (Interferon), respectively, that are used as readouts for bi-directionally transcribed genes.

RNA Interference

The initial discovery of RNA interference was based on the observation that introduction of dsRNA into *C. elegans* triggered highly specific, lasting gene knock-down (Fire et al., 1998). The strategies to adopt a similar approach in mammalian cells, however, failed almost completely. Only a few cell types, oocytes or certain embryonic cells, seem to tolerate significant levels of dsRNA without triggering an immune response (Wianny and Zernicka-Goetz, 2000; Piatek et al., 2017). As a consequence the contribution of Dicer and RNA interference to the processing of natural sense/antisense transcript pairs is controversial. There are a few reports that have identified short RNAs from endogenous RNA duplexes, so-called endo-siRNAs in vertebrates, predominantly in the germline and only few in somatic cells (Watanabe et al., 2008; Carlile et al., 2009; Xia et al., 2013; Werner et al., 2014; Jha et al., 2015). Nevertheless, the biological function of endo-siRNAs in the context of bi-directional transcription is not well-understood. Remarkably, however, both endo-siRNAs and NATs are predominantly found in mammalian testis in accordance with the proposed role in maintaining sperm genome integrity (Song et al., 2011; Soumillon et al., 2013; Werner et al., 2015).

RNA Editing

Members of the ADAR family of adenosine deaminases recognize long stretches of dsRNA and convert adenosines into inosines. This process can either be site-specific or promiscuous (Mannion et al., 2015). Site-specific RNA editing is predominantly observed in the brain and affects a small number of neurotransmitter receptors. Since inosines pair with cytosines (rather than with thymidines) editing leads to point mutations and, consequently, to receptors with altered physiological properties (Schmauss and Howe, 2002). Promiscuous RNA editing acts on long stretches of dsRNA and involves widespread conversion of A to I. The modifications can resolve the double strand and/or interfere with nuclear export. RNA editing predominantly affects repetitive, intronic structures in a co-transcriptional process (Blow et al., 2004; Levanon et al., 2004). Both timing of hyper-editing and large-scale RNAseq data rule out a general contribution of RNA editing to natural sense/antisense RNA processing, though few gene specific regulatory mechanisms involving RNA editing have been reported (Prasanth et al., 2005; Salameh et al., 2015). Moreover, ADARs are induced by an antiviral interferon response and viral dsRNA was found to be hyper-edited supporting a role in innate immunity (Mannion et al., 2015).

RNA Masking

An mRNA contains sequence motifs that are recognized by regulatory proteins and short RNA molecules to control its translation efficiency and half-life. Hybridization of an antisense transcript can potentially interfere with these regulatory interactions in a process called RNA masking. Both stabilizing and de-stabilizing effects of RNA masking by antisense transcripts have been reported. So far these include competition with miRNA binding sites (Faghihi et al., 2010; Liu et al., 2015) and exposure of mRNA degradation motifs (Cayre et al., 2003). In addition, a well-characterized example of antisense RNA-promoted alternative splicing has been reported (Hastings et al., 2000). Here, the levels of two alternative splice forms of the thyroid hormone receptor TRα1 and TRα2 correlate with the expression of an antisense transcript (RevErbα) complementary to the relevant splice site. RevErbα RNA sterically masks the TRα2-specific splice site and promotes TRα1 expression (Hastings et al., 1997). Interestingly, a genome-wide analysis of splicing events using a comprehensive set of exon array data found extensive correlation between antisense transcription and alternative exon usage (Morrissy et al., 2011). Moreover, genes with multiple splice forms are under-represented on metazoan sex chromosomes (Wegmann et al., 2008), a trend that mirrors the limited bi-directional transcription on mammalian X chromosomes (Kiyosawa et al., 2003; Chen et al., 2004). Somewhat counter intuitively, the apparent association between antisense transcription and alternative splicing has not been followed up and underpinned by examples of detailed analyses of specific loci.

Two well-documented examples of RNA masking focus on the bi-directionally transcribed genes BACE1 and HIF1α, both highly relevant to human disease, BACE1 in Alzheimer's disease and HIF1α in cancer. In the former case, the antisense transcript BACE1-AS stabilizes the mRNA encoding β-secretase by masking the binding site of miR-485-5p (Faghihi et al., 2010). This leads to an increased production and accumulation of Amyloid-β (Faghihi et al., 2008). This particular example could apply to a number of bi-directionally transcribed genes since a significant number of NATs overlap with the 3′ end of the sense transcript (Faghihi et al., 2010). The second example of RNA masking includes the hypoxia-inducible factor 1α and the convergently transcribed antisense transcript aHIF (Rossignol et al., 2002). The antisense transcript is widely expressed in healthy tissues but significantly upregulated in various tumors and was proposed as a prognostic marker for cancer progression (Cayre et al., 2003; Dang et al., 2015). The inverse correlation between HIF1α and aHIF was hypothesized to result from an AU-rich element on the HIF1α RNA that becomes accessible upon antisense interaction.

PKR and Innate Immunity

In a quick, first line response, the innate immune system reacts to specific bacterial and viral structures including glycans, lipopolysaccharides, particular proteins, and dsRNA. The latter is recognized by PKR that, upon binding to long RNA duplexes of >30 bp (Lemaire et al., 2008), undergoes dimerization and auto phosphorylation, reduces host protein synthesis and eventually triggers an interferon (IFN) response. Activation of IFN-α/β stimulates the expression of IFN inducible genes (including PKR), inhibits viral protein synthesis by phosphorylating eIF2-α and potentially triggers apoptosis (Marchal et al., 2014). Despite the fact that both PKR and Dicer process dsRNA in the cytoplasm, PKR activation and the IFN response prevail. This is also the reason why gene silencing by RNA interference

(as established in *C. elegans*) is not applicable in most mammalian cells (Billy et al., 2001). From the viewpoint of natural antisense transcription, this poses a major conceptual dilemma: many of the proposed mechanisms established with particular bi-directionally transcribed genes involve cytoplasmic dsRNA intermediates with the potential to activate PKR.

CONCLUSION

The very nature of natural antisense transcription is enigmatic, as large genomes of complex organisms could comfortably accommodate the relatively small number of genes without much interference. Moreover, convergent transcription and dsRNA cause various levels of cellular stress that may even lead to cell death. Nevertheless, NATs are abundant non-coding RNAs that potentially regulate their corresponding sense transcript through a variety of molecular mechanisms. Detailed research into the interplay of sense/antisense transcripts from specific loci has validated biological roles for antisense transcripts, yet mechanistic insights are still rare. As a consequence, the central question why bi-directionally transcribed loci persist and even expand during evolution is still unclear. A way forward here may link particular mechanisms (transcriptional interference, dsRNA formation and RNA interference, RNA masking, RNA editing) to specific categories of NATs (head-to-head, tail-to-tail) and assess these groups in model systems with or without the enzymatic components potentially involved in antisense RNA processing.

AUTHOR CONTRIBUTIONS

HZ, IN, and AW wrote the manuscript and designed the figures.

FUNDING

HZ is supported by a grant from the Higher Committee for Education Development in Iraq (HCED).

REFERENCES

Abrusan, G., Giordano, J., and Warburton, P. E. (2008). Analysis of transposon interruptions suggests selection for L1 elements on the X chromosome. *PLOS Genet.* 4:e1000172. doi: 10.1371/journal.pgen.1000172

Balbin, O. A., Malik, R., Dhanasekaran, S. M., Prensner, J. R., Cao, X., Wu, Y. M., et al. (2015). The landscape of antisense gene expression in human cancers. *Genome Res.* 25, 1068–1079. doi: 10.1101/gr.180596.114

Baulcombe, D. (2004). RNA silencing in plants. *Nature* 431, 356–363. doi: 10.1038/nature02874

Beiter, T., Reich, E., Williams, R. W., and Simon, P. (2009). Antisense transcription: a critical look in both directions. *Cell Mol. Life Sci.* 66, 94–112. doi: 10.1007/s00018-008-8381-y

Billy, E., Brondani, V., Zhang, H., Muller, U., and Filipowicz, W. (2001). Specific interference with gene expression induced by long, double-stranded RNA in mouse embryonal teratocarcinoma cell lines. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14428–14433. doi: 10.1073/pnas.261562698

Blow, M., Futreal, P. A., Wooster, R., and Stratton, M. R. (2004). A survey of RNA editing in human brain. *Genome Res.* 14, 2379–2387. doi: 10.1101/gr.2951204

Carlile, M., Swan, D., Jackson, K., Preston-Fayers, K., Ballester, B., Flicek, P., et al. (2009). Strand selective generation of endo-siRNAs from the Na/phosphate transporter gene Slc34a1 in murine tissues. *Nucleic Acids Res.* 37, 2274–2282. doi: 10.1093/nar/gkp088

Cayre, A., Rossignol, F., Clottes, E., and Penault-Llorca, F. (2003). aHIF but not HIF-1alpha transcript is a poor prognostic marker in human breast cancer. *Breast Cancer Res.* 5, R223–R230. doi: 10.1186/bcr652

Chan, J., Atianand, M., Jiang, Z., Carpenter, S., Aiello, D., Elling, R., et al. (2015). Cutting edge: a natural antisense transcript, AS-IL1alpha, controls inducible transcription of the proinflammatory cytokine IL-1α. *J. Immunol.* 195, 1359–1363. doi: 10.4049/jimmunol.1500264

Chen, J., Sun, M., Kent, W. J., Huang, X., Xie, H., Wang, W., et al. (2004). Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* 32, 4812–4820. doi: 10.1093/nar/gkh818

Chen, L., Dahlstrom, J. E., Lee, S. H., and Rangasamy, D. (2012). Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. *Epigenetics* 7, 758–771. doi: 10.4161/epi.20706

Conley, A. B., Miller, W. J., and Jordan, I. K. (2008). Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet.* 24, 53–56. doi: 10.1016/j.tig.2007.11.008

Cullen, B. R., Cherry, S., and Tenoever, B. R. (2013). Is RNA interference a physiologically relevant innate antiviral immune response in mammals? *Cell Host Microbe* 14, 374–378. doi: 10.1016/j.chom.2013.09.011

Dang, Y., Lan, F., Ouyang, X., Wang, K., Lin, Y., Yu, Y., et al. (2015). Expression and clinical significance of long non-coding RNA HNF1A-AS1 in human gastric cancer. *World J. Surg. Oncol.* 13, 302. doi: 10.1186/s12957-015-0706-3

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends Genet.* 30, 121–123. doi: 10.1016/j.tig.2014.01.004

Duharcourt, S., Lepere, G., and Meyer, E. (2009). Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. *Trends Genet.* 25, 344–350. doi: 10.1016/j.tig.2009.05.007

Emerson, J. J., Kaessmann, H., Betran, E., and Long, M. (2004). Extensive gene traffic on the mammalian X chromosome. *Science* 303, 537–540. doi: 10.1126/science.1090042

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., et al. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* 14, 723–730. doi: 10.1038/nm1784

Faghihi, M. A., Zhang, M., Huang, J., Modarresi, F., Van Der Brug, M. P., Nalls, M. A., et al. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol.* 11:R56. doi: 10.1186/gb-2010-11-5-r56

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391, 806–811. doi: 10.1038/35888

Garcia-Lopez, J., Hourcade Jde, D., Alonso, L., Cardenas, D. B., and Del Mazo, J. (2014). Global characterization and target identification of piRNAs and endo-siRNAs in mouse gametes and zygotes. *Biochim. Biophys. Acta* 1839, 463–475. doi: 10.1016/j.bbagrm.2014.04.006

Goff, L. A., and Rinn, J. L. (2015). Linking RNA biology to lncRNAs. *Genome Res.* 25, 1456–1465. doi: 10.1101/gr.191122.115

Grzechnik, P., Tan-Wong, S. M., and Proudfoot, N. J. (2014). Terminate and make a loop: regulation of transcriptional directionality. *Trends Biochem. Sci.* 39, 319–327. doi: 10.1016/j.tibs.2014.05.001

Gu, W., Shirayama, M., Conte, D. Jr., Vasale, J., Batista, P. J., Claycomb, J. M., et al. (2009). Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol. Cell* 36, 231–244. doi: 10.1016/j.molcel.2009.09.020

Hastings, M. L., Ingle, H. A., Lazar, M. A., and Munroe, S. H. (2000). Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA. *J. Biol. Chem.* 275, 11507–11513. doi: 10.1074/jbc.275.15.11507

Hastings, M. L., Milcarek, C., Martincic, K., Peterson, M. L., and Munroe, S. H. (1997). Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation

but correlates with antisense RNA levels. *Nucleic Acids Res.* 25, 4296–4300. doi: 10.1093/nar/25.21.4296

Heubach, J., Monsior, J., Deenen, R., Niegisch, G., Szarvas, T., Niedworok, C., et al. (2015). The long noncoding RNA HOTAIR has tissue and cell type-dependent effects on HOX gene expression and phenotype of urothelial cancer cells. *Mol. Cancer* 14:108. doi: 10.1186/s12943-015-0371-8

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122. doi: 10.1016/j.celrep.2015.04.023

Ilott, N. E., Heward, J. A., Roux, B., Tsitsiou, E., Fenwick, P. S., Lenzi, L., et al. (2014). Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nat. Commun.* 5:3979. doi: 10.1038/ncomms4979

Jeffrey, K. L., Li, Y., and Ding, S. W. (2017). Reply to 'Questioning antiviral RNAi in mammals'. *Nat. Microbiol.* 2:17053. doi: 10.1038/nmicrobiol.2017.53

Jha, A., Panzade, G., Pandey, R., and Shankar, R. (2015). A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Res.* 43, 8713–8724. doi: 10.1093/nar/gkv871

Kanduri, C. (2016). Long noncoding RNAs: lessons from genomic imprinting. *Biochim. Biophys. Acta* 1859, 102–111. doi: 10.1016/j.bbagrm.2015.05.006

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., et al. (2005). Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566. doi: 10.1126/science.1112009

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., and Hayashizaki, Y. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* 13, 1324–1334. doi: 10.1101/gr.982903

Korhonen, H. M., Meikar, O., Yadav, R. P., Papaioannou, M. D., Romero, Y., Da Ros, M., et al. (2011). Dicer is required for haploid male germ cell differentiation in mice. *PLOS ONE* 6:e24821. doi: 10.1371/journal.pone.0024821

Kornienko, A. E., Guenzl, P. M., Barlow, D. P., and Pauler, F. M. (2013). Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* 11:59. doi: 10.1186/1741-7007-11-59

Laiho, A., Kotaja, N., Gyenesei, A., and Sironen, A. (2013). Transcriptome profiling of the murine testis during the first wave of spermatogenesis. *PLOS ONE* 8:e61558. doi: 10.1371/journal.pone.0061558

Lee, J. T. (2000). Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. *Cell* 103, 17–27. doi: 10.1016/S0092-8674(00)00101-X

Lee, T. L., Pang, A. L., Rennert, O. M., and Chan, W. Y. (2009). Genomic landscape of developing male germ cells. *Birth Defects Res. C Embryo Today* 87, 43–63. doi: 10.1002/bdrc.20147

Lehner, B., Williams, G., Campbell, R. D., and Sanderson, C. M. (2002). Antisense transcripts in the human genome. *Trends Genet.* 18, 63–65. doi: 10.1016/S0168-9525(02)02598-2

Lemaire, P. A., Anderson, E., Lary, J., and Cole, J. L. (2008). Mechanism of PKR Activation by dsRNA. *J. Mol. Biol.* 381, 351–360. doi: 10.1016/j.jmb.2008.05.056

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* 22, 1001–1005. doi: 10.1038/nbt996

Ligtenberg, M. J., Kuiper, R. P., Chan, T. L., Goossens, M., Hebeda, K. M., Voorendt, M., et al. (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1. *Nat. Genet.* 41, 112–117. doi: 10.1038/ng.283

Ling, K. H., Brautigan, P. J., Moore, S., Fraser, R., Cheah, P. S., Raison, J. M., et al. (2016). Derivation of an endogenous small RNA from double-stranded Sox4 sense and natural antisense transcripts in the mouse brain. *Genomics* 107, 88–99. doi: 10.1016/j.ygeno.2016.01.006

Ling, M. H., Ban, Y., Wen, H., Wang, S. M., and Ge, S. X. (2013). Conserved expression of natural antisense transcripts in mammals. *BMC Genomics* 14:243. doi: 10.1186/1471-2164-14-243

Liu, J., Wu, W., and Jin, J. (2015). A novel mutation in SIRT1-AS leading to a decreased risk of HCC. *Oncol. Rep.* 34, 2343–2350. doi: 10.3892/or.2015.4205

Mannion, N., Arieti, F., Gallo, A., Keegan, L. P., and O'connell, M. A. (2015). New insights into the biological role of mammalian ADARs; the RNA editing proteins. *Biomolecules* 5, 2338–2362. doi: 10.3390/biom5042338

Marchal, J. A., Lopez, G. J., Peran, M., Comino, A., Delgado, J. R., Garcia-Garcia, J. A., et al. (2014). The impact of PKR activation: from neurodegeneration to cancer. *FASEB J.* 28, 1965–1974. doi: 10.1096/fj.13-248294

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* 2, 986–991. doi: 10.1093/embo-reports/kve230

Matzke, M. A., and Birchler, J. A. (2005). RNAi-mediated pathways in the nucleus. *Nat. Rev. Genet.* 6, 24–35. doi: 10.1038/nrg1500

Mohammad, F., Mondal, T., Guseva, N., Pandey, G. K., and Kanduri, C. (2010). Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development* 137, 2493–2499. doi: 10.1242/dev.048181

Morrissy, A. S., Griffith, M., and Marra, M. A. (2011). Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res.* 21, 1203–1212. doi: 10.1101/gr.113431.110

Munir, M., and Berg, M. (2013). The multiple faces of protein kinase R in antiviral defense. *Virulence* 4, 85–89. doi: 10.4161/viru.23134

Ning, Q., Li, Y., Wang, Z., Zhou, S., Sun, H., and Yu, G. (2017). The evolution and expression pattern of human overlapping lncRNA and protein-coding gene pairs. *Sci. Rep.* 7:42775. doi: 10.1038/srep42775

Okamura, K., and Lai, E. C. (2008). Endogenous small interfering RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 9, 673–678. doi: 10.1038/nrm2479

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573. doi: 10.1038/nature01266

Osato, N., Suzuki, Y., Ikeo, K., and Gojobori, T. (2007). Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* 176, 1299–1306. doi: 10.1534/genetics.106.069484

Piatek, M. J., Henderson, V., Fearn, A., Chaudhry, B., and Werner, A. (2017). Ectopically expressed Slc34a2a sense-antisense transcripts cause a cerebellar phenotype in zebrafish embryos depending on RNA complementarity and Dicer. *PLOS ONE* 12:e0178219. doi: 10.1371/journal.pone.0178219

Prasanth, K. V., Prasanth, S. G., Xuan, Z., Hearn, S., Freier, S. M., Bennett, C. F., et al. (2005). Regulating gene expression through RNA nuclear retention. *Cell* 123, 249–263. doi: 10.1016/j.cell.2005.08.033

Prescott, E. M., and Proudfoot, N. J. (2002). Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8796–8801. doi: 10.1073/pnas.132270899

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

Rossignol, F., Vache, C., and Clottes, E. (2002). Natural antisense transcripts of hypoxia-inducible factor 1α are detected in different normal and tumour human tissues. *Gene* 299, 135–140. doi: 10.1016/S0378-1119(02)01049-1

Salameh, A., Lee, A. K., Cardo-Vila, M., Nunes, D. N., Efstathiou, E., Staquicini, F. I., et al. (2015). PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8403–8408. doi: 10.1073/pnas.1507882112

Schmauss, C., and Howe, J. R. (2002). RNA editing of neurotransmitter receptors in the mammalian brain. *Sci. STKE* 2002:pe26. doi: 10.1126/stke.2002.133.pe26

Shendure, J., and Church, G. M. (2002). Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.* 3:RESEARCH0044. doi: 10.1186/gb-2002-3-9-research0044

Sleutels, F., Zwart, R., and Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* 415, 810–813. doi: 10.1038/415810a

Song, R., Hennig, G. W., Wu, Q., Jose, C., Zheng, H., and Yan, W. (2011). Male germ cells express abundant endogenous siRNAs. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13159–13164. doi: 10.1073/pnas.1108567108

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* 3, 2179–2190. doi: 10.1016/j.celrep.2013.05.031

Stein, P., Rozhkov, N. V., Li, F., Cardenas, F. L., Davydenko, O., Vandivier, L. E., et al. (2015). Essential Role for endogenous siRNAs during meiosis in mouse oocytes. *PLOS Genet.* 11:e1005013. doi: 10.1371/journal.pgen.1005013

Svobodova, E., Kubikova, J., and Svoboda, P. (2016). Production of small RNAs by mammalian Dicer. *Pflugers Arch.* 468, 1089–1102. doi: 10.1007/s00424-016-1817-6

tenOever, B. R. (2017). Questioning antiviral RNAi in mammals. *Nat. Microbiol.* 2:17052. doi: 10.1038/nmicrobiol.2017.52

Thakur, N., Tiwari, V. K., Thomassin, H., Pandey, R. R., Kanduri, M., Gondor, A., et al. (2004). An antisense RNA regulates the bidirectional silencing property

of the Kcnq1 imprinting control region. *Mol. Cell. Biol.* 24, 7855–7862. doi: 10.1128/MCB.24.18.7855-7862.2004

Tufarelli, C., Stanley, J. A., Garrick, D., Sharpe, J. A., Ayyub, H., Wood, W. G., et al. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.* 34, 157–165. doi: 10.1038/ng1157

Uesaka, M., Nishimura, O., Go, Y., Nakashima, K., Agata, K., and Imamura, T. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC Genomics* 15:35. doi: 10.1186/1471-2164-15-35

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R., and Makalowska, I. (2004). Mammalian overlapping genes: the comparative perspective. *Genome Res.* 14, 280–286. doi: 10.1101/gr.1590904

Verona, R. I., Mann, M. R., and Bartolomei, M. S. (2003). Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annu. Rev. Cell Dev. Biol.* 19, 237–259. doi: 10.1146/annurev.cellbio.19.111401.092717

Wang, L., Jiang, N., Wang, L., Fang, O., Leach, L. J., Hu, X., et al. (2014). 3′ Untranslated regions mediate transcriptional interference between convergent genes both locally and ectopically in *Saccharomyces cerevisiae*. *PLOS Genet.* 10:e1004021. doi: 10.1371/journal.pgen.1004021

Wang, Q., and Carmichael, G. G. (2004). Effects of length and location on the cellular response to double-stranded RNA. *Microbiol. Mol. Biol. Rev.* 68, 432–452. doi: 10.1128/MMBR.68.3.432-452.2004

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453, 539–543. doi: 10.1038/nature06908

Wegmann, D., Dupanloup, I., and Excoffier, L. (2008). Width of gene expression profile drives alternative splicing. *PLOS ONE* 3:e3587. doi: 10.1371/journal.pone.0003587

Weinberg, M. S., and Morris, K. V. (2016). Transcriptional gene silencing in humans. *Nucleic Acids Res.* 44, 6505–6517. doi: 10.1093/nar/gkw139

Werner, A., Carlile, M., and Swan, D. (2009). What do natural antisense transcripts regulate? *RNA Biol.* 6, 43–48.

Werner, A., Cockell, S., Falconer, J., Carlile, M., Alnumeir, S., and Robinson, J. (2014). Contribution of natural antisense transcription to an endogenous siRNA signature in human cells. *BMC Genomics* 15:19. doi: 10.1186/1471-2164-15-19

Werner, A., Piatek, M. J., and Mattick, J. S. (2015). Transpositional shuffling and quality control in male germ cells to enhance evolution of complex organisms. *Ann. N. Y. Acad. Sci.* 1341, 156–163. doi: 10.1111/nyas.12608

Werner, A., Schmutzler, G., Carlile, M., Miles, C. G., and Peters, H. (2007). Expression profiling of antisense transcripts on DNA arrays. *Physiol. Genomics* 28, 294–300. doi: 10.1152/physiolgenomics.00127.2006

Werner, A., and Swan, D. (2010). What are natural antisense transcripts good for? *Biochem. Soc. Trans.* 38, 1144–1149. doi: 10.1042/BST0381144

Wianny, F., and Zernicka-Goetz, M. (2000). Specific interference with gene function by double-stranded RNA in early mouse development. *Nat. Cell Biol.* 2, 70–75. doi: 10.1038/35000016

Wight, M., and Werner, A. (2013). The functions of natural antisense transcripts. *Essays Biochem.* 54, 91–101. doi: 10.1042/bse0540091

Wood, E. J., Chin-Inmanu, K., Jia, H., and Lipovich, L. (2013). Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse. *Front. Genet.* 4:183. doi: 10.3389/fgene.2013.00183

Wyler, E., Menegatti, J., Franke, V., Kocks, C., Boltengagen, A., Hennig, T., et al. (2017). Widespread activation of antisense transcription of the host genome during herpes simplex virus 1 infection. *Genome Biol.* 18:209. doi: 10.1186/s13059-017-1329-5

Xia, J., Joyce, C. E., Bowcock, A. M., and Zhang, W. (2013). Noncanonical microRNAs and endogenous siRNAs in normal and psoriatic human skin. *Hum. Mol. Genet.* 22, 737–748. doi: 10.1093/hmg/dds481

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A. P., et al. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451, 202–206. doi: 10.1038/nature06468

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 05 December 2017
doi: 10.3389/fmicb.2017.02395

Viral tRNA Mimicry from a Biocommunicative Perspective

*Ascensión Ariza-Mateos[1,2] and Jordi Gómez[1,3]**

[1] Laboratory of RNA Archaeology, Instituto de Parasitología y Biomedicina "López Neyra" (Consejo Superior de Investigaciones Científicas), Granada, Spain, [2] Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Consejo Superior de Investigaciones Científicas (CSIC), Campus de Cantoblanco, Madrid, Spain, [3] Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Madrid, Spain

RNA viruses have very small genomes which limits the functions they can encode. One of the strategies employed by these viruses is to mimic key factors of the host cell so they can take advantage of the interactions and activities these factors typically participate in. The viral RNA genome itself was first observed to mimic cellular tRNA over 40 years ago. Since then researchers have confirmed that distinct families of RNA viruses are accessible to a battery of cellular factors involved in tRNA-related activities. Recently, potential tRNA-like structures have been detected within the sequences of a 100 mRNAs taken from human cells, one of these being the host defense interferon-alpha mRNA; these are then additional to the examples found in bacterial and yeast mRNAs. The mimetic relationship between tRNA, cellular mRNA, and viral RNA is the central focus of two considerations described below. These are subsequently used as a preface for a final hypothesis drawing on concepts relating to mimicry from the social sciences and humanities, such as power relations and creativity. Firstly, the presence of tRNA-like structures in mRNAs indicates that the viral tRNA-like signal could be mimicking tRNA-like elements that are contextualized by the specific carrier mRNAs, rather than, or in addition to, the tRNA itself, which would significantly increase the number of potential semiotic relations mediated by the viral signals. Secondly, and in particular, mimicking a host defense mRNA could be considered a potential new viral strategy for survival. Finally, we propose that mRNA's mimicry of tRNA could be indicative of an ancestral intracellular conflict in which species of mRNAs invaded the cell, but from within. As the meaning of the mimetic signal depends on the context, in this case, the conflict that arises when the viral signal enters the cell can change the meaning of the mRNAs' internal tRNA-like signals, from their current significance to that they had in the distant past.

Keywords: tRNA-mimics, ceRNA, biosemiotics, biocommunication, RNase P, code, HCV, cetRNA

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria

Reviewed by:
Miguel Angel Martinez,
IrsiCaixa, Spain
Antonio Mas,
Universidad de Castilla-La Mancha, Spain

***Correspondence:**
Jordi Gómez
jgomez@ipb.csic.es

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *07 November 2017*
Accepted: *20 November 2017*
Published: *05 December 2017*

Citation:
Ariza-Mateos A and Gómez J (2017) Viral tRNA Mimicry from a Biocommunicative Perspective. Front. Microbiol. 8:2395. doi: 10.3389/fmicb.2017.02395

INTRODUCTION

In general terms, a mimetic system involves the interaction of three agents: the mimic that simulates the signals or features of a second agent, referred to as the model, in order to confuse a third entity, the operator. This confusion is in some way beneficial for the mimic (Pasteur, 1995). Mimetic similarities are only considered to be those that occur where the carrying agents have no common origin. The most widespread line of enquiry related to mimicry is based on Darwinian

evolution, according to which, mimicry comes about through selective forces that favor false recognition, although this not always so, for example in a case where the variety of possible forms is limited (Maran, 2017). Other research emphasizes the communicative aspect: for Wolfgang Wicker, mimicry is based on the relationship of superficial similarity between organisms, and this relationship enables a flow of information (Wickler, 1998). This concept is rectified by Timo Maran, who places the emphasis on the signal. The important thing for this latter author is the similarity between the messages or signals emitted by the distinct organisms (which normally belong to different species); it is this signal that has mimetic value (Maran, 2017). This interpretative view of mimicry focuses less on the evolution of the mimicry itself and more on the different roles played by the various actors in the mimetic signal in an ecological context **Figure 1A**.

Mimicry was first studied in the 19th century, with research into butterflies and other organisms, but by the second half of the 20th century it had reached the molecular realm (Maran, 2017). A considerable number of biological activities have been described when it comes to molecular mimicry, many of which play important roles in viral reproduction, viral pathogenesis, and autoimmunity (Murphy, 2001; Alcamí, 2003; Christen et al., 2010; Drayman et al., 2013; Kropp et al., 2014; Oldstone, 2014). In viruses, many mimetic activities are achieved thanks to the structural and functional similarity between viral and cellular proteins. This class of mimicry is strategically valuable to a virus as it helps it to evade action by the host's immune system. Various viral products have been described that imitate different defensive factors, such as cytokine cell receptors (Murphy, 2001; Alcamí, 2003), and even factors that are not directly related to antiviral defense, such as the mimicry of protein histones that end up influencing the antiviral gene expression (Marazzi et al., 2012). Another important field of study is viral mimicry based on RNA, with particular emphasis on the structural mimicry of the transfer RNA molecule (tRNA). Viruses from very diverse families and genera adopt tRNA-like structures in their genomes, through which they acquire a wide range of molecular cell functions (Springer et al., 1998).

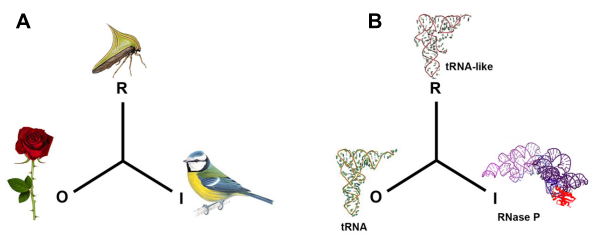


FIGURE 1 | Mimetic relationships. In biosemiotic terms and according to the sign system of Peirce, the mimetic signal corresponds to the representamen (R), the model to the object (O) and the operator to the Interpretant (I) (Kilstrup, 2015). In **(A)**, the plant's thorn is the Object (O), the thorn bug is the Representamen (R) and the predatory bird that the prey eludes is the Interpretant (I). In **(B)**, a molecular example: the $tRNA^{Phe}$ is considered to be the Object (O), the 3′ untranslated end of the genome of plant virus TYMV is the representamen (R), while the interpretant (I) is the tRNA precursor processing enzyme RNase P.

HISTORICAL FRAMEWORK

At the beginning of the 1970s, biologists already knew about the fundamental aspects of the tRNA molecule structure and the way it participates in a code, which we know as the genetic code, as it is the adaptor that operates just at the interface between the polynucleotide sequence and the amino acid sequence. In addition, the complete sequence of $tRNA^{Ala}$ had been elucidated, its clover leaf secondary structure was known (Holley et al., 1965), and there was a three-dimensional model in the form of a boomerang (Witz, 2003), very similar to the L-shape that would be determined years later and which is the standard form of tRNA (Kim et al., 1973). The code was accepted by the scientific community, but only as a metaphor, under the assumption that it would eventually be reduced to a set of favored interactions between the amino acids and their corresponding tRNAs. Surprising, as highlighted by Barbieri, until only a few years ago the genetic code would have been the only biological code that would have been present throughout the history of life on Earth. This had to be so, because the concept of a code is related to the correspondence between signals based on convention and that was something thematically alien to physics and chemistry; unless the codes were considered a way of restricting the different possibilities of the information (Barbieri, 2009). In that case, dealing with the information from the molecular signals represented by nitrogenous bases, which distinguish nucleotides from one another, whether taken one by one, in the form of a triplet in the case of the codons, or longer sequences that constitute the genes, became possible with the Probabilistic Theory of Information (Shannon and Weaver, 1949; Brillouin, 1968). In this theory, the signs are reduced to *bits*, in other words, certain quantities of information devoid of qualitative meaning, which represent a probabilistic value. This theoretical framework has facilitated some work in virology, such as the possibility of describing and comparing the quantity of information contained in viral populations (Wolinsky et al., 1996; Cabot et al., 2000) and evaluating the divergence between viral sequences in the course of evolution (Pan and Deem, 2011). Nevertheless, it should be noted that Information Theory, as a probabilistic theory, is only possible in a closed information space where it is possible to anticipate all the possibilities of this information, equivalent to predicting all the events that could occur or processes that could be carried out. Within this mathematical framework, and clearly influenced by Chomsky (Searls, 2002), Manfred Eigen searched in the rules of the molecular language (syntax) for the element to which all genetic information could be reduced. The shortcomings of these approaches are described in Hoffmeyer and Emmeche (1991) and Witzany (1995).

Some authors have proposed that the responsibility for this treatment of biological data lies in the context in which Information Theory flourished, i.e., World War II, as reviewed

in Hoffmeyer and Emmeche (1991). The way information is dealt with in this theory may be completely valid in a space of military communication saturated with power relations (Foucault, 2001) where interpretation is non-existent, the language being that of a disciplined soldier or automaton. It is also likely that it has something to do with biology's inferiority complex with respect to physics and chemistry, so that biology, in its longing to be perceived as a science of objectivity equal to that of physics and chemistry (Mayr, 1982), has surrounded the question of molecular language with mathematical theory. Recent science has also been strongly influenced by science philosophy, principally logic and mathematics (Witzany, 2010). This has been and continues to be the dominant trend, which has most recently been extended with systems biology.

Influenced by this prejudice, and in this context, the discovery of the genetic code did not lead the way to the study of communication and cell control based on a real molecular language (Witzany and Baluska, 2012). The subject of tRNA mimicry has been developed within this restrictive framework, as has the molecular biology of nucleic acids. Nevertheless, some authors and schools maintain that the reductionist and quantitative treatment of the information is not sufficiently explanatory for molecular biology. These authors, from different perspectives, have proposed: understanding the cell as a duality of digital and analog codes (Hoffmeyer and Emmeche, 1991); that information always is interpretation (Hoffmeyer and Emmeche, 1991); that at the molecular and cellular level there are more codes than just the genetic code (Barbieri, 2008); that even this latter has not been able to codify to itself (Witzany and Baluska, 2012), but requires the participation of a multiplicity of molecular agents; with conflictive and collaborative social relationships at the molecular level (Gómez and Cacho, 2001; Villarreal, 2009, 2015; Villarreal and Witzany, 2013; Witzany, 2014). In short, it is the new areas of knowledge that are interested, firstly, in content communication via the interpretation of signals and molecular codes, an area known as biosemiotics, and secondly, in the fact that this interpretation is context dependent (pragmatic) and classified as biocommunication (Witzany, 2010). We must emphasize that in the latter, the context refers not only to environmental characteristics but also the history of the factors that participate in the interpretation, their identity and particularly their social relationships.

There are some excellent reviews of viral tRNA mimicry in the literature that include the most significant results and interpretations in the field (Giegé et al., 1993, 1998; Springer et al., 1998; Giegé, 2008; Dreher, 2009, 2010). Here we outline results found in the classic works on tRNA mimicry that could be compatible with the biocommunication theory. They include work where the identification of tRNA-like structures has been made directly using tRNA metabolism enzymes, and defines tRNA mimics in an operative way (Springer et al., 1998). We highlight some interpretations these results would have made possible, and also their fall into disuse due to the failure to study the biocommunicative facet in favor of the structural aspect; subsequently, we propose a hypothetical expansion of the subject in that direction.

TRACING VIRAL tRNA MIMICS USING tRNA MODIFYING ENZYMES

tRNA-Like Structures in the 3′ Region of Viral mRNAs

Despite the restrictive way biological information has generally been dealt with, as described above, according to which the genetic code could be treated as a machine code, largely due to the specificity of the aminoacyl-tRNA synthetases, the fortuitous discovery was made of the incorporation reaction of the valine amino acid into the 3′ position of the Turnip Yellow Mosaic Virus (TYMV) genomic RNA (Pinck et al., 1970). It was seen that cellular $tRNA^{val}$ did not participate in this reaction, but it was the viral RNA in the 3′ region of the genome that was being modified (Pinck et al., 1970). This discovery turned out not to be exclusive to the $tRNA^{val}$ and RNA of the TYMV, soon other examples were found among the family of aminoacil-tRNA synthetases, in particular those specific for histidine (Sela, 1972) and tyrosine (Hall et al., 1972) as modifier enzymes and other plant virus RNA substrates, such as Brome Mosaic Virus (BMV) (Hall et al., 1972) and Tobacco Mosaic Virus (TMV) (Sela, 1972). This first set of results shows that one of the enzyme families that is key to decoding the genetic code (known as the operative code) (Schimmel et al., 1993; Ribas de Pouplana and Schimmel, 2001), the aminoacyl-tRNA synthetases, can recognize as their own a substrate with a different sequence from the tRNA sequence. Specific and erroneous recognition of a signal is the paradox indicating that the number of molecular alternatives each aminoacyl-tRNA synthetase can recognize is not known. In other words, the "operative" code is open —it is not a machine code.

In a second study group, factors other than the aminoacyl-tRNA synthetases were discovered, capable of recognizing and modifying the 3′ end of the viral RNA. These factors include: the ACC-tRNA nucleotidyl-transferase (Litvak et al., 1970), responsible for incorporating adenine and cytosine (CCA sequence) at the 3′ end of the tRNA; the translation elongation factor in prokaryotes; and ribonuclease P (RNase P) activity (Prochiantz and Haenni, 1973; Guerrier-Takada et al., 1988), which is able to specifically process the tRNA precursor (pre-tRNA), cutting between two nucleotides that separate the leading 5′ region of pre-tRNA and mature tRNA (Robertson et al., 1972). This second set of results extends this specific and erroneous recognition of the 3′ region of the viral mRNA to a variety of very different cell translation enzymes and factors (Giegé, 2008). This second group of results also indicates that the forms these viral RNAs adopt in the 3′ region are communicative elements, as they reveal a great deal of indetermination and ambiguity, the significance of which necessitates their maturation by a contextualized interpreter.

In a third group of results, it was discovered that the previously mentioned enzymes are able to modify or interact with other

elements of non-viral origin, related to the processing and control of gene expression in cellular mRNA. Examples are the leader sequence of the mRNA of the Threonyl-tRNA synthetase (ThrS) (Springer et al., 1989), the pre-mRNA of several aminoacyl-tRNA synthetases (Guo and Lambowitz, 1992), and the bacterial 10Sa RNA (Komine et al., 1994) [subsequently known as transfer-messenger RNA (tmRNA)]. These data could have at least suggested the possible existence of a context of intracellular tRNA-like signals and recognition factors in which viral tRNA-like signals competed.

Indeed, this entire set of results could have displaced the approach of tRNA-like motifs from a structural and syntactic analysis to a tRNA-like mediated real language, that is the characterization of sequences and structural rules that govern the assignment of aa into each tRNA-like motif to an approach that would have shown the overall contextual aspect that affects the interpretive agents and that definitively gives meaning to the signals (Witzany and Baluska, 2012). If we had continued systematically with this experimental identification of tRNA-like elements with cellular enzymes, such as those described above, we could have compiled a wider collection of correspondences between viral tRNA-like structures and cell recognition factors that would perhaps have enabled the outlining of other possible interrelated codes.

On the contrary, progress in experimental techniques for identifying new tRNA-like elements underwent a change that made this approach obsolete. After the 1980s, techniques for retrotranscribing RNA into cDNA, and subsequent DNA sequencing, largely replaced the direct sequencing of RNA by fingerprinting, a very delicate and painstaking technique (Branch et al., 1989). DNA sequencing became widely available in molecular biology labs, and structural prediction and 3D modeling software began to be developed (Dumas et al., 1987). This changed the way of looking for new tRNA-like motifs (Fechter et al., 2001). The first step was to determine the sequence of an RNA motif of interest and then to check *in silico* if the structure was indeed folded in the form of a clover leaf, and only then to experimentally test whether it really could be modified by a tRNA metabolic enzyme (Konarska and Sharp, 1990; Michel and Westhof, 1990; Pilipenko et al., 1992; Felden et al., 1994; Fechter et al., 2001; Lukavsky et al., 2003). Therefore, although in the 1970s the identification of a tRNA-like motif centered on its recognition through an enzyme of tRNA metabolism (Springer et al., 1989), the fundamental criterion became that of structural similarity according to a human observer and hence the emphasis shifted to the molecular architecture. The development of this new pathway involved an increasing understanding of the different intramolecular interactions of RNA and improved 3D models, as well as the determination of the structures using methods with ever greater resolution (Lukavsky et al., 2003; Boehringer et al., 2005). The objective was achieved with the determination of the TYMV tRNA-like structure using X-ray diffraction (Colussi et al., 2014). This trend represented a departure from the operative search of tRNA-mimic motifs, which would have brought us closer to the signal-interpreter relationship of the language. Structural studies of plant viral tRNA mimicry have identified in tRNA-likes the determinants for specific recognition by Aa-tRNA synthetases and other enzymes (a syntactic level), as well as confirming the similarity with standard tRNA. The next experimental step should have been to demonstrate that the aminoacylated tRNAs serve as Aa donors in the synthesis of viral proteins. In this way, the correlation between structural and functional similarity (semantic level) would have been expanded. This idea has persisted in the field for decades (Haenni et al., 1973; Barends et al., 2003), without success (Matsuda and Dreher, 2006, 2007). The idea would be pertinent if the information space that includes the genetic code were closed, but through the various examples we have looked at we can see the opposite, that the genetic code is one of the communicative possibilities of the factors that manage the information contained in tRNA and tRNA-like structures. This is not only true in the exceptional case of a viral infection, but also in operon regulation in a healthy cell.

tRNA-Like Structures in the 5′ Region of Viral mRNAs

In 2002, more than 30 years after the identification of the tRNA-like motif in the 3′ region of the TYMV genome, it was discovered that human RNase P recognizes and specifically cuts the 5′ region of RNA in the hepatitis C virus (HCV) *in vitro*, a finding that was presented and discussed in relation to the tRNA mimicry of plant RNA viruses (Nadal et al., 2003). Later, it was also observed that this motif could be recognized by the RNase P ribozyme of the cyanobacteria *Synechocystis* sp. under high salinity conditions (Sabariegos et al., 2004). The presence of tRNA-like structures in the 5′ region of viral mRNA is generalized in animal pestiviruses that are phylogenetically related to HCV, such as bovine viral diarrhea virus (BVDV), and classical swine fever virus (CSFV) (Lyons and Robertson, 2003); as well as in the RNA of unrelated viruses, like the cricket paralysis virus (CrPV) (Lyons and Robertson, 2003), picornavirus food and mouth disease virus (Serrano et al., 2007), and polio virus (Andreev et al., 2012). Although there is no similarity in either the nucleotide sequence or secondary structure between the RNA in hepacivirus, picornavirus, and CrPV, these motifs are found in the 5′ region of coding sequences in these viruses, known as internal ribosome entry sites (IRES) (Lozano and Martínez-Salas, 2015). However, in HCV an additional motif was also identified between the region coding for structural and non-structural proteins (Nadal et al., 2003).

Thus, the identification the tRNA-like structure in the 5′ region of viral mRNAs began under the same operative definition as that used in the work to characterize the tRNA-like structure in the 3′ region of plant viruses. However, unlike the 3′ region of RNA, which can be examined by multiple specific and distinct enzymatic activities, the enzymatic identification of a tRNA-like structure located within a longer RNA strand is very limited. The possibilities are limited to RNase P, although in certain specific cases another endonuclease, RNase Z (Wilusz et al., 2008), and an enzyme that modifies tRNA nucleotides have been used successfully (Baumstark and Ahlquist, 2001).

But again, the pathway leading to the communicative aspect would be relatively blocked by two issues. The first relates to the

fact that the subject is strongly influenced by the structuralist idea that it is necessary to determine the RNA structure before deciding whether this may be a tRNA mimetic structure. Therefore, cryo-electron microscopy led to the recognition of the L-shaped tertiary structure of the HCV IRES RNA (Boehringer et al., 2005) and the ribosome-bound CrPV (Fernandez et al., 2014). Subsequently, the resolution of the crystal structure of HCV RNA in the zone recognized by RNase P (Piron et al., 2005) indicated the presence of not a single, but rather a double pseudoknot (Berry et al., 2011). Without a doubt, this approach enables the mimetic resemblance to be visualized and partly explained, but it introduces the matter of the human observer (Maran, 2017). The second issue, which we detail below, relates to the specificity of the RNase P activity. It is a subject that requires particular attention and which we will look at below.

RNase P Recognition Specificity

Various prestigious and influential groups in the field have assessed the specificity of RNase P and come to opposing conclusions. Whereas for some groups it represents a tool with therapeutic potential exactly because of its recognition specificity (Altman, 1995; Yuan and Altman, 1995), for others this represents non-specific activity because it can cut mRNAs in positions where bioinformatic folding indicates it is in a single strand form (Marvin et al., 2011).

Originally, and for many years, the only biological activity attributed to RNase P was the processing of pre-tRNA (Robertson et al., 1972; Altman, 1975; Jarrous and Reiner, 2007). Taking pre-tRNA as a model, the sequence and structure requirements of the RNA substrates of both *Escherichia coli* and human RNase P were carefully studied (Forster and Altman, 1990; Yuan and Altman, 1995; Liu and Altman, 1996; Werner et al., 1997, 1998). The cutting determinants for both endonucleases were found to be similar but not identical. For the RNase P of *E. coli* it is necessary for the T-stem loop next to the acceptor stem to terminate with the sequence –CCA (Liu and Altman, 1996), whereas for the human one, it is necessary to have a sequence of at least one nt between the T-stem and acceptor stem (Werner et al., 1997). Studies where the RNase P RNA domain is used as an antiviral agent against various cytomegalovirus mRNAs in culture, confirm that RNase P expressed in cytoplasm, besides acting as an antiviral agent, is not toxic to cells (Liu and Altman, 1995; Jiang et al., 2012).

Taken together, the results demonstrated that the RNase P was able to cut other substrates, different from tRNA but with which they shared structural similarities, like the RNA of the TYMV (Guerrier-Takada et al., 1988). Nevertheless, it was also observed that the RNase P was able to process substrates that were structurally different from tRNA, particularly when the ribozyme occurred in the presence of the protein subunit, as in the case of 4,5S RNA (Peck-Miller and Altman, 1991). This conferred to the protein the value of extending enzyme's substrate recognition, allowing it to cleave single stranded RNA substrates as small as 5 nt (Hansen et al., 2001). In contrast, other results gave clues as to the requirements for substrate recognition by RNase P. One study found that RNase P processed the ribosomal RNA in yeast (Chamberlain et al., 1996); due to the similarity of the genomic organization of the rRNA precursors in yeast and *E. coli*, where the pre-tRNA are found in the same precursor molecule that contains the rRNA, a phylogenetic relationship explaining this and other cases could not be ruled out. In a completely different vein, there are studies where RNase P processes RNAs that can assume different configurations and whose transition is regulated by small metabolites (riboswitches) (Altman et al., 2005; Seif and Altman, 2008), in such a way that the substrate form is not the most thermodynamically stable (Altman et al., 2005). A systematic study in our laboratory has enabled us to evaluate the number of random mutations that are necessary to be introduced into a population of a standard RNA substrate for RNase P, RNase III and the binding to microRNA-122, and the binding of subunit 40S of the ribosome, and which halves these activities and interactions. It turns out that this number is similar for RNase P, the binding to 40S, and the binding to miR-122, and it is greater for RNase III. Therefore, if it is possible for RNase P to recognize multiple structures, its specificity is greater than that of RNase III and similar to that of the binding of a microRNA to its mRNA substrate and the binding of the HCV IRES to the 40S subunit of the ribosome (Prieto-Vega, 2016).

Based on the evidence presented above, we conclude that RNase P is no more non-specific than other factors for which we are confident of their specificity, and that if a substrate is processed by RNase P, the idea that there must be a structural similarity between this RNA motif and the tRNA, in the view of the human observer, is not correct.

tRNA-LIKE STRUCTURES INSIDE CELLULAR mRNAs

Up to this point, tRNA mimicry been has limited to the interpretations of tRNA as a model (**Figure 1B**). A new study was undertaken to verify whether among the population of hepatic mRNA species there were any that were similar to the tRNA-like structure identified in the 5′ region of HCV RNA (Díaz-Toledano and Gómez, 2015). It was used as recognition factors human RNase P (Bartkiewicz et al., 1989) and the RNase P ribozyme of the cyanobacteria *Synechocystis* sp (Vioque, 1992; Pascual and Vioque, 1999). It was deduced that more than 100 species of mRNA were a substrate of the activity *in vitro*. Of these, three messengers were analyzed separately. These messengers were related to distinct historical periods of life: that of interferon alpha (IFNA), which originated in the vertebrates; that of histone H2A, related to the "DNA world"; and ribosomal protein S9 from the small ribosome subunit, related to the "RNA world." It was observed that the three competed specifically with the pre-tRNA in a standard RNase P processing reaction, that the chemistry of the newly generated 3′OH and 5′PO_3 ends confirmed it was RNase P, and that the minimum size fragment necessary to sustain the cleavage reaction was around a minimum of 120 nts, sufficient to house a tRNA-like structure. In addition, they were recognized by the *Synechocystis sp.* RNAse P ribozyme. The conclusion here should be that mRNAs with tRNA properties have been identified and characterized, thus expanding the repertoire of motifs that viral RNAs can mimic to host mRNAs.

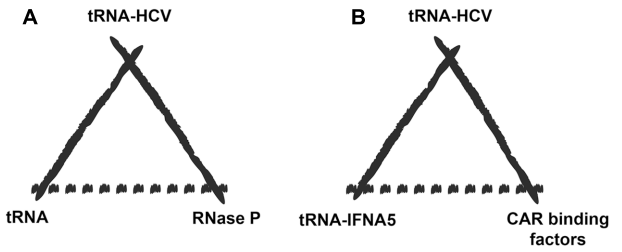


FIGURE 2 | Extended biosemiotic relationships of tRNA mimics. **(A)** A particular example of tRNA mimicry would be the case of hepatitis C virus mRNA carrying a tRNA mimic element (representamen) that is specifically recognized by RNase P (the interpretant), which normally processes the tRNA precursor molecule (the model). **(B)** Potential switch in the tRNA mimetic relationship after the discovery of a tRNA-mimicking element in IFNA5 mRNA. The model for the viral RNA to mimic could now be a cellular mRNA rather than the tRNA molecule itself (in this case the mRNA for liver IFNA5). The deceived interpreter would be those factors that naturally interact with the IFNA5 mRNA-tRNA-mimicking region, such as CAR binding factors.

It has been proposed that each element may mimic more than one structure class of cellular substrate and may carry out more than one function (Alexander-Brett and Fremont, 2017). Thus, a viral tRNA-like mimicking a structure within an mRNA does not necessarily mean that it will not continue to mimic tRNA for other functions, but simply that it expands the number of functions which this structure may support.

In biosemiotic terms, this supposes a paradigm change for the interpretation of the role of viral tRNA mimics due to the possibility that cellular mRNA species may become the "model" for the viral mRNA to mimic (Díaz-Toledano and Gómez, 2015) (**Figure 2**).

On the other hand, the mRNA that codes for the viperin protein, which is also related to antiviral defense, had been identified as a potential substrate for RNase MRP and RNase P (Mattijssen et al., 2011). Although the authors cannot discern which of the two RNases is responsible, this result could indicate the existence of a second family of antiviral defense mRNAs carrying tRNA-like elements.

Structure and Possible Function of IFNA mRNA tRNA-Like Regions

The proposed tRNA-like structure of IFNA mRNA is shown in (**Figure 3**). With respect to the recognition requirements of RNase P, domain 3 would represent the T-stem and T-loop of tRNA and domain 4 the acceptor stem (Díaz-Toledano and Gómez, 2015), without the need for the presence of the pseudoknots located around the helix junction for recognition by RNase P, as is characteristic of plant virus tRNAs. The pseudoknots in IFN5A might be responsible for the adequate orientation of the branches, as in the tRNA-like structure in the tobacco mosaic virus, where the helix junction is integrated within a pseudoknot (Felden et al., 1996).

The comparison of the tRNA-like structure in the 5′ region of HCV and IFNA mRNAs shows that there are remarkable similarities with regard to the polarity of the processing determinants, the sequence, the secondary structure, the presence of two pseudoknots and their location in the mRNA -described in Díaz-Toledano and Gómez (2015). The smallest fragment of IFNA 5required for RNase P processing (positions 242–424) coincides significantly with a region that is able to interact with nuclear export and cytoplasmic stabilization factors (TREX) (Lei et al., 2011). Functional CAR signals involve primary sequence elements (one to four) 10 nts in length, referred to as CAR-E (Lei et al., 2013). Each CAR-E is located at the extremities of domains II and IV of the IFNA tRNA-like motif, in positions 264–273, and 396–405, as depicted in the yellow boxes in **Figure 3**. Apart from this positional symmetry, CAR-E adopts an equivalent secondary structure.

This could be thought of as selective interference of HCV with the expression of IFNA-5 in the liver as a result of direct competition for factors eventually facilitating cytoplasmic accumulation of IFN5A mRNA (Díaz-Toledano and Gómez, 2015). In fact, the INFA signaling pathway is utilized in cell defense against HCV, although the IFNA5 liver-specific sub-type of IFNA mRNA disappears from the liver several days after infection (Castelruiz et al., 1999). Alternatively, or additionally, HCV RNA may benefit from using the cell factor TREX.

Finally, the detection of mimetic elements that are recognized by RNase P in cellular mRNAs, and in particular in IFNA mRNA, leads to a new paradigm. This provides a new way of looking at the function of a variety of viral tRNA-like motifs, as this type of structural mimicry might be related to specific host mRNA species rather than, or in addition to, tRNA itself, an issue which for the moment we can only deal with hypothetically.

WHAT CHANGES WHEN tRNA-LIKE-mRNA IS USED INSTEAD OF tRNA AS A MODEL FOR VIRAL MIMESIS?

Evolutionary Context: Origin of tRNA-Like Structures in Cellular mRNAs

The origins of tRNA go back to the "RNA world" (Rodin et al., 2011; Di Giulio, 2012; Caetano-Anollés and Sun, 2014; Caetano-Anollés and Caetano-Anollés, 2016). In the case of tRNA-like structures in mRNAs, theoretical studies along with certain experimental results seem to insist on an idea first suggested by Eigen and Winkler-Oswatitsch (1981). Eigen asked himself if ancestral tRNA could have worked like mRNA in the RNA world. His idea was that both the populations of RNA capable of aminoacylating and functioning as adapters, such as those that act as messengers, could have been selected from a single population of replicative RNA quasispecies (Eigen and Biebricher, 1988; Domingo, 2007; Mas et al., 2010), positive and negative strands. He identified a series of requirements that would be necessary for this to occur: symmetry between positive and negative chains, a high number of G:C pairings, and small-sized RNAs, among other features, which would guarantee stability and structural complementarity between the two types of molecules. Since then, different groups have suggested and supported this hypothesis,

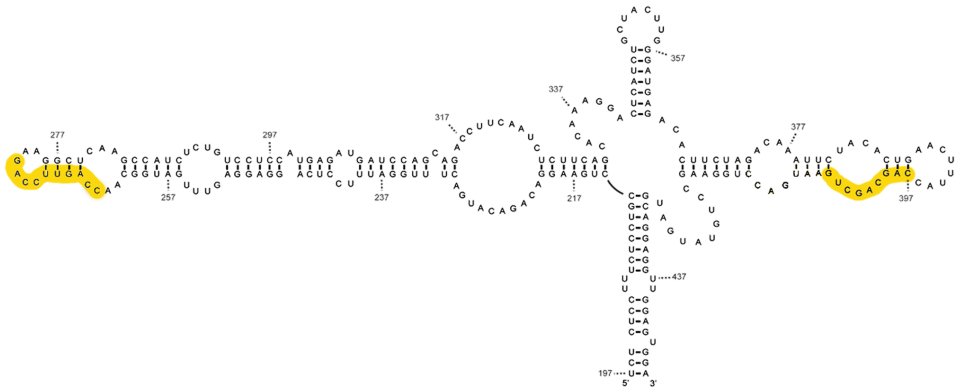


FIGURE 3 | Mimetic structure of the tRNA-like of IFN5A. This image shows the secondary structure of the tRNA-type element of alpha interferon mRNA subtype 5 determined by various chemical and enzymatic methods. The CAR-E sequences are highlighted in yellow.

based on different observations or proposing distinct models (Ohnishi et al., 2002, 2005; Bernhardt and Tate, 2010). The two most recent proposals are that either the original substrates of mRNA were aminoacylated proto-tRNA-like structures with regions of self-complementation (Di Giulio, 2015), or that they were aminoacylated proto-tRNA-like structures with the capacity to take on different configurations (de Farias et al., 2016). These original substrates would have been concatenated to form mRNA. An alternative, and also recent hypothesis is that a population of molecules that already contained tRNA-like and mRNA-like domains (proto-tmRNA) acted as a common ancestor of molecules that elongated and eventually evolved into tRNAs and mRNAs (Macé and Gillet, 2016).

A study by our lab, in which we evaluated the sensitivity of the human hepatic mRNA population to human RNase P, reveals interesting data relating to these proposals on the origin of mRNA. The competition results with Poly-rA, pre-tRNA, and human liver mRNA, in a standard pre-tRNA RNase P processing assay, indicate that the mRNA population competes well with the natural substrate processing of RNase P (Figures 1A–C) in Díaz-Toledano and Gómez (2015). At the molar level, it is six times better competitor than the pre-tRNA itself. The results from the competition are very similar to those obtained by Dr. Engelke's group using RNase P and yeast mRNAs Figures 2B,C in Marvin et al. (2011), although that study did not provide data on the competition with the pre-tRNA. However, direct examination of the products from the mRNA population after incubation with human RNase P did not reflect processing, which was also in line with studies published in the 1970s that had tested the cleavage of hnRNA using RNase P-enriched extracts (Ferrari et al., 1980; Kole and Altman, 1981), and did not observe processing. When instead of examining the mRNA population, human or yeast mRNA species were tested separately (Marvin et al., 2011; Díaz-Toledano and Gómez, 2015), it was seen that these mRNAs were sensitive to being processed by RNase P in specific positions, but there was a very low cutting effectiveness. It is possible the internal structures of mRNA that are able to inhibit the enzyme in a competition test, are nevertheless not sufficiently stable or do not present an optimal configuration to be actively processed. These results could support the hypothesis of the origin of mRNA based on concatenations of tRNA-like molecules, where their structures have been undergoing alterations since their origin. Specifically, the presence of RNase P activity from the "RNA world" to the present would have posed a danger for the proto-tRNA-like polymer because it would have facilitated its fragmentation, making it a selective force that would have tended to alter the proto-tRNA-like structures integrated into the messengers, favoring an imperfect or unstable similarity to tRNA, rather than resembling it perfectly.

Alternative explanations for the origin of tRNA-like structures within mRNA are that these structures have evolved from other forms in the mRNAs, or have been contributed by viral agents that subsequently colonized the genes of the mRNAs at different stages of life, taking advantage of the similarity of some of the tRNA properties. However, if so, these properties do not seem to be related to the tRNAs involved in protein synthesis, firstly, because they are found within a longer molecule that prevents aminoacylation, and principally, because in the examples of plant virus RNA where the tRNA occupies the 3′ position and is able to aminoacylate, this does not act as an Aa donor. Thus, in any case, it returns to the ancestral properties of tRNA.

In summary, the RNase P competition tests reinforce the idea that current mRNAs could have originated from the concatenation of proto-tRNA-likes or the elongation of the original transfer-messenger RNA-type molecules (tmRNAs), which have degenerated, but which in some cases would still contain vestiges of the proto-tRNA-like originals.

Environmental Context: tRNA-Like Structures in mRNAs

The presence of tRNA-like structural motifs in cellular mRNAs (tRNA-like-mRNA) extends the structural space of tRNA forms (Gilbert and Labuda, 1999; Wilusz et al., 2008). The molecules of the tRNA family are very similar to one other, they play a central role in the molecular biology of the cell, although they also participate in other pathways (Giegé, 2008; Maute et al.,

2014; Katz et al., 2016), and at a quantitative level this is the most abundant type of molecule; in contrast, tRNA-like-mRNA have distinct sequences, are found in smaller proportions, and the mRNAs that carry them participate in very diverse functions.

THE ARCHEOLOGICAL RECONSTRUCTION OF tRNA-MIMICRY

Communication in the World of RNA and tRNA-Like Structures

In the primordial world, certain forms of communication would have been increasingly necessary and complex as it was necessary to record the external and internal changes of a primitive cell, and to readjust the metabolic processes ever more precisely. This would have required signals, receptors, riboswitches (Nelson and Breaker, 2017), codes (Caetano-Anollés et al., 2013), and communicative networks that integrated and channelized the reception and response to those changes. The vehicles for these signals would have been small molecules derived from ribonucleotides (Nelson and Breaker, 2017) as well as from longer RNAs, such as stem-loops and proto-tRNA-likes (Witzany, 2014; Villarreal, 2015). This supposes a level of communication that had to follow a series of rules according to capacities and biochemical requirements, as well as the necessity to belong to some or other of the participatory consortiums (Witzany, 2017a). Some of these requirements could have been: (i) the Eigen limit (Eigen and Biebricher, 1988), as due to the high rate of mutation in primitive replication, the size a replicative RNA could maintain without losing information is small (50–100 nts); (ii) participation in a molecular consortium (Witzany, 2014; Villarreal, 2015), requiring the prior molecular recognition of itself which determines a language with a repetitive syntax (Witzany, 2017b); (iii) and finally, the relationship between sequence space and structural space (Schuster, 1997) would have degenerated, in such a way that multiple sequences would group into very few distinct structural folds, greatly favoring structural mimicry. All these requirements would favor an RNA world inhabited by mini-helices that either through duplication or accretion could have created an environment very rich in forms of proto-tRNA-likes (Tamura, 2011; Caetano-Anollés and Caetano-Anollés, 2016). These forms could have included the ancestors of both the aminoacylated tRNAs and the ribosomal peptidyl transferase center (de Farias et al., 2017) and the mRNA (de Farias et al., 2016).

The RNA world would have been the context in which those proto-tRNA-likes would have gone from being free to being embedded in mRNA during the concatenation process or via continuous growth, if we accept the hypotheses presented above. The concatenation would have probably followed the selection rules suggested for "conjoined RNAs" by Robertson and Neel (1999):"*in that two independent activities, each embodied in a separately evolved RNA are required to survive the conjunction or capture process that joins them together.*" In other words, in a primitive stage, the formation of each of the proto-mRNAs would have been selectively favored by the sum of the activities of the tRNA-like molecules that are bound together to form a longer RNA. These capabilities could be related to the interaction of factors that confer RNA stability, association with ribosomal subunits, sensitivity to conformational changes in the presence of small molecules, and so on. All these possibilities should have been conferred on the mRNAs with greater communicative capacity, therefore participating in increasingly complex sections in the original cell.

tRNA-Like Structures in mRNAs after the Development of Protein Synthesis

After the complete development of protein synthesis using the mRNA as a template, a new language was incorporated into the cell, whose syntax was mainly non-repetitive, and in which the semantic function was separated from the sequence of RNA bases by being carried by another material support, the protein. The latter allows for post-translational modifications of the proteins while the mRNA remains intact, producing native proteins, or for the RNA to be edited in various ways. All these changes enriched the communication, to which must also be added the enzymatic activities of proteins that greatly expand their communicative ability. Meanwhile, the contextual dimension did not stop expanding, and it did so in a very elaborate way, where the signals became immersed in positive and negative feedback processes, cascade amplification, molecular reorganization, editing, vacuolization, agglomeration, molecular addiction, and so on.

With all this transformation and the continuous selective pressure to increase translation effectiveness, the majority of the communicative capacities that the tRNA-like molecules would have contributed to mRNAs when integrating into these would have disappeared, others would have become obsolete when their primitive receptors vanished during the extinction of ribozymes and RNA-world forms, and others would eventually have become controlled but not completely extinguished. There are examples that provide certain support to this non-total-elimination, both in viral and cellular RNAs.

At present, the interaction of cellular mRNAs and the ribosome during translation initiation is mediated by mRNA signals and a battery of protein factors that are very specific to that process, including translation initiation factors and other proteins with more varied roles (Jackson et al., 2010). To initiate the translation of the second ORF of CrPV none of this is required (Jan and Sarnow, 2002). To do this, the virus contains an internal ribosome entry site (IRES) that imitates tRNA with its anticodon bound to a messenger when this is present in the P site of the ribosome, enabling a $tRNA^{Ala}$ to enter site A, and allowing translocation, even though the peptide bond has not formed (Spahn et al., 2004; Costantino et al., 2008). This happens both *in vitro* and in cells, although in this latter case it can also begin at the P site (Kamoshita et al., 2009).

Another example also related to the initiation of CrPV translation is the ability of the IRES of this eukaryotic virus to operate in the translation system of both eukaryotes and prokaryotes. The authors who determined the crystal structure

of the CrPV IRES bound to the bacterial ribosome, state that the CrPV IRES bridges billions of years of divergent evolution by being recognized by two different domains of life (Colussi et al., 2015). This finding provides new evidence on a proposal from the 1970s that indicated that at least some signals present in the mRNA, leading to recognition by the ribosome, should be common to both prokaryotes and eukaryotes, despite the signal incompatibility between the two translation systems. In these studies, using ribosomes from reticulocyte lysate, it was observed that the authentic translation initiation sites of the bacteriophage f1 mRNA could be protected from RNase A (Legon et al., 1977). In addition, it had been determined that eukaryotic viruses, such as polioviruses, could function in translation extracts from a prokaryotic origin (Rekosh et al., 1970).

We would like to emphasize that there is another series of evidence also based on tRNA-like signals that currently function as part of the cellular mRNA. Such is the case of plants, where the long distance transport of mRNAs between distinct tissues seems to be determined by tRNA-like structures in mRNAs (Zhang et al., 2016). Or the case of IFNA, where the enormous amount of coincidence between the region forming the cloverleaf structure, and the region that signals for TREF binding seems to indicate that the tRNA-like signal plays a role in the transport of the nucleus to the cytoplasm and participates in mRNA stability (Lei et al., 2011, 2013; Díaz-Toledano and Gómez, 2015). Also in bacteria, it has been known for decades that the leader region of Threonyl-tRNA synthetase mRNA (ThrRS), whose structural similarity to the $tRNA^{Thr}$ anticodon allows it to bind to the ThrS protein, reflects a functional activity in the expression control of this mRNA (Springer et al., 1998).

There is an additional example, which although not related to tRNA-like structures, does indicate that certain archaic RNA signals have not lost their primitive function (Jha et al., 2017). Normally in mammal cells, long poly-A sequences located upstream of the AUG initiator work against the translation, the opposite of what happens in translation in plants. Some poxviruses from mammal cells contain those poly-A stretches upstream of the AUG initiator and optimize protein synthesis in the infected cell, mimicking the translation situation in plants. This is achieved by the viral kinase that is responsible for phosphorylating the RACK1 protein, leaving it at a similar level of phosphorylation and electronic charge to that found in the small ribosomal subunit of plants.

Other evidence is negative, in other words, it is based on the potential activities of viral tRNA-like structures that do not occur, which we have referred to previously: RNase P does not cut into the tRNA-like structures of HCV mRNA *in vivo* (Piron et al., 2005), and neither do tRNA-like structures of plant viruses act as amino acid donors in translation (Matsuda and Dreher, 2006, 2007). This reveals cellular functions that prevent the canonical expression of these signals, for example, the compartmentalization of RNase P in the cell nucleus. This suggests that, probably, the communicative potential of tRNA-like structures embedded in cellular mRNAs are also under cellular control so they do not express themselves, or they do so under very specific conditions, or in a very specific sense.

HYPOTHETICAL COMMUNICATIVE MECHANISM INVOLVING tRNA-LIKE-mRNAs

It is well known that one of the elements involved in regulating the expression of mRNAs is mediated by microRNA binding to the mRNA, mainly in the 3′ UTR region, but also in the 5′ UTR region (Lytle et al., 2007). This binding usually inhibits the translation of cellular mRNA and can have various effects on viral mRNAs, including on their stability (Scheel et al., 2016). In recent years, a new form of communication has been revealed between mRNAs through these microRNAs. This communication is based on the existence of other RNAs that also have specific binding sites for microRNAs, including other long non-coding RNAs, circular and pseudogene RNAs, which compete with mRNAs in the binding of microRNAs affecting the translation rate of each of the messengers in the population. The wealth of microRNA signals and the topological ordering of their binding sites in messengers allow the existence of codes to be considered. This new way of communicating is known as competing endogenous RNAs (ceRNA) and has been denominated the "Rosetta Stone" of a hidden RNA language (Salmena et al., 2011; Tay et al., 2014). RNA viruses can also analogously participate in this type of communication. An example is the case of the microRNA specific to the human liver miR-122, which is necessary for the replication of the viral RNA. The binding of microRNA to viral RNA causes miR-122 sequestration from the cytoplasm (sponge effect) and alters the regulation of those cellular mRNAs on which miR-122 exerts its function (Luna et al., 2015). Recently, this type of communication has also been assigned to other microRNAs and a large group of diverse RNA viruses (Scheel et al., 2016).

Hypothetically, the communication mode of mRNAs carrying tRNA-like signals could be analogous to that of mRNAs with microRNA targets in ceRNA. On the one hand, the tRNA-like-mRNAs would act as specific targets for various cell factors that recognize tRNA, and on the other hand, the different forms related to tRNA (pre-tRNA, mature tRNA, tRNA-Aa, and fragments of mature tRNA (Gebetsberger et al., 2016) and SINE elements) would represent the role of competing RNAs in ceRNA communication. In the competition between tRNA-like-mRNA and the derivatives of tRNA for the factors that bond them, stoichiometry among competing RNAs is not the only important factor. The binding of the various factors to their target RNAs does not occur through simple Watson:Crick pairings, as in the case of microRNAs in the ceRNA-mediated communication model, but instead involves complex interactions. Thus, the differences between the affinity and kinetics constants of each of the factors for each of their target substrates in the mRNAs and that of their competitors, will play a much more important role in the case of the microRNA-mediated language, giving this type of communication a great deal of specificity. In addition, the factors would be capable of recognizing different sub-structures in the tRNA-like-mRNA [i.e., the T stem-loop and acceptor stem (Yuan

and Altman, 1995), the elbow region (Zhang and Ferré-D'Amaré, 2016), etc.] and, as a function of this, different subgroups of cellular mRNAs carrying tRNA-like signals would be able to communicate coordinately. In a similar way to ceRNA-mediated communication, infection by a virus whose RNAs contains tRNA-like structures within them can sequester factors that normally bind the tRNA-like-mRNA signal. Releasing mRNA-tRNA-likes from these factors would allow these mRNAs to participate in another way in the communication and control relationships of the infected cell (i.e., enhanced or reduced stability and translation).

Although there is no experimental evidence of this entire process, there are results from parts of it that would provisionally support this hypothesis. One of these parts, that acting in the healthy cell, is represented by the example of negative self-regulation at the translational level of the expression of Threonyl-tRNA sinthetase (ThrS) in *E. coli* by the enzyme itself. Two tRNA-like structures in the mRNA leader sequence of the ThrS enzyme compete with $tRNA^{Thr}$ for ThrS (Springer et al., 1989, 1998). An example of the other part, that of intersection between the biology of viral RNAs and tRNA is found HIV (Liu et al., 2016). The retrotranscription of the HIV-1 virus RNA needs $tRNA^{Lys}$ as a primer for the reverse transcriptase enzyme. The virion contains copies of this tRNA to initiate the reproductive cycle in the newly infected cell, but this tRNA travels as a complex with the Lysyl-tRNA synthetase. So that this $tRNA^{Lys}$ can be released from the complex, it needs the help of a tRNA-like element in the 5' UTR region of the HIV mRNA, which competes with the interaction of that complex (Liu et al., 2016).

Communication in a Social Context and Power Relations

The spread of a viral tRNA-like signal may change the context for the expression of tRNA-like-mRNAs. These changes could be related to a possible restoration of ancestral forms of cellular communication based on RNA signals that would have been silenced, repressed in the healthy cell (Gómez et al., 2015) and that the presence of the viral tRNA-like may release again. All this could extend the molecular "body" of the viral infection far beyond the mere production of more viruses. In any case, the biological interpretation of the tRNA-like-mRNA signals is an action in the conflict between the different agents in the cell, involving at least, antiviral mechanism on the one hand and cellular structures and functions that favor the infection on the other.

The concept of internal conflict and power, found in various branches of the humanities, has recently been adapted by several authors to accommodate molecular phenomena that do not fit clearly into the interpretation of the cell as a machine (Hurst et al., 1996; Gómez and Cacho, 2001; Cacho and Gómez, 2010). In some cases this takes the dialectic form, for example in the case of addiction systems (like the negation of an action that refers to the negation of another action), while in others it acquires more open forms, which reference Foucault's philosophy of power (Cacho and Gómez, 2010) and can be applied to science without metaphysical contamination, and more generally as "actions on actions" (Garnar, 2006). Andrew Garnar summarizes: "power channels the directions in which meaning goes" (Garnar, 2006).

CONCLUSION

The numerous structural elements embedded in cellular mRNA identified by RNase P could derive directly from proto-tRNA-like structures from the origin of life, or be indirectly related to these. In either case, these elements represent a record of primordial signals, a historical repository of memory brought to the present, not so much through the functions they originally represented (the "object" of the signal in the biosemiotic model (**Figure 1**), but by the potential of many archaic factors (RNase P, tRNA-aa-synthetase, ribosome subunits, etc.) to recognize and interpret these signals in the context of a present-day cell. Interaction or interpretive mediation does not have to be identical to that which occurred in a remote past, when these proto- tRNA-like signals were circulating as independent elements, but there is a possible analogy. We propose a competitive endogenous mechanism dependent on the tRNA-like structure (cetRNA) that would be compatible with primordial communication. This communicative web is the context that affects and is affected by the interpretation of the viral tRNA-like signal, but which is also modified by other signals or viral activities that repress or favor different connections in this network (Sharov et al., 2015).

Considering that the RNA virus may be originated back in the "RNA world," that the tRNA-like structures are present in the virus infecting the three branches of life, and that tRNA-likes within mRNAs may have the same ancient origin, then a cell's interpretation of both the viral tRNA-likes and its own tRNA-like-mRNAs during the infection, reveal the activity of a language at the structural RNA level that could ultimately refer to the "RNA world." These tRNA-like structural elements embedded in cellular and viral mRNAs have been adaptive and not eliminated by live evolution, otherwise random drift and "neutral" synonymous mutations (Martínez et al., 2016). Trends in Microbiology would have easily disrupted these RNA structural motifs (Martínez et al., 2016). During an infection it could be possible to recover, at least partially, interactions from the remote past that have not been completely lost, but rather controlled. In this sense, the virus is a signal for the host cell (Gómez et al., 2015).

The importance of going experimentally deeper into this hypothesis involves verifying the simultaneous presence in the cell of historical layers of signaling and recognition. Viral tRNA-like structures could cross these layers and, simply through competition, displace factors that bind tRNA-like signals embedded in the messengers that would have been under control since the distant past, putting them in contact with present-day cell factors. In this sense, viruses represent a tool that allows submersion in the past through a non-phylogenetic method. It signifies direct immersion into other temporal layers where chronology is broken down, and it simultaneously establishes continuity between the viral mimetic signal and the context of the cellular interpretation of tRNA-like-mRNAs, whose agents refer

to the origin of life. With this hypothesis in mind, it is possible that viruses can provide information on the possible molecular and consortial relationships of that primitive era.

AUTHOR CONTRIBUTIONS

The manuscript was written, discussed, edited, and reviewed by AA-M and JG.

FUNDING

The work was funded by the Spanish National Ministry of Economics and Competiveness (MINECO), grant SAF-52400-R, Instituto de Salud Carlos III, CIBERehd (Centro de Investigación en Red de Enfermedades Hepáticas y Digestivas). Consorcio Centro de Investigacion Biomedica en Red, M.P. (CIBER) Instituto de Salud Carlos III C/Monforte de Lemos, 3-5 Pabellon 11 28029-Madrid G85296226.

REFERENCES

Alcamí, A. (2003). Viral mimicry of cytokines, chemokines and their receptors. *Nat. Rev. Immunol.* 3, 36–50. doi: 10.1038/nri980

Alexander-Brett, J. M., and Fremont, D. H. (2017). Dual GPCR and GAG mimicry by the M3 chemokine decoy receptor. *J. Exp. Med.* 204, 3157–3172. doi: 10.1084/jem.20071677

Altman, S. (1975). Biosynthesis of transfer RNA in *Escherichia coli*. *Cell* 4, 21–29. doi: 10.1016/0092-8674(75)90129-4

Altman, S. (1995). RNase P in research and therapy. *Biotechnology* 13, 327–329. doi: 10.1038/nbt0495-327

Altman, S., Wesolowski, D., Guerrier-Takada, C., and Li, Y. (2005). RNase P cleaves transient structures in some riboswitches. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11284–11289. doi: 10.1073/pnas.0505271102

Andreev, D. E., Hirnet, J., Terenin, I. M., Dmitriev, S. E., Niepmann, M., and Shatsky, I. N. (2012). Glycyl-tRNA synthetase specifically binds to the poliovirus IRES to activate translation initiation. *Nucleic Acids Res.* 40, 5602–5614. doi: 10.1093/nar/gks182

Barbieri, M. (2008). Biosemiotics: a new understanding of life. *Naturwissenschaften* 95, 577–599. doi: 10.1007/s00114-008-0368-x

Barbieri, M. (2009). A short history of biosemiotics. *Biosemiotics* 2, 221–245. doi: 10.1007/s12304-009-9042-8

Barends, S., Bink, H. H., van den Worm, S. H., Pleij, C. W., and Kraal, B. (2003). Entrapping ribosomes for viral translation: tRNA mimicry as a molecular Trojan horse. *Cell* 112, 123–129. doi: 10.1016/S0092-8674(02)01256-4

Bartkiewicz, M., Gold, H., and Altman, S. (1989). Identification and characterization of an RNA molecule that copurifies with RNase P activity from HeLa cells. *Genes Dev.* 3, 488–499. doi: 10.1101/gad.3.4.488

Baumstark, T., and Ahlquist, P. (2001). The brome mosaic virus RNA3 intergenic replication enhancer folds to mimic a tRNA TpsiC-stem loop and is modified in vivo. *RNA* 7, 1652–1670.

Bernhardt, H. S., and Tate, W. P. (2010). The transition from noncoded to coded protein synthesis: did coding mRNAs arise from stability-enhancing binding partners to tRNA? *Biol. Direct* 5:16. doi: 10.1186/1745-6150-5-16

Berry, K. E., Waghray, S., Mortimer, S. A., Bai, Y., and Doudna, J. A. (2011). Crystal structure of the HCV IRES central domain reveals strategy for start-codon positioning. *Structure* 19, 1456–1466. doi: 10.1016/j.str.2011.08.002

Boehringer, D., Thermann, R., Ostareck-Lederer, A., Lewis, J. D., and Stark, H. (2005). Structure of the hepatitis C virus IRES bound to the human 80S ribosome: remodeling of the HCV IRES. *Structure* 13, 1695–1706. doi: 10.1016/j.str.2005.08.008

Branch, A. D., Benenfeld, B. J., and Robertson, H. D. (1989). RNA fingerprinting. *Methods Enzymol.* 180, 130–154. doi: 10.1016/0076-6879(89)80098-9

Brillouin, L. (1968). *Life, Thermodynamics and Cybernetics*. Chicago, IL: Aldine.

Cabot, B., Martell, M., Esteban, J. I., Sauleda, S., Otero, T., Esteban, R., et al. (2000). Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. *J. Virol.* 74, 805–811. doi: 10.1128/JVI.74.2.805-811.2000

Cacho, I., and Gómez, J. (2010). Biopuvoir et virus à ARN. De l' usage des métaphores en biologie moleculaire, à contre-courant des dogmes. *Cah. Int. Symb.* 125-127, 89–100.

Caetano-Anollés, D., and Caetano-Anollés, G. (2016). Piecemeal buildup of the genetic code, ribosomes, and genomes from primordial tRNA building blocks. *Life* 6:E43. doi: 10.3390/life6040043

Caetano-Anollés, G., and Sun, F.-J. (2014). The natural history of transfer RNA and its interactions with the ribosome. *Front. Genet.* 5:127.

Caetano-Anollés, G., Wang, M. H., and Caetano-Anollés, D. (2013). Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PLOS ONE* 8:e72225. doi: 10.1371/journal.pone.0072225

Castelruiz, Y., Larrea, E., Boya, P., Civeira, M. P., and Prieto, J. (1999). Interferon alfa subtypes and levels of type I interferons in the liver and peripheral mononuclear cells in patients with chronic hepatitis C and controls. *Hepatology* 29, 1900–1904. doi: 10.1002/hep.510290625

Chamberlain, J. R., Pagán-Ramos, E., Kindelberger, D. W., and Engelke, D. R. (1996). An RNase P RNA subunit mutation affects ribosomal RNA processing. *Nucleic Acids Res.* 24, 3158–3166. doi: 10.1093/nar/24.16.3158

Christen, U., Hintermann, E., Holdener, M., and von Herrath, M. G. (2010). Viral triggers for autoimmunity: is the 'glass of molecular mimicry' half full or half empty? *J. Autoimmun.* 34, 38–44. doi: 10.1016/j.jaut.2009.08.001

Colussi, T. M., Costantino, D. A., Hammond, J. A., Ruehle, G. M., Nix, J. C., and Kieft, J. S. (2014). The structural basis of transfer RNA mimicry and conformational plasticity by a viral RNA. *Nature* 511, 366–369. doi: 10.1038/nature13378

Colussi, T. M., Costantino, D. A., Zhu, J., Donohue, J. P., Korostelev, A. A., Jaafar, Z. A., et al. (2015). Initiation of translation in bacteria by a structured eukaryotic IRES RNA. *Nature* 519, 110–113. doi: 10.1038/nature14219

Costantino, D. A., Pfingsten, J. S., Rambo, R. P., and Kieft, J. S. (2008). tRNA-mRNA mimicry drives translation initiation from a viral IRES. *Nat. Struct. Mol. Biol.* 15, 57–64. doi: 10.1038/nsmb1351

de Farias, S., Gaudêncio-Rêgo, T., and José, M. V. (2017). Peptidyl transferase center and the emergence of the translation system. *Life* 7:E21. doi: 10.3390/life7020021

de Farias, S. T., Rêgo, T. G., and José, M. V. (2016). tRNA Core hypothesis for the transition from the RNA World to the ribonucleoprotein World. *Life* 6:E15. doi: 10.3390/life6020015

Di Giulio, M. (2012). The origin of the tRNA molecule: Independent data favor a specific model of its evolution. *Biochimie* 94, 1464–1466. doi: 10.1016/j.biochi.2012.01.014

Di Giulio, M. (2015). A model for the origin of the first mRNAs. *J. Mol. Evol.* 81, 10–17. doi: 10.1007/s00239-015-9691-y

Díaz-Toledano, R., and Gómez, J. (2015). Messenger RNAs bearing tRNA-like features exemplified by interferon alfa 5 mRNA. *Cell Mol. Life Sci.* 37, 3747–3768. doi: 10.1007/s00018-015-1908-0

Domingo, E. (2007). "Virus evolution," in *Fields Virology*, 5th Edn, eds D. M. Knipe and P. M. Howley (Philadelphia, PA: Lippincott Williams & Wilkins), 389–421.

Drayman, N., Glick, Y., Ben-nun-shaul, O., Zer, H., Zlotnick, A., Gerber, D., et al. (2013). Pathogens use structural mimicry of native host ligands as a mechanism for host receptor engagement. *Cell Host Microbe* 14, 63–73. doi: 10.1016/j.chom.2013.05.005

Dreher, T. W. (2009). Role of tRNA-like structures in controlling plant virus replication. *Virus Res.* 139, 217–229. doi: 10.1016/j.virusres.2008.06.010

Dreher, T. W. (2010). Viral tRNAs and tRNA-like structures. *Wiley Interdiscip. Rev. RNA* 1, 402–414. doi: 10.1002/wrna.42

Dumas, P., Moras, D., Florentz, C., Giegé, R., Verlaan, P., Van Belkum, A., et al. (1987). 3-D graphics modelling of the tRNA-like 3′-end of turnip yellow mosaic virus RNA: structural and functional implications. *J. Biomol. Struct. Dyn.* 4, 707–728. doi: 10.1080/07391102.1987.10507674

Eigen, M., and Biebricher, C. K. (1988). "Sequence space and quasispecies distribution," in *RNA Genetics*, eds E. Domingo, J. J. Holland, and P. Ahlquist (Boca Raton FL: CRC press).

Eigen, M., and Winkler-Oswatitsch, R. (1981). Transfer-RNA, an early gene? *Naturwissenschaften* 68, 282–292.

Fechter, P., Rudinger-Thirion, J., Florentz, C., and Giegé, R. (2001). Novel features in the tRNA-like world of plant viral RNAs. *Cell Mol. Life. Sci.* 58, 1547–1561. doi: 10.1007/PL00000795

Felden, B., Florentz, C., Giegé, R., and Westhof, E. (1996). A central pseudoknotted three-way junction imposes tRNA-like mimicry and the orientation of three 5' upstream pseudoknots in the 3' terminus of tobacco mosaic virus RNA. *RNA* 2, 201–212.

Felden, B., Florentz, C., McPherson, A., and Giegé, R. (1994). A histidine accepting tRNA-like fold at the 3'-end of satellite tobacco mosaic virus RNA. *Nucleic Acid Res.* 22, 2882–2886. doi: 10.1093/nar/22.15.2882

Fernandez, I. S., Bai, X.-C., Murshudov, G., Scheres, S., and Ramakrishnan, V. (2014). Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* 157, 823–831. doi: 10.1016/j.cell.2014.04.015

Ferrari, S., Yehle, C. O., Robertson, H. D., and Dickson, E. (1980). Specific RNA-cleaving activities from HeLa cells. *Proc. Natl. Acad. Sci. U.S.A.* 77, 2395–2399. doi: 10.1073/pnas.77.5.2395

Forster, A. C., and Altman, S. (1990). External guide sequences for an RNA enzyme. *Science* 249, 783–786. doi: 10.1126/science.1697102

Foucault, M. (2001). "El sujeto y el poder ¿en que consiste la naturaleza específica del poder?," in *Arte Después de la Modernidad*, ed. B. Wallis (Madrid: Ediciones Akal), 431.

Garnar, A. (2006). Power, action, signs: between peirce and foucault. *Trans. Charles S Pierce Soc.* 42, 347–366. doi: 10.2979/TRA.2006.42.3.347

Gebetsberger, J., Wyss, L., Mleczko, A. M., Reuther, J., and Polacek, N. (2016). A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. *RNA Biol.* doi: 10.1080/15476286.2016.1257470 [Epub ahead of print].

Giegé, R. (2008). Toward a more complete view of tRNA biology. *Nat. Struct. Mol. Biol.* 15, 1007–1014. doi: 10.1038/nsmb.1498

Giegé, R., Florentz, C., and Dreher, T. W. (1993). The TYMV tRNA-like structure. *Biochimie* 75, 569–582. doi: 10.1016/0300-9084(93)90063-X

Giegé, R., Frugier, M., and Rudinger, J. (1998). tRNA mimics. *Curr. Opin. Struct. Biol.* 8, 286–293. doi: 10.1016/S0959-440X(98)80060-2

Gilbert, N., and Labuda, D. (1999). CORE-SINEs: Eukaryotic short interspersed retroposing elements with common sequence motifs. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2869–2874. doi: 10.1073/pnas.96.6.2869

Gómez, J., Ariza-Mateos, A., and Cacho, I. (2015). Virus is a signal for the host cell. *Biosemiotics* 8, 483–491. doi: 10.1007/s12304-015-9245-0

Gómez, J. C., and Cacho, I. (2001). Can Nietzsche power relationships be experimentally approached with theoretical and viral quasispecies? *Contrib. Sci.* 2, 103–108.

Guerrier-Takada, C., van Belkum, A., Pleij, C. W., and Altman, S. (1988). Novel reactions of RNAase P with a tRNA-like structure in turnip yellow mosaic virus RNA. *Cell* 53, 267–272. doi: 10.1016/0092-8674(88)90388-1

Guo, Q., and Lambowitz, A. (1992). A tyrosyl-tRNA synthetase binds specifically to the group I intron catalytic core. *Genes Dev.* 6, 1357–1372. doi: 10.1101/gad.6.8.1357

Haenni, A.-L., Prochiantz, A., Bernard, O., and Chapeville, F. (1973). TYMV valyl-RNA as an amino-acid donor in protein biosynthesis. *Nature New Biol.* 241, 166–168. doi: 10.1038/newbio241166a0

Hall, T. C., Shih, D. S., and Kaesberg, P. (1972). Enzyme-mediated binding of tyrosine to brome-mosaic-virus ribonucleic acid. *Biochem. J.* 129, 969–976. doi: 10.1042/bj1290969

Hansen, A., Pfeiffer, T., Zuleeg, T., Limmer, S., Ciesiolka, J., Feltens, R., et al. (2001). Exploring the minimal substrate requirements for trans-cleavage by RNase P holoenzymes from *Escherichia coli* and *Bacillus subtilis*. *Mol. Microbiol.* 41, 131–143. doi: 10.1046/j.1365-2958.2001.02467.x

Hoffmeyer, J., and Emmeche, C. (1991). "Code-duality and the semiotics of nature," in *On Semiotic Modeling*, eds M. Anderson and F. Merrell (Berlin: Mouton de Gruyter), 117–166.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisse, M., Merrill, S. H., et al. (1965). Structure of a ribonucleic acid. *Science* 147, 462–1465. doi: 10.1126/science.147.3664.1462

Hurst, L. D., Atlan, A., and Bengtsson, B. O. (1996). Genetic conflicts. *Q. Rev. Biol.* 71, 317–364. doi: 10.1086/419442

Jackson, R. J., Hellen, C. U., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* 11, 113–127. doi: 10.1038/nrm2838

Jan, E., and Sarnow, P. (2002). Factorless ribosome assembly on the internal ribosome entry site of cricket paralysis virus. *J. Mol. Biol.* 324, 889–902. doi: 10.1016/S0022-2836(02)01099-9

Jarrous, N., and Reiner, R. (2007). Human RNase P: a tRNA-processing enzyme and transcription factor. *Nucleic Acids Res.* 35, 3519–3524. doi: 10.1093/nar/gkm071

Jha, S., Rollins, M. G., Fuchs, G., Procter, D. J., Hall, E. A., Cozzolino, K., et al. (2017). Trans-kingdom mimicry underlies ribosome customization by a poxvirus kinase. *Nature* 546, 651–655. doi: 10.1038/nature22814

Jiang, X., Chen, Y. C., Gong, H., Trang, P., Lu, S., and Liu, F. (2012). Ribonuclease P-mediated inhibition of human cytomegalovirus gene expression and replication induced by engineered external guide sequences. *RNA Biol.* 9, 1186–1195. doi: 10.4161/rna.21724

Kamoshita, N., Nomoto, A., and RajBhandary, U. L. (2009). Translation initiation from the ribosomal A site or the P site, dependent on the conformation of RNA pseudoknot I in dicistrovirus RNAs. *Mol. Cell* 35, 181–190. doi: 10.1016/j.molcel.2009.05.024

Katz, A., Elgamal, S., Rajkovic, A., and Ibba, M. (2016). Non-canonical roles of tRNAs and tRNA mimics in bacterial cell biology. *Mol. Microbiol.* 101, 545–558. doi: 10.1111/mmi.13419

Kilstrup, M. (2015). Naturalizing semiotics: the triadic sign of Charles Sanders Peirce as a systems property. *Prog. Biophys. Mol. Biol.* 119, 563–575. doi: 10.1016/j.pbiomolbio.2015.08.013

Kim, S. H., Quigley, G. J., Suddath, F. L., McPherson, A., Sneden, D., Kim, J. J., et al. (1973). The three-dimensional structure of yeast phenylalanine tRNA: folding of the polynucleotide chain. *Science* 179, 285–288. doi: 10.1126/science.179.4070.285

Kole, R., and Altman, S. (1981). Properties of purified ribonuclease P from *Escherichia coli*. *Biochemistry* 20, 1902–1906. doi: 10.1021/bi00510a028

Komine, Y., Kitabatake, M., Yokogawa, T., Nishikawa, K., and Inokuchi, H. (1994). A tRNA-like structure is present in 10Sa RNA, a small stable RNA from *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 91, 9223–9227. doi: 10.1073/pnas.91.20.9223

Konarska, M. M., and Sharp, P. A. (1990). Structure of RNAs replicated by the DNA-dependent T7 RNA polymerase. *Cell* 63, 609–618. doi: 10.1016/0092-8674(90)90456-O

Kropp, K. A., Angulo, A., and Ghazal, P. (2014). Viral enhancer mimicry of host innate-immune promoters. *PLOS Pathog.* 10:e1003804. doi: 10.1371/journal.ppat.1003804

Legon, S., Model, P., and Robertson, H. D. (1977). Interaction of rabbit reticulocyte ribosomes with bacteriophage f1 mRNA and of *Escherichia coli* ribosomes with rabbit globin mRNA. *Proc. Natl. Acad. Sci. U.S.A.* 74, 2692–2696. doi: 10.1073/pnas.74.7.2692

Lei, H., Dias, A. P., and Reed, R. (2011). Export and stability of naturally intronless mRNAs require specific coding region sequences and the TREX mRNA export complex. *Proc. Natl. Acad. Sci. U.S.A.* 108, 17985–17990. doi: 10.1073/pnas.1113076108

Lei, H., Zhai, B., Yin, S., Gygi, S., and Reed, R. (2013). Evidence that a consensus element found in naturally intronless mRNAs promotes mRNA export. *Nucleic Acids Res.* 41, 2517–2525. doi: 10.1093/nar/gks1314

Litvak, S., Carr, D. S., and Chapeville, F. (1970). TYMV RNA As a substrate of the tRNA nucleotidyltransferase. *FEBS Lett.* 11, 316–319. doi: 10.1016/0014-5793(70)80557-9

Liu, F., and Altman, S. (1995). Inhibition of viral gene expression by the catalytic RNA subunit of RNase P from *Escherichia coli*. *Genes Dev.* 9, 471–480. doi: 10.1101/gad.9.4.471

Liu, F., and Altman, S. (1996). Requirements for cleavage by a modified RNase P of a small model substrate. *Nucleic Acids Res.* 24, 2690–2696. doi: 10.1093/nar/24.14.2690

Liu, S., Comandur, R., Jones, C. P., Tsang, P., and Musier-Forsyth, K. (2016). Anticodon-like binding of the HIV-1 tRNA-like element to human lysyl-tRNA synthetase. *RNA* 22, 1828–1835. doi: 10.1261/rna.058081.116

Lozano, G., and Martínez-Salas, E. (2015). Structural insights into viral IRES-dependent translation mechanisms. *Curr. Opin. Virol.* 12, 113–120. doi: 10.1016/j.coviro.2015.04.008

Lukavsky, P. J., Kim, I., Otto, G. A., and Puglisi, J. D. (2003). Structure of HCV IRES domain II determined by NMR. *Nat. Struct. Biol.* 10, 1033–1038. doi: 10.1038/nsb1004

Luna, J. M., Scheel, T. K., Danino, T., Shaw, K. S., Mele, A., Fak, J. J., et al. (2015). Hepatitis C virus RNA functionally sequesters miR-122. *Cell* 160, 1099–1110. doi: 10.1016/j.cell.2015.02.025

Lyons, A. J., and Robertson, H. D. (2003). Detection of tRNA-like structure through RNase P cleavage of viral internal ribosome entry site RNAs near the AUG start triplet. *J. Biol. Chem.* 278, 26844–26850. doi: 10.1074/jbc.M304052200

Lytle, J. R., Yario, T. A., and Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proc. Natl. Acad. Sci. U.S.A.* 104, 9667–9672. doi: 10.1073/pnas.0703820104

Macé, K., and Gillet, R. (2016). Origins of tmRNA: the missing link in the birth of protein synthesis? *Nucleic Acids Res.* 44, 8041–8051. doi: 10.1093/nar/gkw693

Maran, T. (2017). *Mimicry and Meaning: Structure and Semiosis of Biological Mimicry*. Cham: Springer International Publishing AG. doi: 10.1007/978-3-319-50317-2

Marazzi, I., Ho, J. S., Kim, J., Manicassamy, B., Dewell, S., Albrecht, R. A., et al. (2012). Suppression of the antiviral response by an influenza histone mimic. *Nature* 483, 428–433. doi: 10.1038/nature10892

Martínez, M. A., Jordan-Paiz, A., Franco, S., and Nevot, M. (2016). Synonymous virus genome recoding as a tool to impact viral fitness. *Trends Microbiol.* 24, 134–147. doi: 10.1016/j.tim.2015.11.002

Marvin, M. C., Walker, S. C., Fierke, C. A., and Engelke, D. R. (2011). Binding and cleavage of unstructured RNA by nuclear RNase P. *RNA* 17, 1429–1440. doi: 10.1261/rna.2633611

Mas, A., López-Galíndez, C., Cacho, I., Gómez, J., and Martínez, M. A. (2010). Unfinished stories on viral quasispecies and Darwinian views of evolution. *J. Mol. Biol.* 397, 865–877. doi: 10.1016/j.jmb.2010.02.005

Matsuda, D., and Dreher, T. W. (2006). Close spacing of AUG initiation codons confers dicistronic character on a eukaryotic mRNA. *RNA* 12, 1338–1349. doi: 10.1261/rna.67906

Matsuda, D., and Dreher, T. W. (2007). Cap- and initiator tRNA-dependent initiation of TYMV polyprotein synthesis by ribosomes: evaluation of the Trojan horse model for TYMV RNA translation. *RNA* 13, 129–137. doi: 10.1261/rna.244407

Mattijssen, S., Hinson, E. R., Onnekink, C., Hermanns, P., Zabel, B., Cresswell, P., et al. (2011). Viperin mRNA is a novel target for the human RNase MRP/RNase P endoribonuclease. *Cell Mol. Life Sci.* 68, 2469–2480. doi: 10.1007/s00018-010-0568-3

Maute, R. L., Dalla-Favera, R., and Basso, K. (2014). RNAs with multiple personalities. *Wiley Interdiscip. Rev. RNA* 5, 1–13. doi: 10.1002/wrna.1193

Mayr, E. (1982). *The Position of Biology within the Sciences. The Growth of Biological Thought*. Cambridge, MA: Harvard University Press, 32–36.

Michel, F., and Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216, 585–610. doi: 10.1016/0022-2836(90)90386-Z

Murphy, P. M. (2001). Viral exploitation and subversion of the immune system through chemokine mimicry. *Nat. Immunol.* 2, 116–122. doi: 10.1038/84214

Nadal, A., Robertson, H. D., Guardia, J., and Gomez, J. (2003). Characterization of the structure and variability of an internal region of hepatitis C virus RNA for M1 RNA guide sequence ribozyme targeting. *J. Gen. Virol.* 84, 1545–1548. doi: 10.1099/vir.0.18898-0

Nelson, J. W., and Breaker, R. R. (2017). The lost language of the RNA World. *Sci. Signal.* 10, 1–10. doi: 10.1126/scisignal.aam8812

Ohnishi, K., Hokari, S., Shutou, H., Ohshima, M., Furuichi, N., and Goda, M. (2002). Origin of most primitive mRNAs and genetic codes via interactions between primitive tRNA ribozymes. *Genome Inform.* 13, 71–81.

Ohnishi, K., Ohshima, M., and Furuichi, N. (2005). Evolution from possible primitive tRNA-viroids to early poly-tRNA-derived mRNAs: a new approach from the poly-tRNA theory. *Genome Inform.* 16, 94–103.

Oldstone, M. B. (2014). Molecular mimicry: its evolution from concept to mechanism as a cause of autoimmune diseases. *Monoclon. Antib. Immunodiagn. Immunother.* 33, 158–165. doi: 10.1089/mab.2013.0090

Pan, K., and Deem, M. W. (2011). Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. *J. R. Soc. Interface* 8, 1644–1653. doi: 10.1098/rsif.2011.0105

Pascual, A., and Vioque, A. (1999). Substrate binding and catalysis by ribonuclease P from cyanobacteria and *Escherichia coli* are affected differently by the 3′ terminal CCA in tRNA precursors. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6672–6677. doi: 10.1073/pnas.96.12.6672

Pasteur, G. (1995). "Camouflages et homotypies," in *Bilogie et Mimétismes*, ed. Nathan (Paris: Nathan).

Peck-Miller, K. A., and Altman, S. (1991). Kinetics of the processing of the precursor to 4.5 S RNA, a naturally occurring substrate for RNase P from *Escherichia coli*. *J. Mol. Biol.* 221, 1–5. doi: 10.1016/0022-2836(91)80194-Y

Pilipenko, E. V., Maslova, S. V., Sinyakov, A. N., and Agol, V. I. (1992). Towards identification of cis-acting elements involved in the replication of enterovirus and rhinovirus RNAs: a proposal for the existence of tRNA-like terminal structures. *Nucleic Acids Res.* 20, 1739–1745. doi: 10.1093/nar/20.7.1739

Pinck, M., Yot, P., Chapeville, F., and Duranton, H. M. (1970). Enzymatic binding of valine to the 3' end of TYMV-RNA. *Nature* 226, 954–956. doi: 10.1038/226954a0

Piron, M., Beguiristain, N., Nadal, A., Martínez-Salas, E., and Gomez, J. (2005). Characterizing the function and structural organization of the 5' tRNA-like motif within the hepatitis C virus quasispecies. *Nucleic Acids Res.* 33, 1487–1502. doi: 10.1093/nar/gki290

Prieto-Vega, S. (2016). *Efecto del Aumento Mutacional Sobre el Reconocimineto de los Motivos Estructurales del ARN de la Región Genómica 5' del Virus de la Hepatitis C por Factores Bioquímicos*. Granada: Universidd de Granada.

Prochiantz, A., and Haenni, A. L. (1973). TYMV RNA as a substrate of tRNA maturation endonuclease. *Nat. New Biol.* 241, 168–170. doi: 10.1038/newbio241168a0

Rekosh, D. M., Lodish, H., and Baltimore, D. (1970). Protein synthesis in *Escherichia coli* extracts programmed by poliovirus RNA. *J. Mol. Biol.* 54, 327–340. doi: 10.1016/0022-2836(70)90433-X

Ribas de Pouplana, L., and Schimmel, P. (2001). Operational RNA code for amino acids in relation to genetic code in evolution. *J. Biol. Chem.* 276, 6881–6884. doi: 10.1074/jbc.R000032200

Robertson, H. D., Altman, S., and Smith, J. D. (1972). Purification and properties of a specific *Escherichia coli* ribonuclease which cleaves a tyrosine transfer ribonucleic acid presursor. *J. Biol. Chem.* 247, 5243–5251.

Robertson, H. D., and Neel, O. D. (1999). "Virus origins: conjoined RNAs genomes as precursor to DNA genomes," in *Origin and Evolution of Virus*, eds E. W. Domingo, R. Webster, and J. Holland (London: Academic Press), 25–37. doi: 10.1016/B978-012220360-2/50003-9

Rodin, A. S., Szathmáry, E., and Rodin, S. N. (2011). On origin of genetic code and tRNA before translation. *Biol. Direct* 6:14. doi: 10.1186/1745-6150-6-14

Sabariegos, R., Nadal, A., Beguiristain, N., Piron, M., and Gómez, J. (2004). Catalytic RNase P RNA from *Synechocystis* sp. cleaves the hepatitis C virus RNA near the AUG start codon. *FEBS Lett.* 577, 517–522. doi: 10.1016/j.febslet.2004.10.059

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

Scheel, T. K., Luna, J. M., Liniger, M., Nishiuchi, E., Rozen-Gagnon, K., Shlomai, A., et al. (2016). A broad RNA virus survey reveals both miRNA dependence and functional sequestration. *Cell Host Microbe* 19, 409–423. doi: 10.1016/j.chom.2016.02.007

Schimmel, P., Giege, R., Moras, D., and Yokoyama, S. (1993). An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl. Acad. Sci. U.S.A.* 90, 8763–8768. doi: 10.1073/pnas.90.19.8763

Schuster, P. (1997). Genotypes with phenotypes: adventures in an RNA toy world. *Biophys. Chem.* 66, 75–110. doi: 10.1016/S0301-4622(97)00058-6

Searls, D. B. (2002). The language of genes. *Nature* 420, 211–217. doi: 10.1038/nature01255

Seif, E., and Altman, S. (2008). RNase P cleaves the adenine riboswitch and stabilizes pbuE mRNA in *Bacillus subtilis*. *RNA* 14, 1237–1243. doi: 10.1261/rna.833408

Sela, I. (1972). Tobacco enzyme-cleaved fragments of TMV-RNA specifically accepting serine and methionine. *Virology* 49, 90–94. doi: 10.1016/S0042-6822(72)80009-6

Serrano, P., Gómez, J., and Martínez-Salas, E. (2007). Characterization of a cyanobacterial RNase P ribozyme recognition motif in the IRES of foot-and-mouth disease virus reveals a unique structural element. *RNA* 13, 849–859. doi: 10.1261/rna.506607

Shannon, C. E., and Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Sharov, A., Maran, T., and Tønnessen, M. (2015). Organisms reshape sign relations. *Biosemiotics* 8, 361–365. doi: 10.1007/s12304-016-9269-0

Spahn, C. M., Jan, E., Mulder, A., Grassucci, R. A., Sarnow, P., and Frank, J. (2004). Cryo-EM visualization of a viral internal ribosome entry site bound to human ribosomes: the IRES functions as an RNA-based translation factor. *Cell* 118, 465–475. doi: 10.1016/j.cell.2004.08.001

Springer, M., Graffe, M., Dondon, J., and Grunberg-Manago, M. (1989). tRNA-like structures and gene regulation at the translational level: a case of molecular mimicry in *Escherichia coli*. *EMBO J.* 8, 2417–2424.

Springer, M., Portier, C., and Grunberg-Manago, M. (1998). "RNA mimicry in the translational apparatus," in *RNA Structure and Function,* eds R. W Simons, and M. Grunberg-Manago (New York, NY: Cold Spring Laboratory Press), 377–413.

Tamura, K. (2011). Ribosome evolution: emergence of peptide synthesis machinery. *J. Biosci.* 36, 921–928. doi: 10.1007/s12038-011-9158-2

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352. doi: 10.1038/nature12986

Villarreal, L. P. (2009). *Origin of Group Identity.Viruses, Addiction and Cooperation*. Irvine, CA: Springer.

Villarreal, L. P. (2015). Force for ancient and recent life: viral and stem-loop RNA consortia promote life. *Ann. N. Y. Acad. Sci.* 1341, 25–34. doi: 10.1111/nyas.12565

Villarreal, P., and Witzany, G. (2013). The DNA habitat and its RNA inhabitants: at the dawn of RNA sociology. *Genomics Insights* 6, 1–12. doi: 10.4137/GEI.S11490

Vioque, A. (1992). Analysis of the gene encoding the RNA subunit of ribonuclease P from cyanobacteria. *Nucleic Acids Res.* 20, 6331–6337. doi: 10.1093/nar/20.23.6331

Werner, M., Rosa, E., and George, S. T. (1997). Design of short external guide sequences (EGSs) for cleavage of target molecules with RNase P. *Nucleic Acids Symp. Ser.* 36, 19–21.

Werner, M., Rosa, E., Nordstrom, J. L., Goldberg, A. R., and George, S. T. (1998). Short oligonucleotides as external guide sequences for site-specific cleavage of RNA molecules with human RNase P. *RNA* 4, 847–855. doi: 10.1017/S1355838298980323

Wickler, W. (1998). Mimicry the new Enciclopedia Britannica. *Macropaedia* 24, 144–151.

Wilusz, J. E., Freier, S. M., and Spector, D. L. (2008). 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 135, 919–932. doi: 10.1016/j.cell.2008.10.012

Witz, J. (2003). 1964: The first model for the shape of a transfer RNA molecule. An account of an unpublished small-angle X-ray scattering study. *Biochimie* 85, 1265–1268. doi: 10.1016/j.biochi.2003.09.018

Witzany, G. (1995). From the "logic of the molecular syntax" to molecular pragmatism. Explanatory deficits in Manfred Eigen's concept of language and communication. *Evolut. Cogn.* 1, 148–168.

Witzany, G. (2010). "Introduction: metaphysical and postmetaphysical relationships of humans with nature and life," in *Biocommunication and Natural Genome Editing*, ed. G. Witzany (Dordrecht: Springer), 1–26.

Witzany, G. (2014). RNA sociology: group behavioral motifs of RNA consortia. *Life* 4, 800–818. doi: 10.3390/life4040800

Witzany, G. (2017a). "Key levels of biocommunication," in *Biocommunication: Sign-Mediated Interactions between Cells and Organisms*, eds R. Gordon and J. Seckbach (London: World Scientific), 37–61.

Witzany, G. (2017b). Two genetic codes: repetitive syntax for active non-coding RNAs; non-repetitive syntax for the DNA archives. *Commun. Integr. Biol.* 10, 1–12. doi: 10.1080/19420889.2017.1297352

Witzany, G., and Baluska, F. (2012). Life's code script does not code itself. The machine metaphor for living organisms is outdated. *EMBO Rep.* 13, 1054–1056. doi: 10.1038/embor.2012.166

Wolinsky, S. M., Korber, B. T., Neumann, A. U., Daniels, M., Kunstman, K. J., Whetsell, A. J., et al. (1996). Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science* 272, 537–542. doi: 10.1126/science.272.5261.537

Yuan, Y., and Altman, S. (1995). Substrate recognition by human RNase P: identification of small, model substrates for the enzyme. *EMBO J.* 14, 159–168.

Zhang, J., and Ferré-D'Amaré, A. R. (2016). The tRNA Elbow in Structure, Recognition and Evolution. *Life* 6:E3. doi: 10.3390/life6010003

Zhang, W., Thieme, C., Kollwig, G., Apelt, F., Yang, L., Winter, N., et al. (2016). tRNA-related sequences trigger systemic mRNA transport in plants. *Plant Cell* 28, 1237–1249. doi: 10.1105/tpc.15.01056

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

MINI REVIEW
published: 05 October 2017
doi: 10.3389/fmicb.2017.01869

Retrotransposon Domestication and Control in *Dictyostelium discoideum*

Marek Malicki[‡], *Maro Iliopoulou*[†‡] *and Christian Hammann*[*]

Ribogenetics Biochemistry Lab, Department of Life Sciences and Chemistry, Jacobs University Bremen, Bremen, Germany

Transposable elements, identified in all eukaryotes, are mobile genetic units that can change their genomic position. Transposons usually employ an excision and reintegration mechanism, by which they change position, but not copy number. In contrast, retrotransposons amplify via RNA intermediates, increasing their genomic copy number. Hence, they represent a particular threat to the structural and informational integrity of the invaded genome. The social amoeba *Dictyostelium discoideum*, model organism of the evolutionary Amoebozoa supergroup, features a haploid, gene-dense genome that offers limited space for damage-free transposition. Several of its contemporary retrotransposons display intrinsic integration preferences, for example by inserting next to transfer RNA genes or other retroelements. Likely, any retrotransposons that invaded the genome of the amoeba in a non-directed manner were lost during evolution, as this would result in decreased fitness of the organism. Thus, the positional preference of the *Dictyostelium* retroelements might represent a domestication of the selfish elements. Likewise, the reduced danger of such domesticated transposable elements led to their accumulation, and they represent about 10% of the current genome of *D. discoideum*. To prevent the uncontrolled spreading of retrotransposons, the amoeba employs control mechanisms including RNA interference and heterochromatization. Here, we review TRE5-A, DIRS-1 and Skipper-1, as representatives of the three retrotransposon classes in *D. discoideum,* which make up 5.7% of the *Dictyostelium* genome. We compile open questions with respect to their mobility and cellular regulation, and suggest strategies, how these questions might be addressed experimentally.

Keywords: Skipper-1, DIRS-1, TRE5-A, RNA interference, retrovirus

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria

Reviewed by:
Yukihito Ishizaka,
National Center for Global Health and Medicine, Japan
Akio Kanai,
Keio University, Japan

***Correspondence:**
Christian Hammann
c.hammann@jacobs-university.de

[†]***Present Address:***
Maro Iliopoulou,
Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

[‡]*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *01 August 2017*
Accepted: *13 September 2017*
Published: *05 October 2017*

Citation:
Malicki M, Iliopoulou M and Hammann C (2017) Retrotransposon Domestication and Control in Dictyostelium discoideum. Front. Microbiol. 8:1869. doi: 10.3389/fmicb.2017.01869

TRANSPOSABLE ELEMENTS AND THEIR HABITAT

A significant fraction of all eukaryotic genomes is scattered with repetitive sequences that are predominantly related to different types of transposable elements (TEs) (Biscotti et al., 2015). At first glance, TEs are selfish or parasitic DNA, imposing the burden of their propagation on the invaded host. However, their ability to move and multiply within host genomes can not only have a negative impact, but can also improve the fitness of their host (McClintock, 1950). It is well established that transposition can alter gene expression, as TE insertion upstream of a gene can lead to its up-regulation, and downstream insertion to its down-regulation (Slotkin and Martienssen, 2007). Furthermore, transposition can promote inversions and deletions of large chromosomal DNA fragments, gene mutation, gene shuffling, transcriptional regulation, dispersion of regulatory sequences, genomic recombination and chromosomal rearrangements (Huang et al., 2012). Thus, TE mobility is a crucial factor driving genome evolution (Gbadegesin, 2012).

The natural urge of TEs to move is confronted with the pressure of their hosts to preserve the genomic integrity, with respect to both, structural and informational stability, a prerequisite to eventually replicate successfully (**Figure 1**). Therefore, as computational modeling suggests, the inhabitation of a genome by TEs happens in two stages: After invasion, for example by horizontal transfer of the TE, an initial transposition burst occurs. To reduce the burden of additional TE copies, the invaded host organism subsequently develops strategies to tightly control TE movement within the genome (Le Rouzic and Capy, 2005). Such regulatory mechanisms employ frequently parts of the RNA interference (RNAi) and epigenetic machineries (Castel and Martienssen, 2013). TE control is often incomplete such that TEs that integrate at random positions might cause deleterious mutations. This results in reduced fitness and might lead to the death of the organisms and thus disappearance of the TE (**Figure 1**). To counteract this threat, several selfish TEs have developed strategies to limit the potential deleterious consequences of their activity. One of them is to target gene-poor or transcriptionally inactive regions of the genome, like telomeres and centromeres (Levis et al., 1993; Gao et al., 2015), which essentially results in their domestication (**Figure 1**). Hence, similar to intracellular pathogens, TEs propagate within their host's genome while exploiting cellular mechanisms to maintain integrity and functionality of their niche.

TWO CLASSES OF TRANSPOSABLE ELEMENTS: RETROTRANSPOSONS AND DNA TRANSPOSONS

In principle, mobile genetic elements are categorized on mechanistic grounds, distinguishing retrotransposons from DNA transposons (Kazazian, 2004; Goodier and Kazazian, 2008). The latter, called class II elements (Wicker et al., 2007), move usually by a cut-and-paste process mediated by a TE-encoded recombinase or transposase. They consist of inverted terminal repeats and at least one open reading frame (ORF) coding for a transposase, which excises the entire transposon allowing for integration into a new locus (Vos et al., 1996). Beyond the transposase gene, class II TEs are usually not transcribed, thus do not move through a full RNA intermediate (Wicker et al., 2007), and therefore are here not further considered.

In contrast to class II TEs, retrotransposons amplify through an RNA intermediate (Wicker et al., 2007). This RNA intermediate is converted to a complementary DNA (cDNA) molecule by a reverse transcriptase (RT) activity, followed by integration of resulting cDNA copies. Thus, these class I elements use a copy-and-paste mechanism that, in the absence of suitable control mechanisms, progressively increases their copy number in the genome (Castro-Diaz et al., 2015). Retroelements can be further divided into long terminal repeat (LTR) and non-LTR retrotransposons. The former have arisen from retroviruses that once integrated into the genome of the host or its ancestor. This can be inferred from their similar structural organization, with all or parts of the prototypic retrovirus coding sequences flanked by two LTRs (Friedli and Trono, 2015). Non-LTR elements are subdivided into autonomous and non-autonomous TEs. The autonomous elements have usually two ORFs that encode the proteins required for transposition. These can also act *in trans* on non-autonomous TEs, if these still have the sequences necessary for transposition (Wicker et al., 2007; Kapitonov and Jurka, 2008). In addition, autonomous retroelements are also categorized based on the enzyme that performs the integration: YR elements encode a tyrosine recombinase, while other retroelements employ integrases or endonucleases (Goodwin and Poulter, 2001; Duncan et al., 2002; Wicker et al., 2007).

DICTYOSTELIUM DISCOIDEUM AND ITS MOBILE GENETIC ELEMENTS

Dictyostelium discoideum serves as a model organism helping to understand biological phenomena like motility, chemotaxis, phagocytosis and cytokinesis (Unal and Steinert, 2006; Calvo-Garrido et al., 2010; Surcel et al., 2010; Cai and Devreotes, 2011). The amoeba belongs to the evolutionary supergroup of Amoebozoa (Adl et al., 2012), and it likely is its best-studied representative. In response to starvation, *Dictyostelium* forms differentiated multicellular structures upon aggregation of thousands of solitary amoebae (Fets et al., 2010). At the genomic level, *D. discoideum* amazed the research community with an unexpected diversity of mobile genetic elements (Glöckner et al., 2001) in a gene-dense arrangement (Eichinger et al., 2005). Roughly two thirds of its genome (34 Mb) are protein-coding genes, and 10% are TEs (Glöckner et al., 2001; Eichinger et al., 2005). In such a gene-dense genome, a high frequency of TEs appears unlikely to persist during evolution, unless (a) the TEs have developed strategies for damage-free transposition, (b) the invaded host has developed measures to control TE mobility, or (c) the host benefits from the presence of the TEs. In view of the genome composition of *D. discoideum,* it appears likely, that both the host and the invading TEs have developed strategies to allow for co-existence, possibly even a mutually beneficial co-evolution.

The TEs in the *D. discoideum* genome have been charted in seminal work by Glöckner et al. (2001). Its DNA transposons represent 1.5% of the genome content and interestingly, none of their transposases share significant similarity with known transposases (Winckler et al., 2011). They fall into three main families, the Tdd elements, the DDT elements and the Thug elements with genomic frequencies of 0.5, 0.9, and 0.1%, respectively. To our knowledge, for none of these DNA transposons, expression or cellular control have been studied in detail. It would be of particular interest to analyze whether any of the control mechanisms that begin to emerge for the retrotransposons, discussed below, might also act on DNA transposons. The genome of *D. discoideum* contains retrotransposons that fall into three major classes: non-LTR, LTR, and YR retrotransposons (Glöckner et al., 2001; Winckler et al., 2011). Although the YR retrotransposons feature LTRs, they are considered their own class due to unique characteristics, like the presence of a tyrosine recombinase (Poulter and Goodwin, 2005). In *D. discoideum*, retrotransposons make up about 8% of the genome, which is a significant expansion compared to

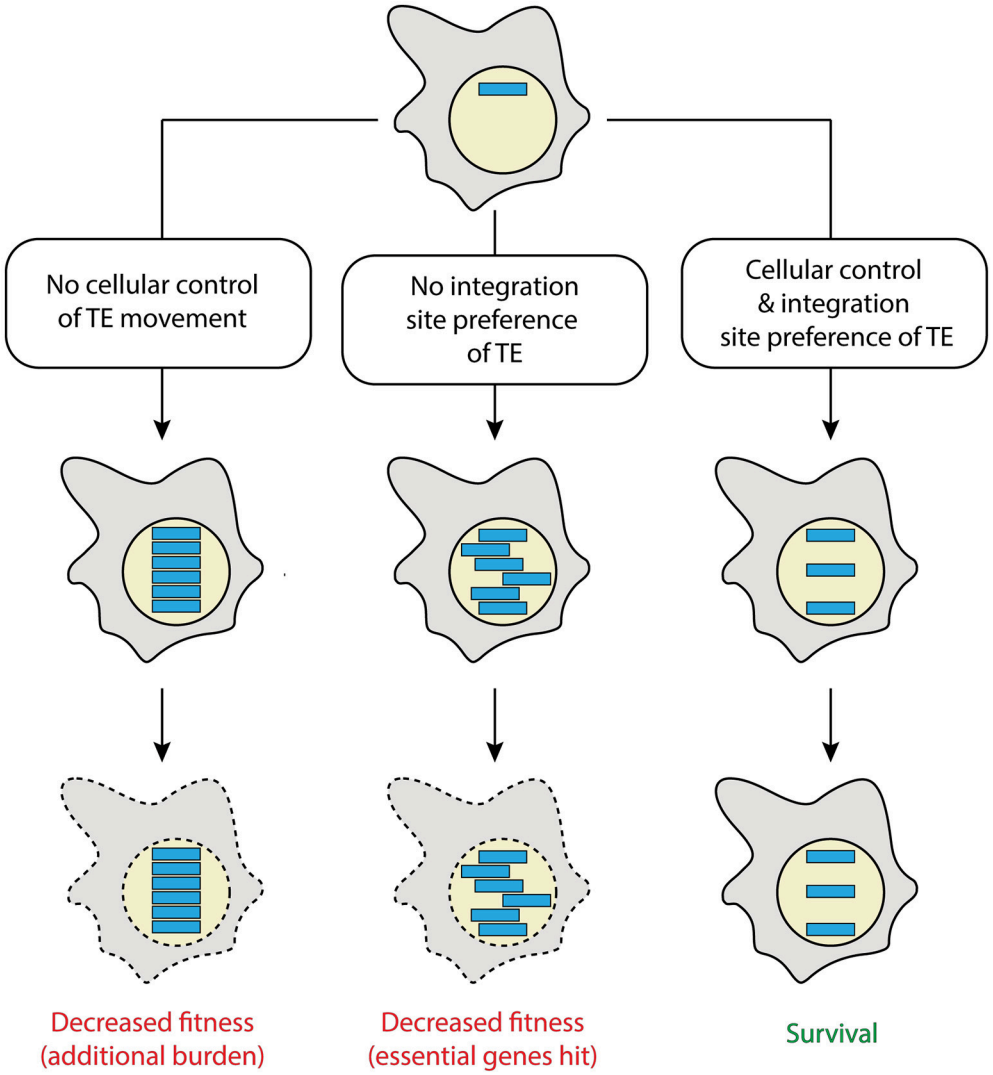


FIGURE 1 | Possible fates of transposable elements. Shown is a cell (gray) with a nucleus (yellow), in which a transposable element (TE; blue) invaded. Depending on the TE features and its control by cellular elements, different scenarios can be envisaged. Lack of cellular control of the TE will lead to its significant accumulation, posing additional burden to the host (left). Random integration (center) by lack of TE site preference is prone to hit essential genes. Compared to a cellularly controlled TE that additionally was domesticated by displaying a safe integration site preference (right), the former two scenarios would result in cells with decreased fitness. This is expected to eventually lead to their loss during evolution (indicated by dashed lines). A combination of the first two scenarios (not shown) is thought to result in a particularly disfavored cell.

other dictyostelid genomes (Spaller et al., 2016). In recent years, representatives of each class of *D. discoideum* retrotransposons have been investigated, as detailed next.

THE NON-LTR RETROTRANSPOSONS

The genome of the amoeba features a comparably large number of 418 transfer RNA (tRNA) genes. The two subfamilies of non-LTR retrotransposons in *Dictyostelium* target the up- and downstream regions of tRNA genes and have accordingly been named TRE5 and TRE3. The TREs represent 3.6% of the genome content (Glöckner et al., 2001). Both TRE subfamilies contain two ORFs (**Figure 2A**). A distinct integration distance of about 50 bp upstream and about 100 bp downstream of tRNA genes is observed for TRE5 and TRE3, respectively.

TRE5-A was the first identified TRE in the genome of *D. discoideum* (Marschalek et al., 1989; Winckler, 1998) and is understood best, mainly by work in the Winckler lab. The autonomous TRE5-A.1 contains two overlapping ORFs and three regulatory sequence modules A, B, and C (**Figure 2A**). TRE5-A ORF1 protein (ORF1p) physical interacts with subunits of the tRNA-gene specific transcription factors IIIB (TFIIIB), indicating that it might be involved in target site selection (Chung et al., 2007). Although it lacks sequence homology to ORF1p of other non-LTR retroelements (Glöckner et al., 2001), there is a certain functional correlation with ORF1p in the mammalian TE L1,

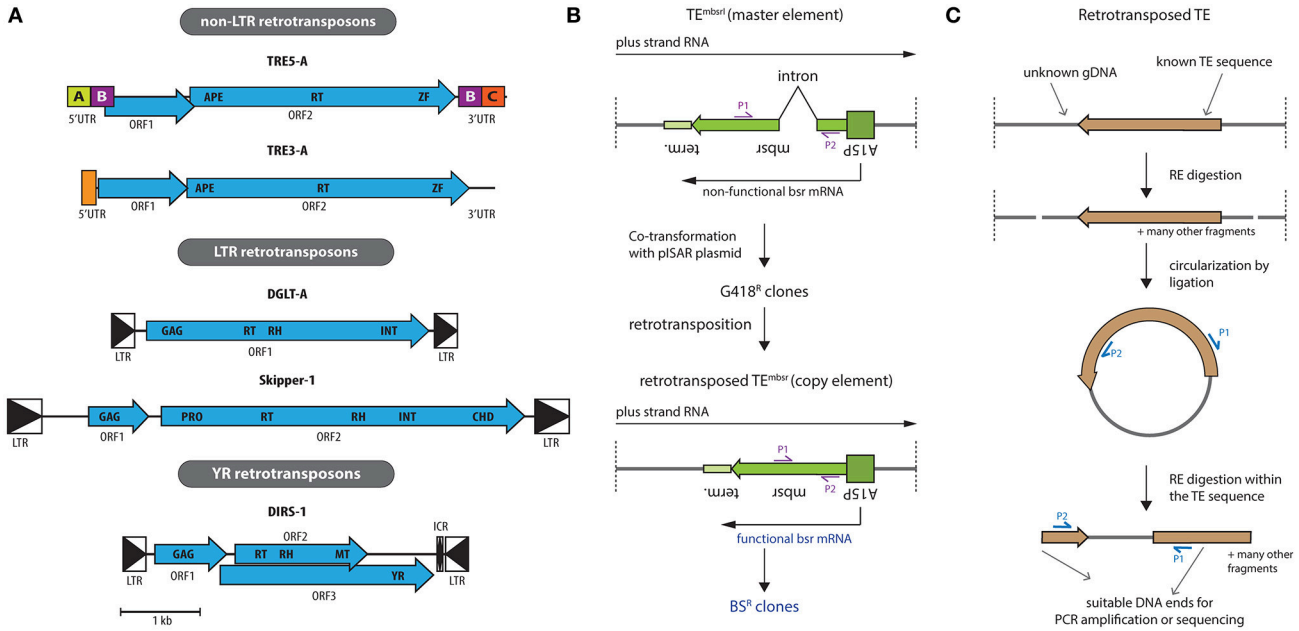


FIGURE 2 | Retrotransposons encoded in the *D. discoideum* genome and strategies for their investigation. **(A)** Shown are the structures of the consensus elements with open reading frames (ORFs) in blue, drawn to scale (see bottom). The non-LTR retrotransposon family is separated into two subgroups, namely TRE5 and TRE3 (for 5′ and 3′ tRNA gene targeted retroelement, respectively), based on their integration preferences upstream or downstream of tRNA genes. Typical for all TRE elements is the presence of two ORFs, where the second one encodes for an apurinic/apyrimidinic endonuclease (APE), a reverse transcriptase (RT), and a zinc-finger domain (ZF). The bordering short untranslated regions (UTR) are indicated, which contain in the TRE5 subfamily distinct modules A–C. The LTR retrotransposons represent Ty3/gypsy-like elements that are surrounded by directed long terminal repeats (LTR, boxed triangles). DGLT-A has a single ORF encoding for a group-specific antigen (GAG) protein with a zinc finger-like signature followed by RT, ribonuclease H (RH), and integrase (INT) domains. Skipper-1 contains two ORFs. ORF1 codes for a GAG protein with a defined zinc finger-like motif. ORF2 codes for a protease (PRO), RT, RH, INT. Additionally, ORF2 of Skipper-1 contains a chromo domain (CHD) at the C-terminus. Note that all copies of DGLT-A and Skipper-2 in the genome of *D. discoideum* are incomplete, contrary to other dictyostelids (Spaller et al., 2016). DIRS-1 is the founding member of the class of tyrosine recombinase (YR) retrotransposons. Its three ORFs are surrounded by inverted LTRs. ORF1 encodes a putative GAG protein, ORF2 overlaps with ORF3 and encodes RT, RH and methyl transferase (MT) domains, while ORF3 contains the YR. **(B)** Schematic representation of the retrotransposition assay with a genetically tagged transposable element TE^{mbsrI} (mbsrI: minus-strand blasticidin S resistance (bsr) gene disrupted by an inverse intron), as applied successfully in the investigation of *Ty1* from *Saccharomyces cerevisiae*, mammalian LINEs and TRE5-A from *D. discoideum* (Boeke et al., 1985; Esnault et al., 2000; Ostertag et al., 2000; Siol et al., 2011). Shown is the orientation of the resistance cassette (referred to as "*master element*" when disrupted by an intron) with respect to the TE sequence, in which the cassette is embedded. A15P denotes the actin15 promoter and terminator sequence (term.) is shown. Upon co-transformation with a marker plasmid, e.g., pISAR, which confers G418 resistance to *D. discoideum*. Transformants with stably integrated plasmids are selected with G418. The *master element* is not able to generate blasticidin S (BS) resistant clones, because the respective gene is inactivated by an intron (from the *S17* gene, 74 bp in size). After a complete retrotransposition cycle, however, the intron is spliced out, resulting in TE^{mbsr} (minus-strand bsr gene), referred to as "*copy element*" upon loosing the intron in the resistance gene sequence. This enables expression of a functional bsr messenger RNA (mRNA), thereby conferring BS resistance (BS^R). BS resistant clones thus are cells, in which at least one full retrotransposition cycle of the TE under investigation has been performed. Primers P1 and P2 can be used to confirm the splicing of the intron by PCR, and the size of the PCR product is indicative for the *master element* or the *copy element*. **(C)** Mapping of novel integration sites using tagged TEs. To address integration sites, genomic DNA (gDNA) is digested using restriction enzymes (RE) that cut outside the TE. Upon circularization of resulting fragments by ligation and RE digestion inside the TE sequence, linear fragments are obtained. These can be analyzed by PCR employing primers P1 and P2, or directly subjected to next generation sequencing.

which is also involved in genomic integration, though not at tRNA genes (Kolosha and Martin, 2003; Martin et al., 2008). Regulatory sequences of tRNA genes, in particular B-boxes at their 5′ end, are sufficient for TRE5-A targeting, even in the absence of a tRNA gene (Siol et al., 2006b). As TFIIIB binds to these sequences, it seems plausible that TRE5-A has hijacked RNA polymerase III transcription factors for its targeting.

ORF2 encodes for a polyprotein containing the enzymatic activities that are required for the retrotransposition cycle (**Figure 2A**; Winckler et al., 2011). The regulatory A module has RNA polymerase II promoter activity, the B module harbors the translation start site for ORF1 and the C module is required for retrotransposition (Marschalek et al., 1992; Schumann et al., 1994; Siol et al., 2011). The C module also harbors an internal promoter for the generation of antisense transcripts (Schumann et al., 1994). In recent years, the Winckler lab has established a genetically traceable version, TRE5-A^{bsr}, schematically shown in **Figure 2B**. This proved most helpful in studying various aspects of TRE5-A retrotransposition (Siol et al., 2011). The A module could be replaced by an artificial promoter, indicating that this is its sole function. TRE5-A^{bsr} is based on the non-autonomous TRE5-A.2 variant, which lacks the ORF2 sequence, but is mobilized *in trans* by the ORF2p of endogenous TRE5-A.1 (Beck et al., 2002). Subsequently, cloning and sequencing of

de novo integration sites of TRE5-A^{bsr} revealed the authentic positioning around 50 nucleotides upstream of tRNA genes (Beck et al., 2002; Siol et al., 2006a). Additionally, this construct was also instrumental to realize that, in principle, all tRNA genes can be targeted (Spaller et al., 2017). Whether integration alters tRNA gene expression cannot be easily investigated due to their high abundance and redundancy in *D. discoideum*. Unexpected TRE5-A^{bsr} integration sites in the extrachromosomal rDNA palindrome of the amoeba (Sucgang et al., 2003) were also uncovered, which are characterized by perfect B box sequences (Siol et al., 2011). Subsequently, additional integration sites in the vicinity of the RNA polymerase III-transcribed ribosomal 5S gene were characterized (Spaller et al., 2017). Taken together, these data strongly point towards a general coupling of TRE5-A integration with active RNA polymerase III transcription. As such, integration of TRE5-A would depend on the presence of regulatory A/B box sequences of tRNA genes, rather than the tRNA genes themselves.

Additionally, a host factor (CbfA for C-module binding Factor A; Geier et al., 1996) supports active TRE5-A retrotransposition, by stabilizing or upregulating TRE5-A sense and antisense transcripts (Bilzer et al., 2011). This transcription factor is also essential for the multicellular development of *D. discoideum* by transcriptionally activating the aggregation-specific adenylyl cyclase ACA (Winckler et al., 2004; Siol et al., 2006b). Later, CbfA was characterized as a general transcriptional regulator, with more than 1000 genes being differentially regulated at least 3-fold in a strain with largely reduced CbfA protein amounts. Amongst these was *agnC,* the gene encoding the Argonaute protein C, which experienced a more than 200-fold upregulation (Schmith et al., 2013). Argonaute proteins are key components of the RNAi machinery (Hutvagner and Simard, 2008), and therefore, this observation was of particular interest, as it opened the possibility that TRE5-A might be regulated by RNAi components. This holds particularly true, as complementary sense and antisense TRE5-A RNAs are present (Bilzer et al., 2011). Next to Argonaute proteins, RNA-dependent RNA polymerases (RdRPs) and Dicer proteins are key components of RNAi, and several representatives of these families exist in *D. discoideum* (Martens et al., 2002; Kuhlmann et al., 2005; Boesler et al., 2014; Wiegand et al., 2014; Kruse et al., 2016; Meier et al., 2016). Indeed, TRE5-A was found overexpressed in an *agnC* deletion strain, and downregulated in an AgnC overexpressing strain, resulting in a reduced retrotransposition rate (Schmith et al., 2015). No indication was found for an involvement of the three RdRPs of the amoeba, nor of its two Dicer proteins. This suggests that a distinct, AgnC-dependent RNAi pathway controls TRE5-A amplification, which, however, is counteracted by CbfA, resulting effectively in an active TRE5-A population in wildtype *Dictyostelium* (Beck et al., 2002; Siol et al., 2006a).

THE LTR RETROTRANSPOSONS

The *D. discoideum* genome features two related LTR retrotransposon families, Skipper and DGLT-A (Spaller et al., 2016). Both are Ty3/gypsy-like retotransposons that share the enzymatic activities required for retrotransposition (**Figure 2A**). These are, however, organized as one ORF in DGLT-A, and spread over two ORFs in Skipper. The *D. discoideum* genome does not feature any full-length DGLT-A copies, indicating that the TE might no longer be able to amplify (Winckler et al., 2005). Intriguingly, the DGLT-A elements in *D. discoideum* are also found 13–33 bp upstream of tRNA genes. Thus, two unrelated TEs, DGLT-A and TRE5 both integrate upstream of tRNA genes, indicating convergent evolution. A recent study expanded this view: in the evolution of dictyostelids, selection of tRNA genes as TE target was invented independently at least six times (Spaller et al., 2016).

Skipper is distinct from DGLT-A not only by the structural organization of its ORFs, but also by the presence of a chromo domain (CHD). Recent data indicated that Skipper retrotransposons in dictyostelids come in two varieties. Skipper-1 contains a conventional CHD and is found as largely fragmented elements in centromeric regions of the chromosomes (Glöckner and Heidel, 2009); also the related DGLT-P element harbors a CHD, resulting in a name change to Skipper-2, despite the fact that this CHD has somewhat diverged. Skipper-2 is found downstream of tRNA genes, similar to the TRE3 elements, another example of the convergent evolution of this target selection (Spaller et al., 2016).

CHDs are known to target retrotransposons to heterochromatin (Gao et al., 2008). In line with this, centromeric sequences in *D. discoideum* are characterized by heterochromatic H3K9 methylation marks (Kaller et al., 2007), and Skipper-1 co-localizes with the centromeric histone variant cenH3 (Dubin et al., 2010). At present, it is unknown though, whether centromeric Skipper-1 targeting is an active CHD-mediated process. Alternatively, the apparent centromeric accumulation might be an indirect effect, resulting from loss of such cells from the population, in which Skipper-1 integrated in other genomic positions, as this might cause mutations in the gene-dense genome.

Skipper-1 was previously shown to be under the transcriptional control by DNA methylation, as Skipper-1 transcripts accumulated in the *dnmA* gene deletion strain, resulting in an increase of its genomic copy numbers. Additionally, components of the RNAi machinery appear to control Skipper-1 post-transcriptionally (Kuhlmann et al., 2005). Söderbom and co-workers noticed an extended hairpin derived of a Skipper-1 fragment, which might be the source of the observed small Skipper-1 RNAs (Hinas et al., 2007). Mechanistic details of Skipper-1 integration into centromeric heterochromatin are currently not available, nor models on how the transcriptional and post-transcriptional control mechanisms might be intertwined to result in only two intact Skipper-1 copies in the *D. discoideum* genome (Spaller et al., 2016).

DIRS-1

Albeit featuring LTRs, the Dictyostelium Intermediate Repeat Sequence (DIRS-1) is the founding member of its own class

of TEs as it features a tyrosine recombinase (YR) instead of a canonical integrase (INT) (Cappello et al., 1985; Poulter and Goodwin, 2005). The enzyme is thought to integrate into the genome circular intermediates (Poulter and Goodwin, 2005), the existance of which we recently verified experimentally (Boesler et al., 2014). Full length DIRS-1 contains three, partially overlapping ORFs that are surrounded by two inverted LTRs (**Figure 2A**). DIRS-1 is the most frequently occurring retrotransposon in *D. discoideum* and this expansion appears unique amongst dictyostelids (Spaller et al., 2016). As seen for Skipper-1, DIRS-1 localizes to centromers (Dubin et al., 2010), of which it constitutes 50% of sequence content (Glöckner and Heidel, 2009). This accumulation has been attributed to the YR, which might facilitate homologous recombination into existing copies (Cappello et al., 1984).

A potentially important feature of DIRS-1 is the internal complementary region (ICR; **Figure 2A**), a non-coding sequence that displays complementary to the 5′ end of the left LTR and to the 3′ end of the right LTR (Cappello et al., 1985; Poulter and Goodwin, 2005). DIRS-1 is transcriptionally active (Cappello et al., 1985) and like for many other retroelements, its LTR sequences serve as promoters (Wiegand et al., 2014), a feature that was recently applied in a knock-down system (Friedrich et al., 2015). For DIRS-1, the inverted orientation of the LTRs results in both sense and antisense transcripts (Wiegand et al., 2014). The sense transcript represents an incomplete copy of DIRS-1, with a small fragment of the left LTR and most of the right LTR missing (Cappello et al., 1985). A mechanism for DIRS-1 replication was proposed (Cappello et al., 1985; Poulter and Goodwin, 2005), but so far experimentally not fully proven. In this, the missing LTR sequences would be reconstituted by using the complementary ICR as template during cDNA synthesis. Upon self-ligation and formation of circular cDNA, a double-stranded molecule would be generated, allowing for site-specific recombination. The last step of this model is indirectly supported by DIRS-1 preferentially targeting existing genomic copies of itself, without apparent sequence preference (Cappello et al., 1984).

Unlike Skipper-1, DIRS-1 appears to be exclusively under post-transcriptional control by components of the RNAi machinery, in particular the RdRP RrpC and the argonaute AgnA (Boesler et al., 2014; Wiegand et al., 2014). In the absence of these two proteins, the amounts of endogenous small (21mer) DIRS-1 RNAs are largely reduced, concurrent with an accumulation of full length and shorter DIRS-1 mRNAs. Southern blot analysis suggested novel DIRS-1 integrations as consequence of the missing post-transcriptional silencing. The small DIRS-1 RNAs observed in the wildtype represent the majority of the small RNA population in *D. discoideum* (Hinas et al., 2007). They are asymmetrically distributed over the DIRS-1 element, and in particular the region of ORF1 appears devoid of significant amounts of small RNAs (Wiegand et al., 2014). As a consequence, GFP fusions of only ORF1, but not of the other two ORFs are translated in the wildtype, while all three GFP fusions can be readily obtained in strains lacking RrpC or AgnA (Boesler et al., 2014; Wiegand et al., 2014). The molecular phenotype with respect to DIRS-1 thus appears to be highly similar in strains lacking these two proteins. The presence of a circular cDNA copy, that is part of the proposed replication mechanism (Cappello et al., 1985; Poulter and Goodwin, 2005), however, has so far only been experimentally shown for an *agnA* gene deletion strain (Boesler et al., 2014), but not yet addressed in strains lacking RrpC.

FUTURE PERSPECTIVES

Based on the highly successful TRE5-A^{bsr} element (Siol et al., 2011), we suggest that similar constructs might be instructive in studying Skipper-1 and DIRS-1 (**Figure 2B**). To clone the consensus sequence of either element might represent a challenge due to the A/T-richness of the *D. discoideum* genome. In line with this, Siol et al. observed instability of tagged TRE5-A.1 sequences on plasmids (Siol et al., 2011). Potentially, however, this might be overcome by gene synthesis. While we had reported DIRS-1 and Skipper-1 transcript accumulation in respective mutant strains (Kuhlmann et al., 2005; Boesler et al., 2014; Wiegand et al., 2014), this is not necessarily indicative for retrotransposition competence.

Having tagged versions (**Figure 2B**) would not only allow to investigate whether the elements can perform full transposition cycles in wildtype and RNAi mutant strains. Additionally, the functional relevance of specific sequence elements in the individual retrotransposons could be addressed. For DIRS-1, such experiments might employ a tagged element lacking the ICR (**Figure 2A**), to address its requirement for the generation of a full length circular cDNA (Cappello et al., 1985; Poulter and Goodwin, 2005; Boesler et al., 2014). Likewise, the functionality of the two Skipper CHDs might be investigated (**Figure 2A**) to determine if they are important for the observed integration sites.

Finally, also novel DIRS-1 and Skipper-1 integration sites might be mapped (**Figure 2C**). For this, the sequence of the inserted resistance cassette might be experimentally addressed by Southern blotting or inverse PCR (**Figure 2C**), thereby discriminating between old and novel integration sites.

CONCLUSION

The contemporary retroelements present in the genome of *D. discoideum* are all found in comparably safe integration sites, either in the vicinity of tRNA genes, or in centromeric sequences, thereby largely preventing mutational insertions. Presumably, the gene-dense genome of the amoeba did not tolerate any retrotransposon with lacking integration specificity during evolution. The three best-studied retroelements, TRE5-A, Skipper-1, and DIRS-1, representing the major retroelement classes of this amoeba (**Figure 2A**), are all cellularly controlled by distinct components of the RNAi machinery. This points towards tailor-made cellular responses to the idiosyncrasies of the individual retrotransposon. The observation that the RNAi component AgnC, which acts in the regulation of TRE5-A, is itself regulated by a host factor that is involved in TRE5-A retrotransposition, points toward a complex, interacting

control network, rather than a linear control featuring two components. Whether similar control networks exist also for other retrotransposons stands to be determined in future work.

AUTHOR CONTRIBUTIONS

MM and MI drafted the manuscript and designed the figures. CH wrote the manuscript.

FUNDING

This work is supported by a grant of the Tönjes-Vagt-Stiftung Bremen, Germany.

ACKNOWLEDGMENTS

We thank Dr. Monica Hagedorn for helpful comments on the manuscript.

REFERENCES

Adl, S. M., Simpson, A. G., Lane, C. E., Lukes, J., Bass, D., Bowser, S. S., et al. (2012). The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* 59, 429–493. doi: 10.1111/j.1550-7408.2012.00644.x

Beck, P., Dingermann, T., and Winckler, T. (2002). Transfer RNA gene-targeted retrotransposition of Dictyostelium TRE5-A into a chromosomal UMP synthase gene trap. *J. Mol. Biol.* 318, 273–285. doi: 10.1016/S0022-2836(02)00097-9

Bilzer, A., Dolz, H., Reinhardt, A., Schmith, A., Siol, O., and Winckler, T. (2011). The C-module-binding factor supports amplification of TRE5-A retrotransposons in the *Dictyostelium discoideum* genome. *Eukaryot. Cell* 10, 81–86. doi: 10.1128/EC.00205-10

Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. S. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Res.* 23, 415–420. doi: 10.1007/s10577-015-9499-z

Boeke, J. D., Garfinkel, D. J., Styles, C. A., and Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell* 40, 491–500. doi: 10.1016/0092-8674(85)90197-7

Boesler, B., Meier, D., Foerstner, K. U., Friedrich, M., Hammann, C., Sharma, C. M., et al. (2014). Argonaute proteins affect siRNA levels and accumulation of a novel extrachromosomal DNA from the Dictyostelium retrotransposon DIRS-1. *J. Biol. Chem.* 289, 35124–35138. doi: 10.1074/jbc.M114.612663

Cai, H., and Devreotes, P. N. (2011). Moving in the right direction: how eukaryotic cells migrate along chemical gradients. *Semin. Cell Dev. Biol.* 22, 834–841. doi: 10.1016/j.semcdb.2011.07.020

Calvo-Garrido, J., Carilla-Latorre, S., Kubohara, Y., Santos-Rodrigo, N., Mesquita, A., Soldati, T., et al. (2010). Autophagy in dictyostelium: genes and pathways, cell death and infection. *Autophagy* 6, 686–701. doi: 10.4161/auto.6.6.12513

Cappello, J., Cohen, S. M., and Lodish, H. F. (1984). Dictyostelium transposable element DIRS-1 preferentially inserts into DIRS-1 sequences. *Mol. Cell. Biol.* 4, 2207–2213. doi: 10.1128/MCB.4.10.2207

Cappello, J., Handelsman, K., and Lodish, H. F. (1985). Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell* 43, 105–115. doi: 10.1016/0092-8674(85)90016-9

Castel, S. E., and Martienssen, R. A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.* 14, 100–112. doi: 10.1038/nrg3355

Castro-Diaz, N., Friedli, M., and Trono, D. (2015). Drawing a fine line on endogenous retroelement activity. *Mob. Genet. Elements* 5, 1–6. doi: 10.1080/2159256X.2015.1006109

Chung, T., Siol, O., Dingermann, T., and Winckler, T. (2007). Protein interactions involved in tRNA gene-specific integration of *Dictyostelium discoideum* non-long terminal repeat retrotransposon TRE5-A. *Mol. Cell. Biol.* 27, 8492–8501. doi: 10.1128/MCB.01173-07

Dubin, M., Fuchs, J., Graf, R., Schubert, I., and Nellen, W. (2010). Dynamics of a novel centromeric histone variant CenH3 reveals the evolutionary ancestral timing of centromere biogenesis. *Nucleic Acids Res.* 38, 7526–7537. doi: 10.1093/nar/gkq664

Duncan, L., Bouckaert, K., Yeh, F., and Kirk, D. L. (2002). kangaroo, a mobile element from *Volvox carteri*, is a member of a newly recognized third class of retrotransposons. *Genetics* 162, 1617–1630.

Eichinger, L., Pachebat, J. A., Glöckner, G., Rajandream, M. A., Sucgang, R., Berriman, M., et al. (2005). The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435, 43–57. doi: 10.1038/nature03481

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* 24, 363–367. doi: 10.1038/74184

Fets, L., Kay, R., and Velazquez, F. (2010). Dictyostelium. *Curr. Biol.* 20, R1008–R1010. doi: 10.1016/j.cub.2010.09.051

Friedli, M., and Trono, D. (2015). The developmental control of transposable elements and the evolution of higher species. *Annu. Rev. Cell Dev. Biol.* 31, 429–451. doi: 10.1146/annurev-cellbio-100814-125514

Friedrich, M., Meier, D., Schuster, I., and Nellen, W. (2015). A simple retroelement based knock-down system in Dictyostelium: further insights into RNA interference mechanisms. *PLoS ONE* 10:e0131271. doi: 10.1371/journal.pone.0131271

Gao, D., Jiang, N., Wing, R. A., Jiang, J., and Jackson, S. A. (2015). Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Front. Plant Sci.* 6:216. doi: 10.3389/fpls.2015.00216

Gao, X., Hou, Y., Ebina, H., Levin, H. L., and Voytas, D. F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359–369. doi: 10.1101/gr.7146408

Gbadegesin, M. A. (2012). Transposable elements in the genomes, parasites, junks or drivers of evolution? *Afr. J. Med. Med. Sci.* 41(Suppl.), 13–25.

Geier, A., Horn, J., Dingermann, T., and Winckler, T. (1996). A nuclear protein factor binds specifically to the 3′-regulatory module of the long-interspersed-nuclear-element-like Dictyostelium repetitive element. *Eur. J. Biochem.* 241, 70–76. doi: 10.1111/j.1432-1033.1996.0070t.x

Glöckner, G., and Heidel, A. J. (2009). Centromere sequence and dynamics in *Dictyostelium discoideum*. *Nucleic Acids Res.* 37, 1809–1816. doi: 10.1093/nar/gkp017

Glöckner, G., Szafranski, K., Winckler, T., Dingermann, T., Quail, M. A., Cox, E., et al. (2001). The complex repeats of *Dictyostelium discoideum*. *Genome Res.* 11, 585–594. doi: 10.1101/gr.GR-1622RR

Goodier, J. L., and Kazazian, H. H. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* 135, 23–35. doi: 10.1016/j.cell.2008.09.022

Goodwin, T. J., and Poulter, R. T. (2001). The DIRS1 group of retrotransposons. *Mol. Biol. Evol.* 18, 2067–2082. doi: 10.1093/oxfordjournals.molbev.a003748

Hinas, A., Reimegard, J., Wagner, E. G., Nellen, W., Ambros, V. R., and Soderbom, F. (2007). The small RNA repertoire of *Dictyostelium discoideum* and its regulation by components of the RNAi pathway. *Nucleic Acids Res.* 35, 6714–6726. doi: 10.1093/nar/gkm707

Huang, C. R., Burns, K. H., and Boeke, J. D. (2012). Active transposition in genomes. *Annu. Rev. Genet.* 46, 651–675. doi: 10.1146/annurev-genet-110711-155616

Hutvagner, G., and Simard, M. J. (2008). Argonaute proteins: key players in RNA silencing. *Nat. Rev. Mol. Cell Biol.* 9, 22–32. doi: 10.1038/nrm2321

Kaller, M., Foldesi, B., and Nellen, W. (2007). Localization and organization of protein factors involved in chromosome inheritance in *Dictyostelium discoideum*. *Biol. Chem.* 388, 355–365. doi: 10.1515/BC.2007.047

Kapitonov, V. V., and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* 9, 411–412; author reply: 414. doi: 10.1038/nrg2165-c1

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science* 303, 1626–1632. doi: 10.1126/science.1089670

Kolosha, V. O., and Martin, S. L. (2003). High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J. Biol. Chem.* 278, 8112–8117. doi: 10.1074/jbc.M210487200

Kruse, J., Meier, D., Zenk, F., Rehders, M., Nellen, W., and Hammann, C. (2016). The protein domains of the Dictyostelium microprocessor that are required for correct subcellular localization and for microRNA maturation. *RNA Biol.* 13, 1000–1010. doi: 10.1080/15476286.2016.1212153

Kuhlmann, M., Borisova, B. E., Kaller, M., Larsson, P., Stach, D., Na, J., et al. (2005). Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. *Nucleic Acids Res.* 33, 6405–6417. doi: 10.1093/nar/gki952

Le Rouzic, A., and Capy, P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* 169, 1033–1043. doi: 10.1534/genetics.104.031211

Levis, R. W., Ganesan, R., Houtchens, K., Tolar, L. A., and Sheen, F. M. (1993). Transposons in place of telomeric repeats at a Drosophila telomere. *Cell* 75, 1083–1093. doi: 10.1016/0092-8674(93)90318-K

Marschalek, R., Brechner, T., Amon-Bohm, E., and Dingermann, T. (1989). Transfer RNA genes: landmarks for integration of mobile genetic elements in *Dictyostelium discoideum*. *Science* 244, 1493–1496. doi: 10.1126/science.2567533

Marschalek, R., Hofmann, J., Schumann, G., Gosseringer, R., and Dingermann, T. (1992). Structure of DRE, a retrotransposable element which integrates with position specificity upstream of *Dictyostelium discoideum* tRNA genes. *Mol. Cell. Biol.* 12, 229–239. doi: 10.1128/MCB.12.1.229

Martens, H., Novotny, J., Oberstrass, J., Steck, T. L., Postlethwait, P., and Nellen, W. (2002). RNAi in Dictyostelium: the role of RNA-directed RNA polymerases and double-stranded RNase. *Mol. Biol. Cell* 13, 445–453. doi: 10.1091/mbc.01-04-0211

Martin, S. L., Bushman, D., Wang, F., Li, P. W., Walker, A., Cummiskey, J., et al. (2008). A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. *Nucleic Acids Res.* 36, 5845–5854. doi: 10.1093/nar/gkn554

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.* 36, 344–355. doi: 10.1073/pnas.36.6.344

Meier, D., Kruse, J., Buttlar, J., Friedrich, M., Zenk, F., Boesler, B., et al. (2016). Analysis of the microprocessor in Dictyostelium: the role of RbdB, a dsRNA binding protein. *PLoS Genet.* 12:e1006057. doi: 10.1371/journal.pgen.1006057

Ostertag, E. M., Prak, E. T., DeBerardinis, R. J., Moran, J. V., and Kazazian, H. H. (2000). Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res.* 28, 1418–1423. doi: 10.1093/nar/28.6.1418

Poulter, R. T., and Goodwin, T. J. (2005). DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenet. Genome Res.* 110, 575–588. doi: 10.1159/000084991

Schmith, A., Groth, M., Ratka, J., Gatz, S., Spaller, T., Siol, O., et al. (2013). Conserved gene regulatory function of the carboxy-terminal domain of dictyostelid C-module-binding factor. *Eukaryot. Cell* 12, 460–468. doi: 10.1128/EC.00329-12

Schmith, A., Spaller, T., Gaube, F., Fransson, A., Boesler, B., Ojha, S., et al. (2015). A host factor supports retrotransposition of the TRE5-A population in Dictyostelium cells by suppressing an Argonaute protein. *Mob. DNA* 6:14. doi: 10.1186/s13100-015-0045-5

Schumann, G., Zundorf, I., Hofmann, J., Marschalek, R., and Dingermann, T. (1994). Internally located and oppositely oriented polymerase II promoters direct convergent transcription of a LINE-like retroelement, the Dictyostelium repetitive element, from *Dictyostelium discoideum*. *Mol. Cell. Biol.* 14, 3074–3084. doi: 10.1128/MCB.14.5.3074

Siol, O., Boutliliss, M., Chung, T., Glöckner, G., Dingermann, T., and Winckler, T. (2006a). Role of RNA polymerase III transcription factors in the selection of integration sites by the dictyostelium non-long terminal repeat retrotransposon TRE5-A. *Mol. Cell. Biol.* 26, 8242–8251. doi: 10.1128/MCB.01348-06

Siol, O., Dingermann, T., and Winckler, T. (2006b). The C-module DNA-binding factor mediates expression of the Dictyostelium aggregation-specific adenylyl cyclase ACA. *Eukaryot. Cell* 5, 658–664. doi: 10.1128/EC.5.4.658-664.2006

Siol, O., Spaller, T., Schiefner, J., and Winckler, T. (2011). Genetically tagged TRE5-A retrotransposons reveal high amplification rates and authentic target site preference in the *Dictyostelium discoideum* genome. *Nucleic Acids Res.* 39, 6608–6619. doi: 10.1093/nar/gkr261

Slotkin, R. K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285. doi: 10.1038/nrg2072

Spaller, T., Groth, M., Glockner, G., and Winckler, T. (2017). TRE5-A retrotransposition profiling reveals putative RNA polymerase III transcription complex binding sites on the Dictyostelium extrachromosomal rDNA element. *PLoS ONE* 12:e0175729. doi: 10.1371/journal.pone.0175729

Spaller, T., Kling, E., Glockner, G., Hillmann, F., and Winckler, T. (2016). Convergent evolution of tRNA gene targeting preferences in compact genomes. *Mob. DNA* 7:17. doi: 10.1186/s13100-016-0073-9

Sucgang, R., Chen, G., Liu, W., Lindsay, R., Lu, J., Muzny, D., et al. (2003). Sequence and structure of the extrachromosomal palindrome encoding the ribosomal RNA genes in Dictyostelium. *Nucleic Acids Res.* 31, 2361–2368. doi: 10.1093/nar/gkg348

Surcel, A., Kee, Y. S., Luo, T., and Robinson, D. N. (2010). Cytokinesis through biochemical-mechanical feedback loops. *Semin. Cell Dev. Biol.* 21, 866–873. doi: 10.1016/j.semcdb.2010.08.003

Unal, C., and Steinert, M. (2006). *Dictyostelium discoideum* as a model to study host-pathogen interactions. *Methods Mol. Biol.* 346, 507–515. doi: 10.1385/1-59745-144-4:507

Vos, J. C., De Baere, I., and Plasterk, R. H. (1996). Transposase is the only nematode protein required for *in vitro* transposition of Tc1. *Genes Dev.* 10, 755–761. doi: 10.1101/gad.10.6.755

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. doi: 10.1038/nrg2165

Wiegand, S., Meier, D., Seehafer, C., Malicki, M., Hofmann, P., Schmith, A., et al. (2014). The *Dictyostelium discoideum* RNA-dependent RNA polymerase RrpC silences the centromeric retrotransposon DIRS-1 post-transcriptionally and is required for the spreading of RNA silencing signals. *Nucleic Acids Res.* 42, 3330–3345. doi: 10.1093/nar/gkt1337

Winckler, T. (1998). Retrotransposable elements in the *Dictyostelium discoideum* genome. *Cell. Mol. Life Sci.* 54, 383–393. doi: 10.1007/s000180050168

Winckler, T., Iranfar, N., Beck, P., Jennes, I., Siol, O., Baik, U., et al. (2004). CbfA, the C-module DNA-binding factor, plays an essential role in the initiation of *Dictyostelium discoideum* development. *Eukaryot. Cell* 3, 1349–1358. doi: 10.1128/EC.3.5.1349-1358.2004

Winckler, T., Schiefner, J., Spaller, T., and Siol, O. (2011). Dictyostelium transfer RNA gene-targeting retrotransposons: studying mobile element-host interactions in a compact genome. *Mob. Genet. Elements* 1, 145–150. doi: 10.4161/mge.1.2.17369

Winckler, T., Szafranski, K., and Glöckner, G. (2005). Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact *Dictyostelium discoideum* genome. *Cytogenet. Genome Res.* 110, 288–298. doi: 10.1159/000084961

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

HYPOTHESIS AND THEORY
published: 04 December 2017
doi: 10.3389/fmicb.2017.02401

Sperm-Mediated Transgenerational Inheritance

*Corrado Spadafora**

Institute of Translational Pharmacology, National Research Council of Italy, Rome, Italy

Spermatozoa of virtually all species can spontaneously take up exogenous DNA or RNA molecules and internalize them into nuclei. In this article I review evidence for a key role of a reverse transcriptase (RT) activity, encoded by LINE-1 retrotransposons, in the fate of the internalized nucleic acid molecules and their implication in transgenerational inheritance. LINE-1-derived RT, present in sperm heads, can reverse-transcribe the internalized molecules in cDNA copies: exogenous RNA is reverse-transcribed in a one-step reaction, whereas DNA is first transcribed into RNA and subsequently reverse-transcribed. Both RNA and cDNA molecules can be delivered from sperm cells to oocytes at fertilization, further propagated throughout embryogenesis and inherited in a non-Mendelian fashion in tissues of adult animals. The reverse-transcribed sequences are extrachromosomal, low-abundance, and mosaic distributed in tissues of adult individuals, where they are variably expressed. These "retrogenes" are transcriptionally competent and induce novel phenotypic traits in animals. Growing evidence indicate that cancer tissues produce DNA- and RNA-containing exosomes. We recently found that these exosomes are released in the bloodstream and eventually taken up into epididymal spermatozoa, consistent with the emerging view that a transgenerational flow of extrachromosomal RNA connects soma to germline and, further, to next generation embryos. Spermatozoa play a crucial bridging role in this process: they act as collectors of somatic information and as delivering vectors to the next generation. On the whole, this phenomenon is compatible with a Lamarckian-type view and closely resembles Darwinian pangenesis.

Keywords: spermatozoa, LINE-1 retrotransposons, reverse transcriptase, exosomes, transgenerational inheritance, evolution

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria

Reviewed by:
Masato Tsurudome,
Mie University, Japan
Carlos M. Guerrero-Bosagna,
Linköping University, Sweden

***Correspondence:**
Corrado Spadafora
corrado.spadafora@gmail.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *15 August 2017*
Accepted: *20 November 2017*
Published: *04 December 2017*

Citation:
Spadafora C (2017) Sperm-Mediated Transgenerational Inheritance. Front. Microbiol. 8:2401. doi: 10.3389/fmicb.2017.02401

SPERMATOZOA AS A SOURCE OF REVERSE TRANSCRIPTASE-MEDIATED EXTRACHROMOSOMAL INFORMATION: A LOOK TO THE PAST

It is a well-established notion that mature spermatozoa have the spontaneous ability to take up exogenous DNA molecules and to internalize them in their nuclei (reviewed by Spadafora, 1998). This permeability is a distinctive feature of spermatozoa, both epididymal and ejaculated (after wash-off of seminal fluid), from virtually all animal species including humans

(Smith and Spadafora, 2005). Thus, in spite of the highly compact and impenetrable structure of their nuclei, sperm cells are in fact highly permeable to foreign molecule intrusion. Intense investigations of this phenomenon revealed that the interaction of exogenous DNA molecules with sperm cells, as well as their subsequent nuclear internalization, are well-regulated processes mediated by a network of specific factors (Spadafora, 1998). Parallel studies have revealed that spermatozoa can also take up RNA molecules and internalize them in their nuclei. Somewhat unexpectedly, these RNAs are reverse-transcribed into cDNA copies by a biologically active reverse transcriptase (RT) activity encoded by LINE-1 retrotransposons and present in sperm nuclei (Giordano et al., 2000; Spadafora, 2008). The LINE-1-derived RT interplays with a DNA-dependent RNA polymerase, also present in spermatozoa (Fuster et al., 1977), which together amplify the cDNAs copy number, mimicking a "natural" PCR/RT-PCR process. Most newly generated cDNA copies are released from spermatozoa into the medium and can be taken up again by further spermatozoa and internalized in their nuclei. Through this continuously cycling process, cDNA copies are evenly distributed among the vast majority of sperm cells suspension. Work with murine models showed that the RT-generated cDNAs are: delivered to oocytes at fertilization (Giordano et al., 2000; Pittoggi et al., 2006), maintained as low-copy number (below one copy/genome) non-integrated extrachromosomal sequences throughout development, mosaic propagated in the tissues of founder individuals, eventually transmitted in a non-Mendelian fashion to the next generation, transcriptionally competent and able to generate phenotypic variations in animals of both generations (Sciamanna et al., 2003; Pittoggi et al., 2006). These results suggest that spermatozoa provide a previously unrecognized source of RT-mediated information, not linked to chromosomal genes, and, at the same time, act as propagating vectors throughout generations.

These findings raise several puzzling questions. First, does the ability of sperm cells to take up foreign nucleic acid molecules reflects an enforced behavior when they come in contact with RNA under conditions of *in vitro* assays, or else do spermatozoa naturally collect and carry foreign molecules under physiological conditions *in vivo*? Second, does the RT activity stored in spermatozoa represent a functionless remnant of ancestral retrotransposon activity, brought to new life in response to occasional intrusions of foreign molecules, or does it exert an extant physiological role in development? These two issues, i.e., the sperm permeability to exogenous RNA, and the sperm RT that uses the latter as a substrate for retrotranscription, raise the third key question of whether these phenomena are physiologically relevant or, in other words, whether they occur in nature to generate a source of novel information. To begin to address these issues, it was imperative to characterize the RNA population stored in spermatozoa and possibly identify its origin. In recent years, high-throughput technologies and next generation sequence analysis have revealed a highly complex composition of spermatozoal RNA, whose components are increasingly emerging as key players in epigenetic inheritance processes, as will be seen in more depth in the following paragraphs.

THE COMPLEX TRANSCRIPTIONAL LANDSCAPE OF MATURE SPERMATOZOA

Traditional views considered spermatozoa as transcriptionally silent cells (Grunewald et al., 2005) and sperm RNAs as negligible remnants produced during spermatogenesis. More recent data however contrast with these views, showing that mature spermatozoa in fact contain a complex population of coding RNAs, small non-coding RNA classes, and, finally, LINE-1, SINE/Alu, and LTR repeat-associated transcripts (Jodar et al., 2013; Sendler et al., 2013; Miller, 2014). Small non-coding RNAs account for a considerable proportion of spermatozoal RNA (Krawetz et al., 2011; Kawano et al., 2012), mainly represented by piRNAs produced during spermatogenesis, tsRNA (tRNA-derived), and to a lesser extent, microRNAs (miRNAs) are instead predominant in epididymal spermatozoa (Chen et al., 2016b). Importantly, the composition of the spermatozoal RNA population varies in response to paternal exposure to a variety of stressing conditions (Rodgers et al., 2013; Brieno-Enriquez et al., 2015), a circumstance that can have crucial consequences for the fate and health of the progeny. Most importantly, growing data are revealing that RNAs of somatic origin also contribute to the composition of the sperm RNA cargo in the form of selectively retained RNAs derived from soma-to-spermatozoa intercellular communication. This flow is mediated by a special class of epididymis-derived nanovesicles, called epididymosomes, which shuttle miRNAs and tRNA fragments from the epididymal tissue to mature sperm cells (Belleannée et al., 2013; Vojtech et al., 2014; Sharma et al., 2016). The shuttled sperm RNA, containing several 100s of developmentally relevant small RNAs, are the product of a "sieving" process, as their profiles are distinct from those of the surrounding soma (Reilly et al., 2016). The modulation of the sperm RNA content occurs during maturation of spermatozoa between the proximal and distal epididymal segments, and identifies the epididymis as a key site for the establishment of the sperm epigenome (Nixon et al., 2015).

We recently reported that epididymal spermatozoa can incorporate RNA from somatic cell-released exosomes: indeed, we found that human melanoma cells, engineered to express EGFP and inoculated in nude mice, release EGFP RNA-containing nanovesicles in the bloodstream of the animals; a proportion of that RNA reaches the epididymis and becomes internalized in sperm heads (Cossetti et al., 2014). This finding shows that the flow of RNA delivered to spermatozoa originates not only from the surrounding epididymal soma, but also from distant, unrelated districts of the body. Nanovesicles act as the ideal vectors of such delivery. Sperm heads are the final recipients of this extrachromosomal information due to their ability to spontaneously take up exogenous molecules, as mentioned above. On the whole, these data indicate that the

impenetrable Weismann barrier, considered for a long time as a cornerstone of modern genetics, can in fact be breached by nanovesicle-mediated flows of extrachromosomal RNA (Eaton et al., 2015).

BREAKING THE WEISMANN BARRIER: A SPERM-MEDIATED RNA-BASED FLOW CONNECTS SOMA TO THE NEXT GENERATION EMBRYOS

In a seminal article, Krawetz and collaborators (Ostermeier et al., 2004) first reported that the sperm-specific RNA cargo is delivered to oocytes at fertilization. That finding proved that not only the male genome, but also extrachromosomal RNA carried by sperms, contribute to the zygote formation. However, the sperm RNA *per se* is not strictly required for embryonic development, as parthenogenetic mice can be successfully generated by microinjecting haploid, or bimaternal embryonic stem cells in murine oocytes (Li et al., 2016; Zhong et al., 2016). The latter finding indicates that all the fundamental information to support the developmental program, from fertilization to adulthood, is linked to chromosomal genes.

A novel turn to the field is being provided by recent data indicating that the composition of sperm RNA reflects the lifestyle habits and carry the "memory" of paternal experiences; that RNA-based memory is transmissible to the offspring as paternally acquired characteristics, with the potential to affect the health and overall biological fate of the progeny (reviewed by Liebers et al., 2014; Klosin and Lehner, 2016). Of remarkable interest are recent experiments that have assessed the potential of sperm RNAs as transgenerational modifiers in response to parental environmental or stressing conditions (Carone et al., 2010; Rodgers et al., 2013, 2015), including diet (Fullston et al., 2013; Chen et al., 2016a; Huypens et al., 2016), cigarette smoke (Marczylo et al., 2012), odor sensitivity (Dias and Ressler, 2014), and cognitive and behavioral conditioning (Rodgers et al., 2013; Gapp et al., 2014). RNA was unambiguously identified as the transgenerational modifier in a large set of compelling experimental data, showing that offspring generated from normal zygotes microinjected with sperm RNA recapitulate the phenotypical traits of the RNA donor animals (Rassoulzadegan et al., 2006; Gapp et al., 2014; Grandjean et al., 2015; Chen et al., 2016a).

Together, these data show that inheritance is not exclusively linked to chromosomal genes. Indeed, a subtle yet effective flow of RNA is established between somatic tissues and the next generation embryos. Spermatozoa are the pivots, playing a dual role both as collectors of paternal extrachromosomal RNA and as their vectors to the offspring. The emerging evidence that RNA-based information can travel from soma to germline subvert the Weismann's theory and provide a foundation for the inheritance of acquired traits with far reaching implications for evolutionary processes.

RT ENCODED BY LINE-1 RETROTRANSPOSONS AS MODULATOR OF EARLY EMBRYONIC DEVELOPMENT

In addition to being stored in mature spermatozoa, LINE-1-encoded RT is also abundantly expressed in early embryos and is implicated in the genesis and propagation of extrachromosomal information. We have found that LINE-1 retrotransposon-encoded RT is triggered soon after fertilization in both zygotic pronuclei, predominantly in the paternal pronucleus (Vitullo et al., 2012), and remains active in early preimplantation embryos (Pittoggi et al., 2003). RT plays a crucial role in early development: indeed, RT inhibition, induced either by pharmacological RT inhibitors (Pittoggi et al., 2003), or by downregulating LINE-1 expression by microinjecting antisense oligonucleotides in zygotic pronuclei (Beraldi et al., 2006), causes a drastic arrest of embryo development at the 2- or 4-cell stages. These results suggest that RT is strictly necessary for the unfolding of the developmental program from the second cell division, as the first cleavage exploits the maternal RNA stored in oocytes (Tang et al., 2007).

Although neither the specific role(s) nor the mechanism of action of embryonic RT are yet fully clarified, emerging data suggest that RT controls the biogenesis of miRNAs, a class of RNA that is globally, yet transiently, suppressed in early embryogenesis (Suh et al., 2010), concomitant with the up-regulation of RT expression (Pittoggi et al., 2003; Vitullo et al., 2012).

The link between LINE-1-encoded RT and miRNA biogenesis has been investigated in some depth in cancer cells. In striking analogy with early embryos, RT is also highly expressed in most cancer types from very early stages (reviewed by Sinibaldi-Vallebona et al., 2011; De Luca et al., 2016). In parallel with high RT activity, the biogenesis of LINE-1-derived miRNAs (Lu et al., 2005) and siRNAs (Chen et al., 2012) is globally reduced in cancer compared to normal cells, with ensuing alterations of the gene expression regulatory network. Exposure of cancer cells to RT inhibitors restores the normal expression profile of miRNAs, with a direct impact on global gene expression (Sciamanna et al., 2013, 2014). These lines of evidence suggest therefore that high RT expression exerts: (i) a physiological control on the biogenesis of miRNA in early embryogenesis, and (ii) a pathological role in cells conveyed toward tumorigenesis, by impairing the production of miRNAs, with the ensuing dysregulation of downstream targets and the increase of transcriptional fluctuations.

THE REMODELING OF THE EMBRYONIC EPIGENETIC LANDSCAPE AND ITS IMPLICATION IN EVOLUTION. A MODEL

In recapitulating the aspects discussed so far a framework is beginning to emerge: (1) RNA-containing nanovesicles are released from somatic tissues into the bloodstream; (2)

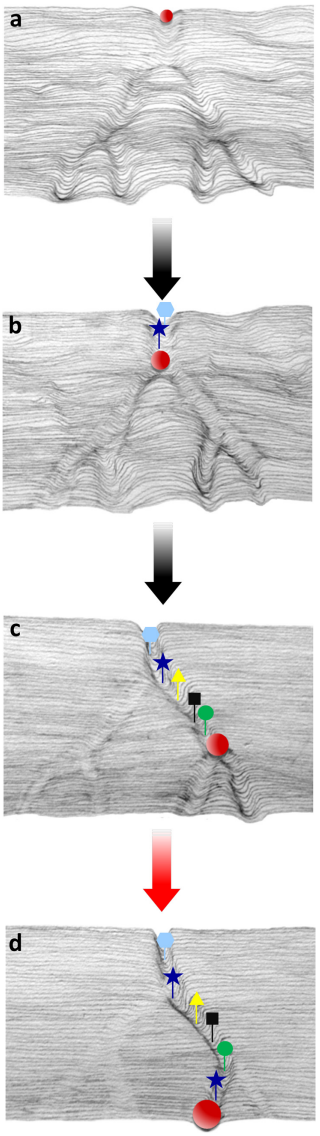


FIGURE 1 | Schematic illustration, inspired by the Waddington's model, representing the canalization process of the embryonic epigenetic landscape. The red ball rolling down from the top of the hill **(a–d)** throughout the landscape represents the preferential trajectory of the progressing epigenetic pathway driven by sperm RNAs delivered at fertilization throughout generations. Targeted retrotransposal insertions (colored shapes) follow the progressing groove **(b,c)**, and provide DNA recognition sites for gene function (transcription initiation, transcript splicing, etc.), leading to the assembly of functional regulatory networks. Insertions also favor the copy number expansion of miRNA and tsRNA coding sequences. The deeper groove with colored symbols represents the functionally active path, while the shallower grooves are not in use. The red ball in **(d)** indicates the completion of the remodeling path and the functional activation of a novel canalized pattern (red arrow).

Epididymal spermatozoa take up nanovesicles and internalize them in their nuclei; (3) The internalized RNA molecules are processed and their copy number is amplified, via the RT/DNA-dependent RNA polymerase interplay; (4) Somatic RNAs, or their cDNA copies, are delivered from sperm to embryos at fertilization.

The first three steps continuously renew the RNA storage in sperm heads. The last one, i.e., the delivery of processed somatic RNA to oocytes, can recur at each round of fertilization. Through this process sperm RNA is transmitted from one generation to the next, which can contribute to the embryo fitness and in principle expand the adaptation of the newborn to diverse environmental conditions. It is reasonable to assume that a large proportion of the males living in a same ecological niche and exposed to the same stimuli, produce sperm RNA cargos of similar composition; under constant environmental conditions, these RNA cargo would be continuously delivered to the progeny via fertilization throughout generations. It is not unreasonable to hypothesize that, in the long run (i.e., after a "sufficient" number of generations), the sperm RNAs promote the "assimilation" of new trait(s), to use Waddington's concept (Waddington, 1959). In other words, the cumulative effects of RNA delivery through generations may promote the emergence of novel functional "canalized" pathways (Waddington, 1959), via remodeling of the embryonic chromatin architecture; in consequence, novel genetic circuits might be activated and/or pre-existing ones might be "rewired." Mechanistically, the cumulative effects of regulatory miRNAs and tsRNAs delivered by sperm upon fertilization would drive the emergence of novel canalized pathways through two sequential steps: (i) first, by "rewiring" the expression profile of genes constituting canalized genetic circuits, and (ii) second, via targeted retrotransposition events that provide new regulatory sequences, which brings to completion the newly canalized circuits.

The first (epigenetic) step builds on the established regulatory functions of miRNAs and tsRNAs, which can modulate the expression of relevant genes. It is reasonable to hypothesize that RNAs delivered by sperm cells at fertilization also exert these regulatory functions and remodulate the gene expression profile in early embryos. Consequently, new genetic circuits (canalized circuits) become functionally active, or/and pre-existing ones are rewired.

Canalized circuits would then reach their final state through targeted retrotransposition events (genetic step). New retrotranspositions can provide additional layers of control, by inserting protein-binding sites (e.g., for transcription factors, hormones, splicing factors), as well as enhancers, promoters, insulators, etc. in new sites within the genome.

Thus, a hybrid epigenetic/genetic process drives the remodeling of the embryonic chromatin architecture, as schematized in **Figure 1**.

The process is seen as progressive in nature, based on the assumption that the "quanta" of sperm-derived regulatory RNAs delivered to the embryo at fertilization (represented by the red ball rolling down the groove in **Figure 1**) constitute minimal contributions toward the epigenetic activation of novel canalization(s) by "deepening the canal" (**Figures 1a–c**, right branch), while non-active pathways become shallower (**Figures 1a–d**, left branch). When the cumulative effects of the

sperm-delivered RNA overcome the buffering capacity of the embryos, the novel canalized genomic circuit (**Figure 1d**) has the potential to redirect the embryonic ontogenesis and generate phenotypic novelties.

Targeted retrotransposition events constitute the genetic component of the model, contributing to the functional reshaping of the embryo regulatory circuits. Targeted insertions (symbolized by different colored symbols in **Figure 1**) contribute to establish novel regulatory circuits in at least three ways: (i) they provide new protein-binding and regulatory sites; (ii) they contribute new miRNA-coding sequences that expand the overall diversity in the RNA population, and (iii) they stabilize the newly remodeled landscape by fixing the chromatin architecture.

Three considerations suggest that these changes could be permanently assimilated. First, zygotes and early embryos are thought to provide permissive, change-prone environments, consistent with the finding that the early embryonic genome is largely unstructured before zygotic genome activation, showing a low level of chromatin organization over long genomic distances (Hug et al., 2017). Second, as mentioned, LINE-1-encoded RT activity is high in preimplantation embryos (Vitullo et al., 2012) and, in parallel, the miRNA-based control system is globally suppressed (Suh et al., 2010). This is relevant in the light of evidence that miRNA-mediated control reduces random fluctuations in differentiated cells and in development, hence conferring robustness to genetic pathways (Li et al., 2009; Ebert and Sharp, 2012); on the contrary, miRNA suppression increases instability and random fluctuations in the developmental program (Hornstein and Shomron, 2006; Li et al., 2009; Ebert and Sharp, 2012). Moreover, retrotransposon families (i.e., LINE-1, Alus, LTRs) are de-repressed in embryos concomitant with global genomic hypomethylation (Lee et al., 2014; Smith et al., 2014), and constitute a potential source of both genetic and epigenetic variations (Macia et al., 2011; Vitullo et al., 2012; Fadloun et al., 2013). Overall, no massive retrotransposition events are required, with the exception of some crucial insertions that provide regulatory sequences to the newly formed canalized circuits. These crucial events would be targeted to specific hypersensitive sites generated during embryonic chromatin remodeling. Third, the RNA population that spermatozoa deliver to oocytes contains regulatory miRNAs and tsRNAs (Chen et al., 2016b), which can reshape the embryonic expression landscape and reprogram the transcription profiles of 100s of embryonic genes. Indeed, even small amounts of delivered regulatory RNAs can generate an ample spectrum of epigenetic variations with a potential impact on the phenotype. Thus, inheritable variations may be driven by small regulatory RNAs, assimilated in the change-prone genome architecture of embryos and translated into new phenotypic variants, with no major adverse effect on the permissive embryonic context.

It is worth recalling that small RNAs are involved both in macroevolutionary processes – as their number increases over time in parallel with complexity, while their loss is associated with morphological simplification (Wheeler et al., 2009; Erwin et al., 2011) – and with the canalization of genetic programs (Hornstein and Shomron, 2006; Li et al., 2009; Vidigal and Ventura, 2015).

CONCLUSION

The present model of transgenerational inheritance attempts to integrate data from different sources in a biologically coherent framework. Most aspects implicated in the process are experimentally tested and are potentially able to generate transgenerationally relevant novelties. The mechanism is predominantly epigenetic and independent of genomic mutations. Importantly, the "sieving force" of natural selection is not necessary in conventional terms, because the canalization process driven by sperm RNA would generate specific pathways, leaving little space for random variations, and favoring the ultimate emergence of one or few new phenotype(s). In analogy with Lamarckism, this hypothesis is based on the assumption that extrachromosomal transgenerational inheritance can affect ontogenesis and generate evolutionarily significant, stably acquired variations. Darwinian pangenesis (Holterhoff, 2014) is the other theory with which the model has significant overlap. The hypothesis that "gemmules" containing parental characters are released from tissues and transferred to the next generation via the germline can now be reinterpreted in the light of our current knowledge of circulating nanovesicles and exosomes, carrying nucleic acids and released from somatic tissues, which can be taken up by sperm cells, thus providing a foundation for spermatozoa-mediated transgenerational inheritance. It is amazing that so-called obsolete concepts, developed in the context of two historically rejected theories, are re-emerging from modern experimental data based on next generation genomic methodologies, thus confirming that history sometime repeats itself.

AUTHOR CONTRIBUTIONS

CS has conceived the model and wrote the manuscript.

FUNDING

This work was supported by a grant from Fondazione Roma to the Project "Investigating the cellular endogenous Reverse Transcriptase as a novel therapeutic target and an early tumor marker" to CS.

ACKNOWLEDGMENT

The author grateful to Michele Spadafora and Graziano Bonelli for skilful assistance with drawing preparation.

REFERENCES

Belleannée, C., Calvo, E., Caballero, J., and Sullivan, R. (2013). Epididymosomes convey different repertoires of microRNAs throughout the bovine epididymis. *Biol. Reprod.* 89:30. doi: 10.1095/biolreprod.113.110486

Beraldi, R., Pittoggi, C., Sciamanna, I., Mattei, E., and Spadafora, C. (2006). Expression of LINE-1 retroposons is essential for murine preimplantation development. *Mol. Reprod. Dev.* 3, 279–287. doi: 10.1002/mrd.20423

Brieno-Enriquez, M. A., Garcia-Lopez, J., Cardenas, D. B., Guibert, S., Cleroux, E., Ded, L., et al. (2015). Exposure to endocrine disruptor induces transgenerational epigenetic deregulation of microRNAs in primordial germ cells. *PLOS ONE* 10:e0124296. doi: 10.1371/journal.pone.0124296

Carone, B. R., Fauquier, L., Habib, N., Shea, J. M., Hart, C. E., Li, R., et al. (2010). Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell* 143, 1084–1096. doi: 10.1016/j.cell.2010.12.008

Chen, L., Dahlstrom, J. E., Lee, S. H., and Rangasamy, D. (2012). Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. *Epigenetics* 7, 758–771. doi: 10.4161/epi.20706

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., et al. (2016a). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science* 351, 397–400. doi: 10.1126/science.aad7977

Chen, Q., Yan, W., and Duan, E. (2016b). Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. *Nat. Rev. Genet.* 17, 733–743. doi: 10.1038/nrg.2016.106

Cossetti, C., Lugini, L., Astrologo, L., Saggio, I., Fais, S., and Spadafora, C. (2014). Soma-to-germline transmission of RNA in mice xenografted with human tumour cells: possible transport by exosomes. *PLOS ONE* 9:e101629. doi: 10.1371/journal.pone.0101629

De Luca, C., Guadagni, F., Sinibaldi-Vallebona, P., Sentinelli, S., Gallucci, M., Hoffmann, A., et al. (2016). Enhanced expression of LINE-1-encoded ORF2 protein in early stages of colon and prostate transformation. *Oncotarget* 7, 4048–4061. doi: 10.18632/oncotarget.6767

Dias, B. G., and Ressler, K. J. (2014). Parental olfactory experience influences behavior and neural structure in subsequent generations. *Nat. Neurosci.* 17, 89–96. doi: 10.1038/nn.3594

Eaton, S. A., Jayasooriah, N., Buckland, M. E., Martin, D. I. K., Cropley, J. E., and Suter, C. M. (2015). Roll over weismann: extracellular vesicles in the transgenerational transmission of environmental effects. *Epigenomics* 7, 1165–1171. doi: 10.2217/epi.15.58

Ebert, M. S., and Sharp, P. A. (2012). Roles for MicroRNAs in conferring robustness to biological processes. *Cell* 149, 515–524. doi: 10.1016/j.cell.2012.04.005

Erwin, D. H., Laflamme, M., Tweedt, S. M., Sperling, E. A., Pisani, D., and Peterson, K. J. (2011). The cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* 334, 1091–1097. doi: 10.1126/science.1206375

Fadloun, A., Le Gras, S., Jost, B., Ziegler-Birling, C., Takahashi, H., Gorab, E., et al. (2013). Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat. Struct. Mol. Biol.* 20, 332–338. doi: 10.1038/nsmb.2495

Fullston, T., Ohlsson Teague, E. M., Palmer, N. O., DeBlasio, M. J., Mitchell, M., Corbett, M., et al. (2013). Paternal obesity initiates metabolic disturbances in two generations of mice with incomplete penetrance to the F2 generation and alters the transcriptional profile of testis and sperm microRNA content. *FASEB J.* 27, 4226–4243. doi: 10.1096/fj.12-224048

Fuster, C. D., Farrell, D., Stern, F. A., and Hecht, N. B. (1977). RNA polymerase activity in bovine spermatozoa. *J. Cell Biol.* 74, 698–706. doi: 10.1083/jcb.74.3.698

Gapp, K., Jawaid, A., Sarkies, P., Bohacek, J., Pelczar, P., Prados, J., et al. (2014). Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nat. Neurosci.* 17, 667–669. doi: 10.1038/nn.3695

Giordano, R., Magnano, A. R., Zaccagnini, G., Pittoggi, C., Moscufo, N., Lorenzini, R., et al. (2000). Reverse transcriptase activity in mature spermatozoa of mouse. *J. Cell Biol.* 148, 1107–1113. doi: 10.1083/jcb.148.6.1107

Grandjean, V., Fourré, S., De Abreu, D. A. F., Derieppe, M.-A., Remy, J.-J., and Rassoulzadegan, M. (2015). RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders. *Sci. Rep.* 5:18193. doi: 10.1038/srep18193

Grunewald, S., Paasch, U., Glander, H. J., and Anderegg, U. (2005). Mature human spermatozoa do not transcribe novel RNA. *Andrologia* 37, 69–71. doi: 10.1111/j.1439-0272.2005.00656.x

Holterhoff, K. (2014). The history and reception of charles darwin's hypothesis of pangenesis. *J. Hist. Biol.* 47, 661–695. doi: 10.1007/s10739-014-9377-9377-0

Hornstein, E., and Shomron, N. (2006). Canalization of development by microRNAs. *Nat. Genet.* 38(Suppl.), S20–S24. doi: 10.1038/ng1803

Hug, C. B., Grimaldi, A. G., Kruse, K., and Vaquerizas, J. M. (2017). Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* 169, 216–228. doi: 10.1016/j.cell.2017.03.024

Huypens, P., Sass, S., Wu, M., Dyckhoff, D., Tschoep, M., Theis, F., et al. (2016). Epigenetic germline inheritance of diet-induced obesity and insulin resistance. *Nat. Genet.* 48, 497–499. doi: 10.1038/ng.3527

Jodar, M., Selvaraju, S., Sendler, E., Diamond, M. P., Krawetz, S. A., and Reproductive Medicine Network (2013). The presence, role and clinical use of spermatozoal RNAs. *Hum. Reprod. Update* 19, 604–624. doi: 10.1093/humupd/dmt031

Kawano, M., Kawaji, H., Grandjean, V., Kiani, J., and Rassoulzadegan, M. (2012). Novel small noncoding RNAs in mouse spermatozoa, zygotes and early embryos. *PLOS ONE* 7:e44542. doi: 10.1371/journal.pone.0044542

Klosin, A., and Lehner, B. (2016). Mechanisms, timescales and principles of trans-generational epigenetic inheritance in animals. *Curr. Opin. Genet. Dev.* 36, 41–49. doi: 10.1016/j.gde.2016.04.001

Krawetz, S. A., Kruger, A., Lalancette, C., Tagett, R., Anton, E., Draghici, S., et al. (2011). A survey of small RNAs in human sperm. *Hum. Reprod.* 26, 3401–3412. doi: 10.1093/humrep/der329

Lee, H. J., Hore, T. A., and Reik, W. (2014). Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* 14, 710–719. doi: 10.1016/j.stem.2014.05

Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S., and Carthew, R. W. (2009). A microRNA imparts robustness against environmental fluctuation during development. *Cell* 137, 273–282. doi: 10.1016/j.cell.2009.01.058

Li, Z., Wan, H., Feng, G., Wang, L., He, Z., Wang, Y., et al. (2016). Birth of fertile bimaternal offspring following intracytoplasmic injection of parthenogenetic haploid embryonic stem cells. *Cell Res.* 26, 135–138. doi: 10.1038/cr.2015.151

Liebers, R., Rassoulzadegan, M., and Lyko, F. (2014). Epigenetic regulation by heritable RNA. *PLOS Genet.* 10:e1004296. doi: 10.1371/journal.pgen.1004296

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature* 435, 834–838. doi: 10.1038/nature03702

Macia, A., Muñoz-Lopez, M., Cortes, J. L., Hastings, R. K., Morell, S., Lucena-Aguilar, G., et al. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol. Cell. Biol.* 31, 300–316. doi: 10.1128/MCB.00561-10

Marczylo, E. L., Amoako, A. A., Konje, J. C., Gant, T. W., and Marczylo, T. H. (2012). Smoking induces differential miRNA expression in human spermatozoa: a potential transgenerational epigenetic concern? *Epigenetics* 7, 432–439. doi: 10.4161/epi.19794

Miller, D. (2014). Sperm RNA as a mediator of genomic plasticity. *Adv. Biol.* 2014:179701. doi: 10.1155/2014/179701

Nixon, B., Stanger, S. J., Mihalas, B. P., Reilly, J. N., Anderson, A. L., Tyagi, S., et al. (2015). The microRNA signature of mouse spermatozoa is substantially modified during epididymal maturation. *Biol. Reprod.* 93, 91. doi: 10.1095/biolreprod.115.132209

Ostermeier, G. C., Miller, D., Huntriss, J. D., Diamond, M. P., and Krawetz, S. A. (2004). Reproductive biology: delivering spermatozoan RNA to the oocyte. *Nature* 429:154. doi: 10.1038/429154a

Pittoggi, C., Beraldi, R., Sciamanna, I., Barberi, L., Giordano, R., Magnano, A. R., et al. (2006). Generation of biologically active retro-genes upon interaction of mouse spermatozoa with exogenous DNA. *Mol. Reprod. Dev.* 73, 1239–1246. doi: 10.1002/mrd.20550

Pittoggi, C., Sciamanna, I., Mattei, E., Beraldi, R., Lobascio, A. M., Mai, A., et al. (2003). Role of endogenous reverse transcriptase in murine early embryo development. *Mol. Reprod. Dev.* 66, 225–236. doi: 10.1002/mrd.10349

Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., and Cuzin, F. (2006). RNA mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* 441, 469–474. doi: 10.1038/nature04674

Reilly, J. N., McLaughlin, E. A., Stanger, S. J., Anderson, A. L., Hutcheon, K., Church, K., et al. (2016). Characterisation of mouse epididymosomes reveals a complex pro le of microRNAs and a potential mechanism for modi cation of the sperm epigenome. *Sci. Rep.* 6:31794. doi: 10.1038/srep31794

Rodgers, A. B., Morgan, C. P., Bronson, S. L., Revello, S., and Bale, T. L. (2013). Paternal stress exposure alters sperm microRNA content and reprograms offspring HPA stress axis regulation. *J. Neurosci.* 33, 9003–9012. doi: 10.1523/JNEUROSCI.0914-13.2013

Rodgers, A. B., Morgan, C. P., Leu, N. A., and Bale, T. L. (2015). Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13699–13704. doi: 10.1073/pnas.1508347112

Sciamanna, I., Barberi, L., Martire, A., Pittoggi, C., Beraldi, R., Giordano, R., et al. (2003). Sperm endogenous reverse transcriptase as mediator of new genetic information. *Biochem. Biophys. Res. Commun.* 312, 1039–1046. doi: 10.1016/j.bbrc.2003.11.024

Sciamanna, I., Gualtieri, A., Cossetti, C., Osimo, E. F., Ferracin, M., Macchia, G., et al. (2013). A tumor-promoting mechanism mediated by retrotransposon-encoded reverse transcriptase is active in human transformed cell lines. *Oncotarget* 4, 2271–2287. doi: 10.18632/oncotarget.1403

Sciamanna, I., Gualtieri, A., Piazza, P. V., and Spadafora, C. (2014). Regulatory roles of LINE-1-encoded reverse transcriptase in cancer onset and progression. *Oncotarget* 5, 8039–8051. doi: 10.18632/oncotarget.2504

Sendler, E., Johnson, G. D., Mao, S., Goodrich, R. J., Diamond, M. P., Hauser, R., et al. (2013). Stability, delivery and functions of human sperm RNAs at fertilization. *Nucleic Acids Res.* 41, 4104–4117. doi: 10.1093/nar/gkt132

Sharma, U., Conine, C. C., Shea, J. M., Boskovic, A., Derr, A. G., Bing, X. Y., et al. (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* 351, 391–396. doi: 10.1126/science.aad6780

Sinibaldi-Vallebona, P., Matteucci, C., and Spadafora, C. (2011). Retrotransposon-encoded reverse transcriptase in the genesis, progression and cellular plasticity of human cancer. *Cancers* 3, 1141–1157. doi: 10.3390/cancers3011141

Smith, K., and Spadafora, C. (2005). Sperm-mediated gene transfer: applications and implications. *Bioessays* 27, 551–562. doi: 10.1002/bies.20211

Smith, Z. D., Chan, M. M., Humm, K. C., Karnik, R., Mekhoubad, S., Regev, A., et al. (2014). DNA methylation dynamics of the human preimplantation embryo. *Nature* 511, 611–615. doi: 10.1038/nature13581

Spadafora, C. (1998). Sperm cells and foreign DNA: a controversial relation. *Bioessays* 20, 955–964. doi: 10.1002/(SICI)1521-1878(199811)20:11<955::AID-BIES11>3.0.CO;2-8

Spadafora, C. (2008). Sperm-mediated 'reverse' gene transfer: a role of reverse transcriptase in the generation of new genetic information. *Hum. Reprod.* 23, 735–740. doi: 10.1093/humrep/dem425

Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., et al. (2010). MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr. Biol.* 20, 271–277. doi: 10.1016/j.cub.2009.12.044

Tang, F., Kaneda, M., O'Carroll, D., Hajkova, P., Barton, S. C., Sun, Y. A., et al. (2007). Maternal microRNAs are essential for mouse zygotic development. *Genes Dev.* 21, 644–648. doi: 10.1101/gad.418707

Vidigal, J. A., and Ventura, A. (2015). The biological functions of miRNAs: lessons from *in vivo* studies. *Trends Cell Biol.* 25, 137–147. doi: 10.1016/j.tcb.2014.11.004

Vitullo, P., Sciamanna, I., Baiocchi, M., Sinibaldi-Vallebona, P., and Spadafora, C. (2012). LINE-1 retrotransposon copies are amplified during murine early embryo development. *Mol. Reprod. Dev.* 79, 118–127. doi: 10.1002/mrd.22003

Vojtech, L., Woo, S., Hughes, S., Levy, C., Ballweber, L., Sauteraud, L. P., et al. (2014). Exosomes in human semen carry a distinctive repertoire of small non-coding RNAs with potential regulatory functions. *Nucleic Acids Res.* 42, 7290–7304. doi: 10.1093/nar/gku347

Waddington, C. H. (1959). Canalization of development and genetic assimilation of acquired characters. *Nature* 183, 1654–1655. doi: 10.1038/1831654a0

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., et al. (2009). The deep evolution of metazoan microRNAs. *Evol. Dev.* 11, 50–68. doi: 10.1111/j.1525-142X.2008.00302.x

Zhong, C., Xie, Z., Yin, Q., Dong, R., Yang, S., Wu, Y., et al. (2016). Parthenogenetic haploid embryonic stem cells efficiently support mouse generation by oocyte injection. *Cell Res.* 26, 131–134. doi: 10.1038/cr.2015.132

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

MINI REVIEW
published: 10 October 2017
doi: 10.3389/fmicb.2017.01957

Pivotal Impacts of Retrotransposon Based Invasive RNAs on Evolution

Laleh Habibi and Hamzeh Salmani*

Department of Medical Genetics, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

RNAs have long been described as the mediators of gene expression; they play a vital role in the structure and function of cellular complexes. Although the role of RNAs in the prokaryotes is mainly confined to these basic functions, the effects of these molecules in regulating the gene expression and enzymatic activities have been discovered in eukaryotes. Recently, a high-resolution analysis of the DNA obtained from different organisms has revealed a fundamental impact of the RNAs in shaping the genomes, heterochromatin formation, and gene creation. Deep sequencing of the human genome revealed that about half of our DNA is comprised of repetitive sequences (remnants of transposable element movements) expanded mostly through RNA-mediated processes. ORF2 encoded by L1 retrotransposons is a cellular reverse transcriptase which is mainly responsible for RNA invasion of various transposable elements (L1s, Alus, and SVAs) and cellular mRNAs in to the genomic DNA. In addition to increasing retroelements copy number; genomic expansion in association with centromere, telomere, and heterochromatin formation as well as pseudogene creation are the evolutionary consequences of this RNA-based activity. Threatening DNA integrity by disrupting the genes and forming excessive double strand breaks is another effect of this invasion. Therefore, repressive mechanisms have been evolved to control the activities of these invasive intracellular RNAs. All these mechanisms now have essential roles in the complex cellular functions. Therefore, it can be concluded that without direct action of RNA networks in shaping the genome and in the development of different cellular mechanisms, the evolution of higher eukaryotes would not be possible.

Keywords: evolution, retrotransposon, invasive RNA, pseudogene, DNA structure

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos - Philosophische Praxis, Austria

Reviewed by:
Carmen Hernandez,
Instituto de Biología Molecular y Celular de Plantas (CSIC), Spain
Yukihito Ishizaka,
National Center for Global Health and Medicine, Japan

****Correspondence:***
Laleh Habibi
habibi@razi.tums.ac.ir

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *12 July 2017*
Accepted: *22 September 2017*
Published: *10 October 2017*

Citation:
Habibi L and Salmani H (2017) Pivotal Impacts of Retrotransposon Based Invasive RNAs on Evolution. Front. Microbiol. 8:1957. doi: 10.3389/fmicb.2017.01957

INTRODUCTION

Finding the primary molecule that was responsible for the initiation of life on Earth is the goal of many studies in the field of evolution. Regarding the "central dogma," the DNA has been a candidate for the name of the molecule of life. However, fans of the "RNA world theory" explain how life could have been started by the RNAs. The discovery of RNAs with enzymatic activity (Ellington and Szostak, 1990; Robertson and Joyce, 1990; Tuerk and Gold, 1990) and the chemical features of different RNAs—along with the widespread viruses using RNA as their only genetic material—are some clues that help scientists describe the RNA world hypothesis (Pressman et al., 2015). In this theory, it is postulated that RNA and RNA-like molecules, which could fold into a three-dimensional structure with catalytic activities, had played central metabolic roles in the ancient world (Bass and Cech, 1984). Additionally, the double feature of tRNAs to bind with the genetic codes in one loop and their specific binding to amino acids in another stem could further confirm the central role of this molecule in early evolution (Lee et al., 2000;

Saito et al., 2001; Murakami et al., 2003; Chumachenko et al., 2009). In this review, we have briefly discussed the importance of the intracellular RNAs in the DNA expansion and its role in shaping the genome to create higher order structures and mechanisms throughout the course of evolution.

TYPES OF RNAs AND INTRACELLULAR INVASIVE RNAs

RNAs had been primarily known as the mediators of the gene expression. However, the different types of RNAs with various roles in the eukaryotic and prokaryotic cells have been discovered. Based on their functions, these molecules can be categorized into four different types: (1) Encoding RNAs that contain the codons for the synthesis of polypeptides. (2) Structural RNAs [ribonucoeoproteins (RNPs)] that incorporate into the structure of some proteins; thus, they could have played an essential role in maintaining the steady feature and activity of these proteins (Cech and Steitz, 2014). (3) Catalytic RNAs (ribozymes), associated with proteins (RNPs), and mainly involved in the formation of peptide bonds in the peptidyl transferase center of ribosomes, site specific cleavage, ligation of RNAs, and mRNA splicing (Weinger et al., 2004; Keating et al., 2010; Wilson et al., 2016). (4) Regulatory RNAs (riboregulators), which include the non-coding RNAs with various sequences and sizes. These RNAs could regulate the gene expression by targeting mRNAs, leading to the modification of the rRNA, repressions of transposons, and also involved in X-inactivation, chromatin remodeling, and DNA methylation to repress the transcription (Lippman et al., 2004; Esteller, 2011; Cech and Steitz, 2014).

Apart from these functional molecules, the eukaryotic cells also contain RNAs that are exclusively transcribed to be incorporated into the genome by a mechanism called reverse transcription. This process is mainly involved in the construction of telomere (Autexier and Lue, 2006; Lewis and Wuttke, 2012), formation of pseudogenes (Tutar, 2012; Milligan and Lipovich, 2015), and expansion of retrotransposon (Kassiotis and Stoye, 2016). In all these cases, the intracellular RNAs (which we have called "invasive RNAs" in this paper) could be transformed to cDNA in the nucleus and inserted into the genome through the double strand breaks in the DNA. Generally, three types of invasive RNAs can be considered in the eukaryotic cells. Some of these RNAs have been evolved to form specific genomic constructions, such as the telomerase RNA component (TERC), which functions as a template for the extension of telomeres at the end of the eukaryotic chromosomes (Ozturk et al., 2017). Invasive RNAs transcribed from the retrotransposons do not seem to play any pivotal roles in a cells' lifecycle, but have been highly effective during evolution (Cordaux and Batzer, 2009). The DNA might also be attacked by functional RNAs. These RNAs are not naturally invasive, but could be transformed into cDNA by intracellular reverse transcriptase (RTs) and result in the formation of pseudogenes (Tutar, 2012).

The RTs are the key enzymes for RNA invasion. Telomerase and ORF2 (reverse transcriptase produced by retrotransposon) are the two known functional RTs in the eukaryotic cells (Meyer et al., 2017). The role of telomerase is confined to the construction of telomeres by using a specific RNA (TERC) as a template (Lewis and Wuttke, 2012); however, ORF2 uses cytoplasmic RNAs and retroelement transcripts to create pseudogenes and cause retrotransposon expansion respectively (Wei et al., 2001). Interestingly, in some eukaryotes, the retroelement-related RT is responsible for the elongation of the telomere (Biessmann et al., 1992).

INVASIVE RNAs ORIGINATED FROM RETROTRANSPOSONS: STRUCTURAL AND FUNCTIONAL ROLES

Retrotransposons are groups of mobile DNA elements [transposable elements (TEs)] that copy and paste themselves using the RNA molecules (**Figure 1**). As mentioned in the previous section, these RNA molecules are naturally invasive and are basically transcribed to be randomly inserted into the genome and increase the copy numbers of the retrotransposons (Goodier and Kazazian, 2008). All groups of TEs had been active during the early evolution; however, their selfish and mutagenic movements have resulted in the limitation of their activities to specific types of retrotransposons in the modern human (Marchetto et al., 2013). Long Interspersed Elements (LINE, L1) are the most active retroelement in our cells. It is estimated that the human DNA contains around 500,000 copies of the L1 retrotransposon; however, only 80–100 copies of these elements have maintained their mobility (Goodier and Kazazian, 2008). The structure of a complete L1 element includes a promoter located in the 5′ UTR region, an open reading frame (ORF) 1 gene that encodes the RNA binding protein, ORF2 gene that produces a protein with both endonuclease and reverse transcriptase activity in the two different domains, and a 3′ UTR providing poly-A-tail for the L1 RNA (Goodier and Kazazian, 2008). The RNA polymerase II apparatus is responsible for the production of the L1 RNA (Burns and Boeke, 2012). The transcribed RNA is then transported to the cytoplasm to produce the ORF1 and ORF2 proteins. This invasive RNA in the complex with ORF1 and ORF2 is transported to the nucleus, where it invades the DNA using endonuclease and reverse transcriptase activity of the ORF2 protein by a mechanism called target prime reverse transcription (TPRT) (Cost et al., 2002). The invasive RNAs produced by other retrotransposons (Alu and SVAs) are inserted into the genome through the function of the L1 proteins (Raiz et al., 2012).

It seems that more than the other TEs, the retrotransposons have a greater impact on changing the structure of the DNA and developing specific cellular mechanisms through 100s million years of evolution (Habibi et al., 2015). The retroelement RNA invasions that occurred most often early during evolution have been caused by the genomic expansion and when the DNA is given the space to create structures, such as heterochromatin and centromere (Nigumann et al., 2002) (**Figure 2**). The human genome project revealed that more than half of our DNA is comprised of non-coding regions. Further evaluation showed

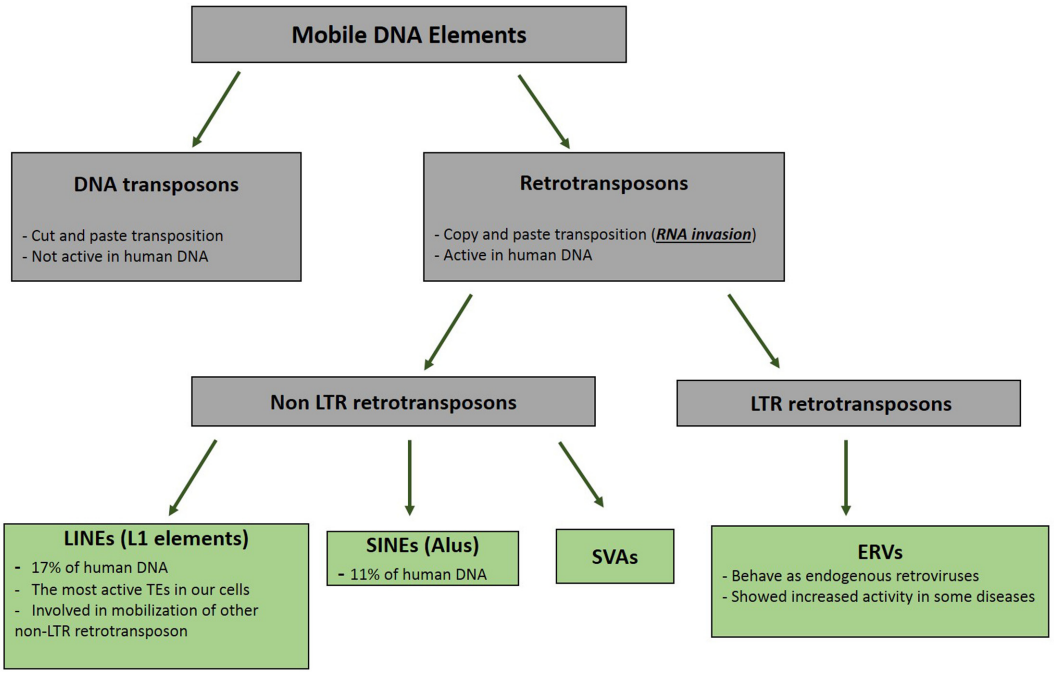


FIGURE 1 | Pedigree of Mobile DNA Elements. Transposable elements are categorized in two distinct groups based on their mode of mobilization. DNA transposons move by cut and paste mechanism, however, retrotransposons (retroelements) mobility are mediated by RNAs. Activity of DNA transposons had been fully shut down during evolution, whereas retroelements still show activities in different types of our cells.

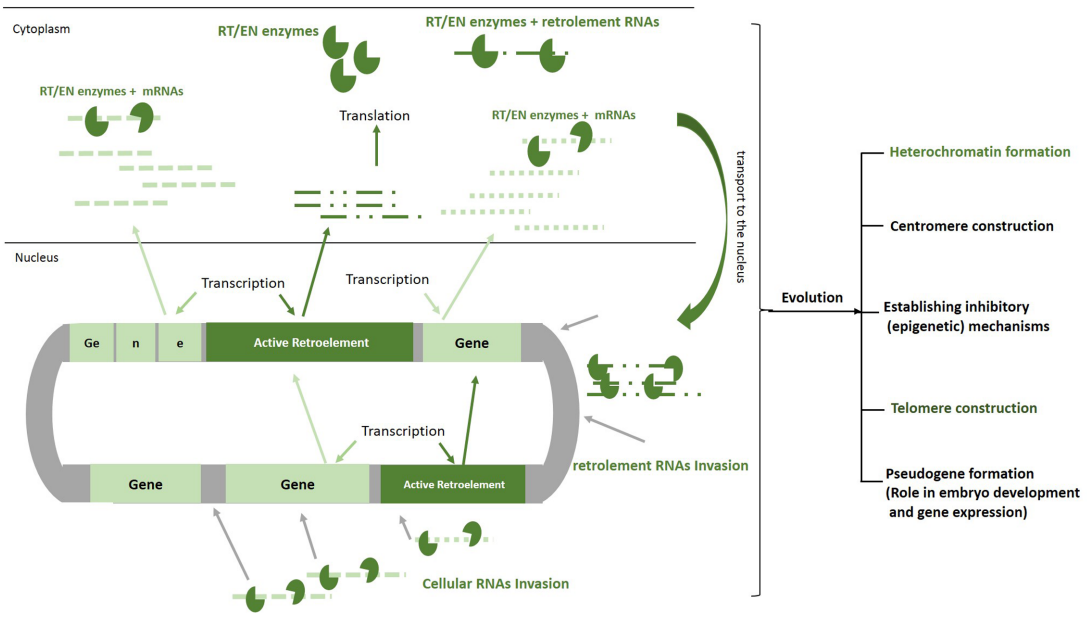


FIGURE 2 | Roles of RNA invasion in shaping human genome. Active retroelements in early evolution have been able to actively transpose and increase their copy number by means of their natural invasive RNAs. Retroelements RNAs similar to mRNAs are transcribed and translated by cellular apparatus. The proteins that are encoded by these elements (EN/RT) can bind to cellular RNAs as well as retroelements RNAs, transport them to the nucleus, create DNA breaks, make cDNA, and finally pasting a copy of each RNA in to the genome. Although the movements of these mobile elements are now inhibited remarkably in our cells, 100 million years of their activities have resulted in formation of heterochromatin, centromere, telomere, and pseudogenes. In order to decrease deleterious effects of retrotransposition, inhibitory mechanisms such as DNA methylation, heterochromatinization, and miRNA production have been established by host cell. EN/RT, endonuclease/reverse transcriptase.

that these parts of our genome, which mainly construct the heterochromatin, centromeres, telomeres, and gene spacers, include repetitive sequences comprising the remnants of the retrotransposition events (Lander et al., 2001). The importance of the heterochromatin and centromeres in the gene expression, senescence, embryo development, and cell cycle in the eukaryotes has been found in different studies (Eberl et al., 1993; Harmon and Sedat, 2005; Hammoud et al., 2009; Chandra et al., 2015). Therefore, one can conclude that without the actions of these ancient invasive RNAs, our cells would not perform genomic expansion to form the heterochromatin region, centromeres, telomeres, introns, or regulatory elements, and would remain in the prokaryotic phase. On the other hand, the RT enzyme produced by the retroelements could transform the functional cytoplasmic RNAs into invasive molecules to create pseudogenes (**Figure 2**). This process was essential in the doubling of genes and generation of new genes with different functions throughout the course of evolution (Tutar, 2012). Additionally, it has been shown that the small interfering RNAs transcribed from these pseudogenes might interact with the functional genes in the eukaryotic cells (Ewing, 2017).

Although retrotransposons had been important in the shaping and evolution of the eukaryotic genome, the selfish mobility of these elements would be harmful for DNA integrity and cell viability (Symer et al., 2002). During the retrotransposons' lifecycle, the invasion of the RNAs by means of the endonuclease/RT enzyme could break the genes, disrupt the open-reading frames, and, finally, affect the production of proteins (Dupuy et al., 2001). The excessive activity of endonuclease produced by the retroelements could also induce excessive DNA double strand breaks (Chen et al., 2006). On the other hand, the promoter region of these elements might also be copied and inserted near the genes and thus influence the quantity and quality of the gene expression (Cordaux and Batzer, 2009). Different lines of studies have shown the increased levels of L1 retrotransposition in different types of cancers (Shpyleva et al., 2017), schizophrenia (Bundo et al., 2014; Doyle et al., 2017), autism (Shpyleva et al., 2017), and Rett syndrome (Muotri et al., 2010), emphasizing the pathogenic role of these elements in the human cells. Regarding these potential threats, the eukaryotic cells have developed repressive mechanisms, including epigenetic modifications (DNA methylation, heterochromatinization), miRNAs, and piRNAs expressions, to inhibit and control the activity of the TEs (mainly retrotransposons) (Habibi et al., 2015). All these repressive pathways have other roles at present rather than the retrotransposons repression inside the cells. Therefore, we can emphasize that the embryo development, differential gene expression, cell differentiation, and specifications would not have occurred without the development of repressive mechanisms against intracellular invasive RNAs.

CONCLUSION

Various types of RNAs have been discovered that play a role in the different aspects of the gene expression. Here, we have described another kind of RNAs that are transcribed to invade the DNA and increase their source (retroelement) copy number. These ancient RNAs have a pivotal role in increasing the size of the DNA, establishing heterochromatin, centromeres, telomeres, methylation processes, epigenetic mechanisms, miRNA production, etc. through 100 million years of evolution.

Regardless of the advantageous evolutionary roles; the activities of retrotransposons and their invasive RNAs are highly inhibited in fully differentiated cells (Wissing et al., 2012) since such invasions could be harmful for the genomic integrity of evolved cells. However, different lines of studies showed increased retroelements movements in neural precursor cells (Muotri et al., 2005), embryonic stem cells (Kano et al., 2009) as well as germ cells (Georgiou et al., 2009). One could suggest two ideas for this exceptional high activity of the retrotransposons; (1) in all these cells we are facing to vast changes in epigenetic status of DNA including hypomethylation which could remove the lock of the retroelements and give them chance to increase their movements as side effect of epigenetic changes. These random RNA insertions might result in neurodevelopmental disorders (McConnell et al., 2017) and different kind of cancers (Kano et al., 2009). (2) These increased retrotransposition might do have functional role such as memory storage in neurons (Habibi et al., 2009) and involving in the survival of the organism (Sciamanna et al., 2011). Totally, all these aspects of intracellular invasive RNAs life cycle could show the importance of these elements in creating complex organisms during the evolution.

AUTHOR CONTRIBUTIONS

LH: collected data, wrote the paper, designed and drew figures; HS: collected data.

FUNDING

This article is funded by LH.

REFERENCES

Autexier, C., and Lue, N. F. (2006). The structure and function of telomerase reverse transcriptase. *Annu. Rev. Biochem.* 75, 493–517. doi: 10.1146/annurev.biochem.75.103004.142412

Bass, B. L., and Cech, T. R. (1984). Specific interaction between the self-splicing RNA of Tetrahymena and its guanosine substrate: implications for biological catalysis by RNA. *Nature* 308, 820–826. doi: 10.1038/308820a0

Biessmann, H., Champion, L. E., O'Hair, M., Ikenaga, K., Kasravi, B., and Mason, J. M. (1992). Frequent transpositions of *Drosophila melanogaster* HeT-A transposable elements to receding chromosome ends. *EMBO J.* 11, 4459–4469.

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., et al. (2014). Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* 81, 306–313. doi: 10.1016/j.neuron.2013.10.053

Burns, K. H., and Boeke, J. D. (2012). Human transposon tectonics. *Cell* 149, 740–752. doi: 10.1016/j.cell.2012.04.019

Cech, T. R., and Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 157, 77–94. doi: 10.1016/j.cell.2014.03.008

Chandra, T., Ewels, P. A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S. W., Kirschner, K., et al. (2015). Global reorganization of the nuclear landscape in senescent cells. *Cell Rep.* 10, 471–483. doi: 10.1016/j.celrep.2014.12.055

Chen, J. M., Férec, C., and Cooper, D. N. (2006). LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. *J. Biomed. Biotechnol.* 2006:56182. doi: 10.1155/JBB/2006/56182

Chumachenko, N. V., Novikov, Y., and Yarus, M. (2009). Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J. Am. Chem. Soc.* 131, 5257–5263. doi: 10.1021/ja809419f

Cordaux, R., and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703. doi: 10.1038/nrg2640

Cost, G. J., Feng, Q., Jacquier, A., and Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *EMBO J.* 21, 5899–5910. doi: 10.1093/emboj/cdf592

Doyle, G. A., Crist, R. C., Karatas, E. T., Hammond, M. J., Ewing, A. D., Ferraro, T. N., et al. (2017). Analysis of LINE-1 elements in DNA from postmortem brains of individuals with schizophrenia. *Neuropsychopharmacology* doi: 10.1038/npp.2017.115 [Epub ahead of print].

Dupuy, A. J., Fritz, S., and Largaespada, D. A. (2001). Transposition and gene disruption in the male germline of the mouse. *Genesis* 30, 82–88. doi: 10.1002/gene.1037

Eberl, D. F., Duyf, B. J., and Hilliker, A. J. (1993). The role of heterochromatin in the expression of a heterochromatic gene, the rolled locus of *Drosophila melanogaster*. *Genetics* 134, 277–292.

Ellington, A. D., and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature* 346, 818–822. doi: 10.1038/346818a0

Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874. doi: 10.1038/nrg3074

Ewing, A. D. (2017). "The mobilisation of processed transcripts in germline and somatic tissues," in *Human Retrotransposons in Health and Disease*, ed. G. Cristofari (Cham: Springer), 95–106.

Georgiou, I., Noutsopoulos, D., Dimitriadou, E., Markopoulos, G., Apergi, A., Lazaros, L., et al. (2009). Retrotransposon RNA expression and evidence for retrotransposition events in human oocytes. *Hum. Mol. Genet.* 18, 1221–1228. doi: 10.1093/hmg/ddp022

Goodier, J. L., and Kazazian, H. H. Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* 135, 23–35. doi: 10.1016/j.cell.2008.09.022

Habibi, L., Ebtekar, M., and Jameie, S. B. (2009). Immune and nervous systems share molecular and functional similarities: memory storage mechanism. *Scand. J. Immunol.* 69, 291–301. doi: 10.1111/j.1365-3083.2008.02215.x

Habibi, L., Pedram, M., AmirPhirozy, A., and Bonyadi, K. (2015). Mobile DNA elements: the seeds of organic complexity on earth. *DNA Cell Biol.* 34, 597–609. doi: 10.1089/dna.2015.2938

Hammoud, S. S., Nix, D. A., Zhang, H., Purwar, J., Carrell, D. T., and Cairns, B. R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. *Nature* 460, 473–478. doi: 10.1038/nature08162

Harmon, B., and Sedat, J. (2005). Cell-by-cell dissection of gene expression and chromosomal interactions reveals consequences of nuclear reorganization. *PLOS Biol.* 3:e67. doi: 10.1371/journal.pbio.0030067

Kano, H., Godoy, I., Courtney, C., Vetter, M. R., Gerton, G. L., Ostertag, E. M., et al. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev.* 23, 1303–1312. doi: 10.1101/gad.1803909

Kassiotis, G., and Stoye, J. P. (2016). Immune responses to endogenous retroelements: taking the bad with the good. *Nat. Rev. Immunol.* 16, 207–219. doi: 10.1038/nri.2016.27

Keating, K. S., Toor, N., Perlman, P. S., and Pyle, A. M. (2010). A structural analysis of the group II intron active site and implications for the spliceosome. *RNA* 16, 1–9. doi: 10.1261/rna.1791310

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Lee, N., Bessho, Y., Wei, K., Szostak, J. W., and Suga, H. (2000). Ribozyme-catalyzed tRNA aminoacylation. *Nat. Struct. Biol.* 7, 28–33. doi: 10.1038/71225

Lewis, K. A., and Wuttke, D. S. (2012). Telomerase and telomere-associated proteins: structural insights into mechanism and evolution. *Structure* 20, 28–39. doi: 10.1016/j.str.2011.10.017

Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476. doi: 10.1038/nature02651

Marchetto, M. C., Narvaiza, I., Denli, A. M., Benner, C., Lazzarini, T. A., Nathanson, J. L., et al. (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* 503, 525–529. doi: 10.1038/nature12686

McConnell, M. J., Moran, J. V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., et al. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: the brain somatic mosaicism network. *Science* 356:eaal1641. doi: 10.1126/science.aal1641

Meyer, T. J., Rosenkrantz, J. L., Carbone, L., and Chavez, S. L. (2017). Endogenous retroviruses: with us and against us. *Front. Chem.* 5:23. doi: 10.3389/fchem.2017.00023

Milligan, M. J., and Lipovich, L. (2015). Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front. Genet.* 5:476. doi: 10.3389/fgene.2014.00476

Muotri, A. R., Chu, V. T., Marchetto, M. C., Deng, W., Moran, J. V., and Gage, F. H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 435, 903–910. doi: 10.1038/nature03663

Muotri, A. R., Marchetto, M. C., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., et al. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 468, 443–446. doi: 10.1038/nature09544

Murakami, H., Kourouklis, D., and Suga, H. (2003). Using a solid-phase ribozyme aminoacylation system to reprogram the genetic code. *Chem. Biol.* 10, 1077–1084. doi: 10.1016/j.chembiol.2003.10.010

Nigumann, P., Redik, K., Mätlik, K., and Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* 79, 628–634. doi: 10.1006/geno.2002.6758

Ozturk, M. B., Li, Y., and Tergaonkar, V. (2017). Current insights to regulation and role of telomerase in human diseases. *Antioxidants* 6:E17. doi: 10.3390/antiox6010017

Pressman, A., Blanco, C., and Chen, I. A. (2015). The RNA world as a model system to study the origin of life. *Curr. Biol.* 25, R953–R963. doi: 10.1016/j.cub.2015.06.016

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., et al. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* 40, 1666–1683. doi: 10.1093/nar/gkr863

Robertson, D. L., and Joyce, G. F. (1990). Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* 344, 467–468. doi: 10.1038/344467a0

Saito, H., Kourouklis, D., and Suga, H. (2001). An in vitro evolved precursor tRNA with aminoacylation activity. *EMBO J.* 20, 1797–1806. doi: 10.1093/emboj/20.7.1797

Sciamanna, I., Vitullo, P., Curatolo, A., and Spadafora, C. (2011). A reverse transcriptase-dependent mechanism is essential for murine preimplantation development. *Genes* 2, 360–373. doi: 10.3390/genes2020360

Shpyleva, S., Melnyk, S., Pavliv, O., Pogribny, I., and Jill James, S. (2017). Overexpression of LINE-1 retrotransposons in autism brain. *Mol. Neurobiol.* doi: 10.1007/s12035-017-0421-x [Epub ahead of print].

Symer, D. E., Connelly, C., Szak, S. T., Caputo, E. M., Cost, G. J., Parmigiani, G., et al. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* 110, 327–338. doi: 10.1016/S0092-8674(02)00839-5

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249, 505–510. doi: 10.1126/science.2200121

Tutar, Y. (2012). Pseudogenes. *Comp. Funct. Genomics* 2012:424526. doi: 10.1155/2012/424526

Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., et al. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* 21, 1429–1439. doi: 10.1128/MCB.21.4.1429-1439.2001

Weinger, J. S., Parnell, K. M., Dorner, S., Green, R., and Strobel, S. A. (2004). Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nat. Struct. Mol. Biol.* 11, 1101–1106. doi: 10.1038/nsmb841

Wilson, T. J., Liu, Y., and Lilley, D. M. J. (2016). Ribozymes and the mechanisms that underlie RNA catalysis. *Front. Chem. Sci. Eng.* 10, 178–185. doi: 10.1007/s11705-016-1558-2

Wissing, S., Muñoz-Lopez, M., Macia, A., Yang, Z., Montano, M., Collins, W., et al. (2012). Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. *Hum. Mol. Genet.* 21, 208–218. doi: 10.1093/hmg/ddr455

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 22 December 2017
doi: 10.3389/fmicb.2017.02483

The Importance of ncRNAs as Epigenetic Mechanisms in Phenotypic Variation and Organic Evolution

Daniel Frías-Lasserre and Cristian A. Villagra*

Instituto de Entomología, Universidad Metropolitana de Ciencias de la Educación, Santiago, Chile

Neo-Darwinian explanations of organic evolution have settled on mutation as the principal factor in producing evolutionary novelty. Mechanistic characterizations have been also biased by the classic dogma of molecular biology, where only proteins regulate gene expression. This together with the rearrangement of genetic information, in terms of genes and chromosomes, was considered the cornerstone of evolution at the level of natural populations. This predominant view excluded both alternative explanations and phenomenologies that did not fit its paradigm. With the discovery of non-coding RNAs (ncRNAs) and their role in the control of genetic expression, new mechanisms arose providing heuristic power to complementary explanations to evolutionary processes overwhelmed by mainstream genocentric views. Viruses, epimutation, paramutation, splicing, and RNA editing have been revealed as paramount functions in genetic variations, phenotypic plasticity, and diversity. This article discusses how current epigenetic advances on ncRNAs have changed the vision of the mechanisms that generate variation, how organism-environment interaction can no longer be underestimated as a driver of organic evolution, and how it is now part of the transgenerational inheritance and evolution of species.

Keywords: non-codingRNAs, phenotypic plasticity, biodiversity, adaptation, evolution

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria

Reviewed by:
Diego Franco,
Universidad de Jaén, Spain
Claes Wahlestedt,
Leonard M. Miller School of Medicine, United States

****Correspondence:***
Daniel Frías-Lasserre
daniel.frias@umce.cl

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *28 July 2017*
Accepted: *29 November 2017*
Published: *22 December 2017*

Citation:
Frías-Lasserre D and Villagra CA (2017) The Importance of ncRNAs as Epigenetic Mechanisms in Phenotypic Variation and Organic Evolution. Front. Microbiol. 8:2483. doi: 10.3389/fmicb.2017.02483

INTRODUCTION

In the Synthetic Theory of Evolution, mutations have been proposed as the principal factor behind the origin of new phenotypic variation and highlighted as the cornerstone of evolutionary process (Nei, 2013). In that approach, phenotype variations related to the environment, such as the reaction norm and phenotypic plasticity, did not influence the genetic background, and were therefore not transmitted to offspring (Mayr, 1985). The central dogma postulates an unidirectional flow of information from DNA, mediated by RNA, to proteins (Crick, 1958, 1970). This pervasive idea consolidated a deterministic and reductionist inheritance (Shapiro, 2009; Frías-Lasserre, 2012), impacting our understanding of all genetic mechanisms that effectively intervene on population genetics and organic evolution (Schreiber, 2005; Weber, 2006; Gillings and Westoby, 2014). As a result, many evolutionary mechanisms have been omitted in Neo-Darwinian theory, including non-coding RNAs (ncRNAs; Frías, 2010). In classic evolutionary theory, genetic code was mainly associated with protein coding DNAs, which only make up ∼2% of the human genome; however,

recently, novel functions have been assigned for non-coding DNA regions for proteins (Lunter et al., 2006; Dunham et al., 2012). The remaining non-coding area of DNA has been revealed to be related to key biological processes and adaptive complexity in eukaryotic life, both in plants and animals, contradicting the paradox of the value C (Creevey and McInerney, 2003; Andolfatto, 2005; Taft et al., 2007; Knowles and McLysaght, 2009; Ling and Wurtele, 2014; Gaiti et al., 2016). In the habitat where the organisms live, there are a variety of stimuli and stressors that could induce rapid modification in the transcription of genes through epigenetic mechanisms capable of generating memory and epigenomic transgenerational inheritance (D'Urso and Brickner, 2014). Gene expression can be differentially influenced by stable epigenetic modifications that can be later kept through ontogenetic development with the aid of ncRNAs and even pass to the following generations (Jaenisch and Bird, 2003; Hanson and Skinner, 2016; Van Otterdijk and Michels, 2016).

The objective of this article is to analyze the importance of ncRNAs in the regulation of gene expression, and their impact at the level of population variability, adaptation, and the evolution of species. Also, we review and discuss how environmental stimuli and ncRNAs may play an important role in inheritance through the epigenome by triggering epigenetically heritable changes that may lead the origin of new species. In transgenerational inheritance caused by environmental stressors, ncRNAs may play an important role among the set of mechanisms that underlie changes in phenotypic variation and organic evolution.

NEW MECHANISMS OF GENETIC VARIABILITY AND PHENOTYPIC NOVELTY

The stability of genes on homologous chromosomes, except for translocation, was a generalized fact for geneticists until 1960, when *mobile genetic elements* were described by Barbara McClintock (1950) in corn. This findings was later verified in other eukaryotes and prokaryotes (Sakaguchi, 1990; Kazazian, 2004). In addition to transposable elements, there are other epigenetic mechanisms explaining allelic instability and phenotypic variation such as: splicing, RNA editing, metastable epialleles, epimutation, and paramutations (Tollefsbol, 2014). Many of these mechanisms involve different ncRNAs capable of making gene regulation in cells of various tissues oriented to a wide range of biological processes (Yan, 2014).

THE EPIGENETIC CONCEPT

Waddington (2012) coined *epigenetics* as *the interaction between genes and their products that allow for phenotypic expression* in order to reveal the mechanisms of development under the classic theory of epigenesis. Waddington also coined the concept of *epigenotype* as these *interphase that connected the genotype with the phenotype during development* (Slack, 2002; Sweatt and Tamminga, 2016). Epigenetics is a heritable change in the epigenotype of cells unchanged in the primary structure of DNA (Tollefsbol, 2011). Epigenetic was, for many years, limited to the understanding of cell differentiation. Now it is known that epigenetics is a hereditary transgenerational mechanism, linked to processes such as paramutations, metastable epialleles, DNA methylation, and chromatin remodeling, wherein there is also participation from different types of ncRNAs (Brink et al., 1968; Grewal and Klar, 1996; Cavalli and Paro, 1999; Kosten and Nielsen, 2014; Mashoodh and Champagne, 2014).

NcRNAs

A major surprise arising from the DNA sequencing of eukaryotes organisms was the limited number of protein-coding genes found in relation to the total size of the genome. This had no correlation with the complexity of organisms, and did not explain the effects of selection pressure during evolution (Lander et al., 2001). In areas of the genome that do not encode for proteins, there is a great deal of information for ncRNAs, which also play a key role in regulating gene expression, working on specific sequences targeting genes, transposons, and viruses where they exert regulation or silencing (Mattick, 2009; Qu and Adelson, 2012). The first small RNAs were those rRNAs and tRNAs that were related to protein synthesis (Choudhuri, 2010). Currently, we know that there are many other classes of ncRNAs, small and long (Eddy, 2001), and know about their biogenesis, function and role in diseases (Choudhuri, 2010; Yu et al., 2014; Li et al., 2015). The most ancient of these small ncRNA is thought to be the ribozyme, which is a catalytic RNA. A ribozyme performs its catalyzing process without the aid of protein factors (Swati, 2017). The hammerhead ribozyme, 50–150 nt, was discovered in subviral plant pathogens and has been found in bacteria, archaea, and in many eukaryotic genomes, such as plants and mammals including the human genome. Some ribozymes, the riboswitches, have the ability to catalyze reactions in the absence of proteins and the capacity to function as switches that regulate gene expression by altering their conformation in response to a ligand or a small molecule. Some riboswitches act as thermosensors, detecting and alerting the organism of a temperature rise due to an infection or climatic change (Przybilski et al., 2005; Serganov and Dinshaw, 2007; Martick et al., 2008; De la Peña and Garcia-Robles, 2010a,b; Seehafer et al., 2011). The first ncRNA (miRNAs) was described in *Caenorhabditis elegans* and associated with embryonic development (Lee et al., 1993). In eukaryotes, they are relatively more abundant than protein-coding RNA (Herbert and Rich, 1999). For instance in humans, ∼98% of all transcriptional output corresponds to ncRNAs, and was previously considered *junk DNA* (Wright and Bruford, 2011). NcRNAs have been detected in viruses, archaea, bacteria, and eucharia and can participate in a great number of cellular activities such as transcription, DNA replication, messenger RNAs stability, RNAs processing (Storz, 2002). There are different types of ncRNAs with varied functions; among the most relevant are:

Micro RNA (miRNA)

Are short (22 bp), and found in animals, plants, and viruses. MiRNAs belongs to a highly conserved post-transcriptional regulatory gene family, with paramount functions across various

cellular and developmental processes such as immunity, cell behavior (including proliferation, differentiation, contractility, inflammation), and host–microorganism interactions (Asgari, 2011; Mendell and Olson, 2012). In insects, miRNAs encoded by viruses interact with the host's defenses and help during virus replication (Asgari, 2013). In eutherian mammals, including humans, miRNAs from trophoblasts are expressed in the placentae of pregnant females and could mediate cross talk between the feto-placental unit and the mother during pregnancy (Ouyang et al., 2013). MiRNAs regulate several cellular processes in relation to pregnancy, such as: placental development, endometrial receptivity, angiogenes, and immune cells at the maternal-fetal interface. MiRNAs, are capable of regulating the immunological balance between the mother and her offspring, and likely help to regulate successful placentation and pregnancy. Also miRNAs, via exosomes, induce viral resistence through autophagy and has a role in the maternal-fetal exchange (Bidarimath et al., 2014). Furthermore, during pregnancy, miRNAs interact with reproductive hormones and are important regulators of mRNA translation (Bidarimath et al., 2014). The miRNAs resolve the paradoxical nature of mammalian pregnancy, in which an intimate immunological relationship exists between the mother and the allogeneic fetus where the mother does not reject the fetus. MiRNAs are packaged in vesicles within cells (nano-packages) and are released to the extracellular space, and circulate in blood and breast milk. These miRNAs carry out on target mRNAs in other distant or nearby cells, providing intercellular communication (Ouyang et al., 2013) and also induce antiviral immunity (Mouillet et al., 2014). In plants, miRNA may also play a critical role in seed development and germination (Pluskota et al., 2011). Moreover, miRNAs can act on animal behavior. It has been found that, in eukariotic organism (amphibian larvae), miRNAs participate in neuroplasticity (attraction/aversion) in relation to social preference to sustained exposure to kinship odorants. Thus, miRNAs act as a switch governing experience-dependent social preference (Dulcis et al., 2017).

MiRNA are capable of silencing RNA in a similar way to siRNA, but differing in terms of origins, as miRNA originate from self-folding regions of RNA transcripts forming short hairpins (Lim et al., 2003; Cuperus et al., 2011). Their action mode consists of an interaction with target mRNAs in a perfect complementary base sequence that results in mRNA cleaveage; furthermore, an interaction in an imperfect base sequence causes a translational repression (Yekta et al., 2004).

Small Interfering RNA (siRNA) or (RNAi)

Measure 20–25 bp, and originate from regions of double-stranded RNA molecules (dsRNA). These molecules are capable of interfering with mRNA translation by degrading it after transcription through perfect base pairing. *In vivo* and *in vitro* experiments suggest that the first RNAi initiating step involves the binding of the RNA nucleases to a large dsRNA and its cleavage into discrete 21- to 25-nucleotide RNA fragments (siRNA). In a second step, these siRNAs join a multinuclease complex (RISC) and degrade the homologous single-stranded mRNAs (Agrawal et al., 2003). SiRNA allows for the silencing of genes from different eukaryotic organisms with great specificity (Sunkar et al., 2007; Ghildiyal and Zamore, 2009). The specificity of siRNA's post-transcriptional gene silencing has been used in the development of therapeutic applications for treating a great variety of diseases (Zhou et al., 2008; Pulukuri et al., 2009).

Small Nuclear RNA (snRNA)

Are molecules around 150 bp in length. They are located in the nucleus of eukaryotic cells where they can be found mainly in the soluble fraction of the nucleoplasm, but also associated with the chromatin (Mondal et al., 2010). SnRNA control pre-messenger RNA and regulate the nuclear level of active positive transcription elongation factor b (P-TEFb), thus regulating RNA polymerase II (RNAPII) transcription in the nucleus (Muniz et al., 2013). Among snRNA, there are small nucleolar RNAs (snoRNAs) found in eukaryotic nucleolus and the Cajal bodies. These have several roles in ribosome synthesis, the regulation of alternative splicing, translation and oxidative stress. Moreover, both snRNA and snoRNAs are related to hereditary disorders and carcinogenesis (Mannoor et al., 2012).

Piwi-Interacting RNA (piRNA)

PiRNA are the most abundant and diverse ncRNA molecules found in animals, and have 26–31 bp lacking sequence conservation. They interact with encoding regulatory proteins piwi, configuring RNA-protein complexes associated with post-transcriptional gene silencing and epigenetic reprograming. In the germ line of several animal lineages, piRNA form the piRNA-induced silencing complex (piRISC), a configuration capable of silencing foreign transposable elements protecting genomic heredity integrity (Siomi et al., 2011). Moreover, piRNA play a critical role in genome rearrangement and transgenerational carriers of epigenetic information for genome programming (Ashe et al., 2012), affecting varied biological processes such as stem-cell functioning, tissue regeneration and pathogenic states such as cancer (Kim, 2006).

Long ncRNA (lncRNA)

Are functionally diverse relatively long (more than 200 bp) regulatory ncRNA molecules (Kurokawa, 2015). Despite being the least-studied ncRNAs, so far, it has been demonstrated that lncRNA are capable of regulating themselves and that they function as transcriptional activators and post-transcriptional regulators in gene expression (Ponting et al., 2009). LncRNA controls protein regulator activity and separate them from their target DNA sequences. LncRNA operates as a scaffold platform for subcellular structures, regulating other ncRNAs. However, several lncRNA manufacture themselves in to small RNAs (Wilusz et al., 2009). For instance, some lncRNA are involved in the regulation of somatic tissue differentiation by associating directly with the protein and mRNA related to these processes (Kretz et al., 2012). Xist is an lncRNA that has an important role in the inactivation of one of the X chromosome in female mammals. X-inactivation is a process that equalizes gene expression between mammalian males and females.

THE ROLE OF ncRNAs IN PHENOTYPIC VARIATION

As a consequence of genome organization, the proteome of higher organisms is relatively conserved. For example, comparing humans and mice in terms of genetic coding for proteins, their structure is 99% similar (Mattick, 2001). Therefore, the principal mechanisms of phenotypic variation between species are located in the non-protein coding area of the genome. This suggests that ncRNAs have an important role contributing toward an explanation for the biological diversity in the evolution of species. Small RNAs receive or transmit information from and to the environment, which is stored in the epigenome (Mattick, 2001). The sequence of the small ncRNAs shows evolutionary conservation that in lncRNAs is smaller with certain exceptions (Louro et al., 2008; Guttman et al., 2009; Mercer et al., 2009).

Next we will refer to several mechanisms where the ncRNAs intervene, regulating genetic expression and generating new phenotypic variation such as: (1) DNA Methylation, Chromatin remodeling, and gene expression, (2) Epiallelic interaction, (3) RNA editing, (4) Splicing, (5) Genome imprinting, (6) Hox genes, homeotic mutations, and development, (7) Transgenerational epigenetic.

NcRNAs AND THEIR ROLE IN DNA METHYLATION, CHROMATIN REMODELING AND GENE EXPRESSION

In eukaryotes, epigenetic mechanisms consist of DNA methylation or chromatin modification such as methylation or acetylation (Weigel and Colot, 2012). SiRNAs and lnc RNAs participate regulating gene expression by heterochromatinization (Richards and Elgin, 2002; Rangwala and Richards, 2004; Vella and Slack, 2005; Kim, 2006; Bird, 2007; Koerner et al., 2009; Luco and Misteli, 2011; Luco et al., 2011; Siomi et al., 2011; Chisholm et al., 2012). SiRNAs regulates DNA methylation in CpG dinucleotide in eukaryotes (Kawasaki and Taira, 2004; Klose and Bird, 2006; Suzuki and Bird, 2008; Lyko et al., 2010; Siegfried and Simon, 2010). Additionally, methylation gives extra regulation to those regions of DNA coding for proteins (Flanagan and Wild, 2007; Guttman et al., 2009; Rinn and Chang, 2012; Kulis et al., 2013; Sabin et al., 2013). Also siRNAs induce DNA histone H3 methylation in human cells (Kawasaki and Taira, 2004). Differential methylations during development are important in cell differentiation during the mitosis (Bird, 2002). LncRNAs intervene in methylation or in demethylation through interaction with various methyl transferase in cis or trans, directly or indirectly through a protein intermediate (Cao, 2014; Zhao et al., 2016).

In eukaryotes, miRNAs, piRNAs, and siRNAs also have a function in gene expression at the level of chromatin through histone methylation, acetylation, ubiquination, sumoylation, and phosphorylation. These epigenetic mechanisms regulate gene action in different parts of the chromosome and have an important role in heterochromatinization, replication, and transcription (Black, 2003; Bannister and Kouzarides, 2011; De Lucia and Dean, 2011; Keller and Bühler, 2013; Joh et al., 2014; Rivera et al., 2014).

NcRNAs IN EPIALLELIC INTERACTION AND IMPRINTING

In Mendelism, alleles remain unchanged and are thus transmitted to offspring. With epigenetics, it has been established that alleles can undergo modifications due to methylations, where ncRNAs can participate (Yan, 2014). Methylation of one of the alleles can change the expression of other alleles and produce an epimutation in a locus and originate an *epiallele* that is a group of otherwise identical genes that differ in the grade of methylation and originate novel phenotype that are heritable across generations (Rakyan et al., 2002; Yan, 2014). In *Arabidopsis thaliana* several epialleles related with siRNAs have been identified that correspond to different *Arabidopsis* ecotypes. These varieties present different gene expression characteristics, which are stably-maintained and transmitted to the offspring (Watson et al., 2014). The use of DNA methylation inhibitors can induce phenotypic variation in epialleles during meiosis, which can then be inherited and produce evolutionary change in the offspring (Weigel and Colot, 2012; House and Lukens, 2014; Ruden et al., 2015).

SiRNAS also explains an unusual allelic interaction where an allele in trans position modifies the expression of that allele, without altering their intimate nucleotide structure. These epigenetic interactions in a locus gave origin to the concepts of *paramutations* (Mahfouz, 2010, reviewed by Hollick, 2010). Furthermore, paramutation also extended the concept of *imprinting* and transgenerational heredity to a allelic interaction (Li et al., 1993). In imprinting, the epiallele has a different expression depending on whether it comes from the father or mother. A paradigm of this situation is what happens in the plant *A. thaliana*, where the MEA gene is only expressed in the phenotype of the endosperm, the maternal epiallele (Mahfouz, 2010). Moreover, paramutation has been described in maize by Brink in 1956 in a b1 locus that encodes for the pigment anthocyanin: the B′ allele of low expressivity that can cause changes in the allele B1 of high expressiveness. This change may be inherited for several generations. Both B′ and B1 have the same nucleotide sequence but differ in their methylation pattern (Coe, 1966; Brink et al., 1968; Hollick, 2010). Recently it has been discovered that siRNAs from a tandem repeat of non-coding DNA located in the b1 gene are involved in the paramutation in maize (Chandler, 2007).

In mice, the induced paramutation *white-tail-tip* has been reported using an insertional mutation in the Kit locus (Yuan et al., 2015). Microinjection into fertilized eggs of Kit-specific miRNAs induced a heritable white tail phenotype; however the specific mechanism of these miRNAs on chromatin remodeling is still unknown (Rassoulzadegan et al., 2006; Hollick, 2010). Maternal miRNAs and piRNAs seem to have an inhibitory effect on the germ line transmission of paramutations, meaning they

are an important tool for understanding the mechanism of epigenetic transgenerational inheritance (Yuan et al., 2015).

RNA EDITING AND THE ncRNAs, AND THEIR IMPACT IN THE REGULATION OF TRANSCRIPTION

RNA editing is a special type of mutation in the primary nucleotide sequences in RNAs of eukaryotes, in the nucleus or in mitochondria where functionally different proteins are processed from a single gene. RNA editing was discovered in mitochondria of the protozoa *Trypanosome* where a special type of deletion or insertion of Uridine occurs (Benne et al., 1986; Feagin et al., 1988; Rubio et al., 2007). RNA editing not only occurs in the RNAs that participate in the protein synthesis, but also in ncRNAs such as miRNAs, siRNAs, and piRNAs (Gott and Emeson, 2000; Blanc and Davidson, 2003; Luciano et al., 2004; Liang and Landweber, 2007). Other similar edition of RNAs have been described, such as cytosine deamination and inosine by adenin substitution (Gommans et al., 2009). In higher eukaryotes, A to I RNA generates RNA and protein diversity, selectively reshaping coding and noncoding sequences in nuclear and mitochondria transcripts. The enzymes involved in this type of editing are adenosine deaminases (ADARs). The ADARs edit the duplex RNAs formed by ncRNAs, and can alter RNA functions, leading to an modified regulatory gene network of mRNAs and miRNAs and also siRNAs, piRNAs, and lncRNAs (Singh, 2013), A to I RNA editing may provide key links between neural development, nervous system function and neurological diseases. The ncRNAs and their alternative expression may alter the regulation of genetic machinery and to cause neurological diseases (Penn et al., 2013; Singh, 2013). The list of ncRNAs and their relation with RNA editing in brain development and disease in mammals is growing (Mehler and Mattick, 2007; Salta and De Strooper, 2012). Therefore, RNA editing could be one of the, previously underappreciated, driving forces for adaptive evolution (Gommans et al., 2009).

NcRNAs AND SPLICING

In 1977, Sharp and Roberts discovered RNA splicing, wherein genes are divided into exons and introns (Sharp, 2005). Thus, the structural genes are fractionated into introns that are spliced out from the precursor-messenger RNA (pre-mRNA) and in exons that are the expressed regions in mature mRNA (Berk and Sharp, 1977; Chow et al., 1977; Gilbert, 1978; Berk, 2016). Introns could self-cleave by acting as an enzyme (ribozymes). Now we know that there is alternative splicing and that specific genes produce different proteins, generating complex proteomes that explain the structural and functional complexity in the eukaryotes organism (Graveley, 2001; Black, 2003; Matlin et al., 2005; Pan et al., 2008; Wang et al., 2008; Nilsen and Graveley, 2010). Splicing from a pre-mRNA is an alternative mechanism for genetic regulation in higher eukaryotes. Variability in splicing model is an important source of protein diversity from the genetic code (Black, 2003).

In eukaryotes, the majority of pre-mRNAs are subject to alternative splicing, which can be regulated according to the developmental stage or cell type, or in response to signal transduction pathways (Black, 2003; Blencowe, 2006; House and Lynch, 2008). A large number of introns are sources of ncRNAs, such as mi RNAs, lncRNAs, piRNAS, and small circular RNAs, revealing the high complexity of the genomes and epigenome of eukaryotes (Tilgner et al., 2012; Yang, 2015). This evidence suggests these ncRNAs are involved in speciation processes (Lei et al., 2016). SnRNAs and proteins constituting spleciosoma, an enzyme that removes the introns, also participate in splicing (Wahl et al., 2009).

NcRNAs AND GENOMIC IMPRINTING

Genomic imprinting is an epigenetic transgenerational process that marks DNA in a sex-dependent manner, resulting in the differential expression of a gene depending on its parent of origin. Achieving an imprint requires establishing meiotically stable male and female imprints during gametogenesis and maintaining the imprinted state through DNA replication in the somatic cells of the embryo (MacDonald, 2012).

The term *imprinting* was taken from Konrad Lorenz who used it in the context of animal behavior. Helen Crouse (1960) used it in relation to dipterans of the Sciaridae family to explain the preferential removal of paternal X sex chromosomes in the somatic and germinative cells of the diptera of these sciarid flies (Crouse, 1960). During meiosis, sex X chromosomes acquire an imprint (mark) throughout the process in their passage toward the paternal line that determines a behavior opposite to that conferred by the maternal germ line (Crouse, 1960). Very similar phenomena, such as the heterochromatinization of paternal chromosomes occur in mealybug insects *Planococcus lilachinus* (Khosla et al., 1996; Bongiorni et al., 1999). For instance, in *P. citri* the haploid set of chromosomes of paternal origin, in males and females, is hypomethylated and heterochromatized, which does not happen with the haploid set derived from the mother (Brown and Nur, 1964; Brown, 1966; Bongiorni et al., 1999). Also, genomic imprinting has been found in mammals, demonstrating that androgenic and gynogenic zygotes were not functionally equivalent (McGrath and Solter, 1984; Feinberg, 2000).

Imprinting explains the inactivation by heterochromatinization of one of the sex X-chromosome in females of mammals, where one lncRNAs is transcripted from de Xist gene acting in cis position (Blignaut, 2012). The establishing of imprinting requires establishing epigenetic meiotically stable tags during meiosis in gametogenesis and also maintaining the imprinted state through DNA replication in the somatic and germinal cells of the embryo (MacDonald, 2012).

NcRNAs AND THEIR RELATION WITH THE HOX GENE, HOMEOTIC MUTATIONS AND DEVELOPMENT

Homeotic mutations are reflected in drastic, often aberrant changes in an organism's phenotypic structures by another

different during development (for example antennae by legs; Goldschmidt, 1945a,b; Dietrich, 2000). In Goldschmidt's opinion, these mutations are important in order to understand the developmental basis for morphological innovations and new species formation (Dietrich, 2003). However, these ideas were not taken into serious consideration the evolutionists of that time (Dobzhansky, 1940). Homeotic mutations are generally not adaptive, but some of them could pass the natural selection filter (Goldschmidt, 1940) and can explain the origin of biological novelties such as new species formation (Scott et al., 1989).

In light of current advances in epigenetic research, homeotic mutation could be a fundamental factor in organic evolution. In the last decade, it has been demonstrated that homeotic mutations that have to do with development in eukaryotes are controlled by ncRNAs (Petruk et al., 2006; Rinn et al., 2007). It has been discovered that miRNAs are encoded in homeotic genes (Hox genes). These miRNA genes are associated with transcription factor-encoding genes, and thus are of particular interest to the changes described above. In Hox genes there is a nucleotide sequence (homeo-domain) that is essential for embryonic development (McGinnis and Krumlauf, 1992). The homology between the homeotic invertebrated gene with vertebrate Hox genes has been demonstrated (Akam, 1989; Schubert et al., 1993; Fried et al., 2004). Therefore, these sequences are highly evolutionarily conserved and very important in the development of organism. The huge quantity of Hox miRNAs suggest that they play a significant role in Hox gene regulation during development through mRNA cleavage and translation inhibition (Yekta et al., 2004; Rinn et al., 2007).

Intergenic regions of the Hox genes in *Drosophila* produce many lncRNAs that regulate Hox gene coding sequences (Petruk et al., 2006). The studies of long ncRNAs have increased in recent times, and have become very important in expanding the knowledge of the regulation of development and other biological processes such as, heterochromatinization or diseases, and also in genomic changes (Kung et al., 2013).

NcRNAs AND TRANSGENERATIONAL EPIGENETICS

One of the great problems that Jacob and Monod solved was to find a mechanism of genetic regulation at the cellular level in *E. coli,* which they called *operon lactose* (Jacob and Monod, 1961, 1963). In the eukaryotes there were similar models that explained cellular differentiation and development (Gann, 2010). With the advances of molecular genetics, and the finding of several new modes of regulation of genetic action such as DNA methylation, histone modification and ncRNAS, the regulation of gene expression and cell differentiation has been better understood in eukaryotic organisms. These epigenetic changes, in differentiated somatic cells, can be transmitted during mitosis. But now we know that cell-to-cell inheritance can also be extended to meiotic generational inheritance between organisms (Tollefsbol, 2014). Traditionally, studies concerning the transfer of information between generations have focused on DNA as the only molecule that contains heritable genetic information, but now we know that in the epigenome there are also epigenetic marks that could be transgenerationally inherited (Jablonka et al., 2005). Epigenetic transgenerational inheritance has been defined as transmission via the germ line (sperm or egg) of epigenetic tags between generations in the absence of direct stimuli or genetic changes that drive phenotypic variation (Skinner, 2011; Yan, 2014; Yohn et al., 2015). Small ncRNAs are influential in transgenerational epigenetic inheritance because they can act as guides to specific genomic location by sequencing homology and also by recruiting various proteins to target sites, including epigenetic modifiers such as methyltransferases that are important in ADN methylation (Castel and Martienssen, 2013; Riddle, 2014).

In basal eukaryotes, such as *C. elegans*, transgenerational epigenetic inheritance mediated by ncRNAs has been described. The gene silencing induced by treatment with dsRNA in the parent is transgenerational, and inherited to the F1 offspring, proving that the silent state is transmitted through gametes to the next generation or past the F1 offspring where RNAi, siRNA, and piRNA pathways participate (Fire et al., 1998; Vastenhouw et al., 2006; Ashe et al., 2012; Riddle, 2014).

It has been found that in mammals there are various types of ncRNAs that can act in epigenetic programs. Epigenetic tags can be transmitted in somatic cells and also transgenerationally, where ncRNAs could correspond to a very important type of epigenetic inheritance mechanism (Larriba and del Mazo, 2016).

CONCLUSION

In recent years, it has been demonstrated that ncRNAs participate in many important biological process in biodiversity that aren't included in classic evolutionary theory, such as phenotypic variation, regulation of gene expression, development and transgenerational epigenetic inheritance. With these new epigenetic mechanisms, several question arose in relation to the origin and maintenance of the biodiversity in populations. In this final section, we will then try to answer some of these questions.

NcRNAs AS INTERPHASE BETWEEN THE EPIGENOTYPE AND ENVIRONMENT. GENETIC OR EPIGENETIC REVOLUTION?

Transposable elements, viruses and the RNA world, in particular the ncRNAs, open a new window into the knowledge of the processes explaining the dynamics of phenotypic changes, biodiversity and evolution. An increasing number of ncRNAS have been found in all life forms: from viruses and the simplest unicellular organisms (bacteria, archea) to the more complex eukaryotes such as mammals. These molecules have been revealed to have most varied functions, challenging the value C paradox, which was not really a paradox, but rather the lack of information regarding the functional values of an important and very dynamic area of an organism's inheritance: the epigenoma, where the different classes of ncRNAs play a fundamental role in generating evolutionary novelties.

NcRNAs participate in many biological processes, both in plants and animals, such as the regulation of transcription, development and adaptation to stressful conditions in the environment. In animals, lncRNAs regulate important processes in the central nervous system such as neurogenesis, neuron formation and synaptic plasticity related to behavior. With the advent of epigenetics and ncRNAs research, new sources of genetic variation and control of gene action have been discovered, such as splicing, RNA editing, metastable epialleles, and paramutations. NcRNAs actively participate in all these cases, generating dynamic responses to the environment and phenotypic novelties and giving rise to new species. Organisms can solve emerging problems that arise from the environment by increasing their epigenetic repertoire and dynamically developing distinct phenotypic variation without the need for new mutations or a *genetic revolution,* as has been postulated in the classic geographical model of speciation within the framework of the Synthetic Theory of Evolution (Mayr, 1949).

Furthermore, ncRNA molecules help to explain, from a molecular point of view, some classic concepts that are sources of phenotypic variation, such as pleiotropy and phenotypic plasticity. RNA splicing and RNA editing, although via different mechanisms, arise as updated explanations for the concept of pleiotropy, which itself is not adequately covered by Neo-Darwinian approaches. Plate, in 1910, describes the concept of pleitropy as a mutant gene with several phenotypic effects. As a consequence of splicing, one gene is capable of originating several proteins with different functions. This process has been proposed to be related with the increase diversity of proteomic and evolutionary diversification (Graveley, 2001; Bush et al., 2017). In RNA editing, epimutation at mRNA produces different versions of proteins with different functions in different cells (Gu et al., 2016). RNA editing increases the functional capacity of a single mRNA in different cells (Harjanto et al., 2016). This pleiotropic capacity of a unique mRNA to express itself in different cells and organs could develop varied organism phenotypes and responses in the face of environmental pressures in a rather adaptive fashion (Eddy, 2001; Mattick, 2001). In this new scenario, ncRNAs may become the artisans of the pleiotropic expression of a living organism's genome, allowing life on earth to thrive and colonize multiple habitats and overcome the boundaries of life (Khraiwesh et al., 2012; Wang et al., 2012). Considering this evidence, ncRNAs could be considered the precursor of speciation (Lake et al., 1988; Landweber and Gilbert, 1993). NcRNAs also provide an up-to-date heuristics tool for the consideration of the ontogenetic and phylogenetic consequences of environmentally inherited influences (Burggren et al., 2016).

Futhermore, it is probable that the evolution of new functional repeated RNAs has been derived from ncRNAs by retrotramposition. NcRNAs can diversify in their structure and adopt new roles (Herbert and Rich, 1999), extending the coding capacity of the genome to the epigenome. Thus, ncRNAs could be a reservoir for speciation and organic evolution (Matylla-Kulinska et al., 2014; Lei et al., 2016).

Splicing and RNA editing may also help to explain other classic concepts of phenotypic plasticity and the *norm of reaction* (Woltereck, 1909; Thoday, 1953); therefore, ncRNA appear to cover these previous definitions and processes with mechanisms.

MENDELIAN OR EPIGENETIC INHERITANCE?

Epigenetic variations in the epigenome would be inherited in a Neo-Lamarckian manner, bypassing the Weismann barrier and thus reviving Baldwin's old ideas (1896, 1897) of organic selection and Waddington's epigenetic heredity (2012) on genetic assimilation and inheritance produced by environmental pressures.

Now we know that phenotypic plasticity not only protects individuals from environmental changes, but also that there is an epigenetic control in these phenotypic changes (Moss, 2001), increasing the phenotypic variability at population level. In addition, new epigenetics tags in the epigenome could be transgenerationally inherited and populations of a species could have a different epigenetic mark but similar protein DNA code regions (Verhoeven et al., 2010; MacDonald, 2012). Experimental studies show that epigenetic variations, environmentally induced in phenotypic changes, could be inherited by future generations (Jablonka and Raz, 2009). Thus, the epigenetic variation in the epigenome corresponds to a new and important mechanism of phenotypic variation with an evolutionary perspective. This evidence has been collected from many species, including microorganisms (e.g., bacteria; Adam et al., 2008), plants (Hauser et al., 2011), and vertebrates. For instance, it has recently been described that populations of bats have different epigenetic marks suggesting that these epigenetic tags could have a correlation with phenotypic variation (Liu et al., 2015) and probably with speciation. In social insects, ncRNA related epigenetic changes have been found playing key roles in varied biological dynamics, from development to behavioral processes (Asgari, 2013). For example, studies of miRNA population diversity among *Apis mellifera* castes demonstrated striking differences between miRNA from nursing and foraging bees. Furthermore, in that study it was found that some of these ncRNA molecules were related to neural functions (Liu et al., 2012). Metastable epialleles and paramutations, which occur at the level of gene alleles, are also a source of novel epigenetic variability that help explain phenotypic variegation phenomena and also previously unknown aspects of classical quantitative genetics. These epigenetics changes would be inherited by genomic imprinting.

Environmental stressors induce epigenetic changes at epigenome level where several ncRNAs motile elements and viruses participate. These can explain some non-Mendelian models of inheredity. NcRNAs process and store a lot of information from environmental signals against unfavorable environmental conditions. In the adult rat it has been described that cells exposed to traumatic conditions during early life have different types and amounts of miRNAs in their blood, brain, and spermatozoids in comparison to the non-traumatized individuals. Some of these miRNAs were produced in excess while others were underrepresented in comparison with control

animals. These changes resulted from deficient regulation of cell processes controlled by these miRNAs (Gapp et al., 2014). These behavioral symptoms were also observed in the offspring of treated groups, despite the fact that these pups were never exposed to stress during their own ontogeny, suggesting that germ line epigenetic marks were alerted due to the paternal stress and that such alteration was then inherited trough the spermatozoids (Gapp et al., 2014). It is becoming increasingly evident that the surrounding environment leaves epigenetic footprints on brains, organs, and also gametes, in which case epigenetic marks may even pass to the next generation (reviewed by Denhardt, 2017; Mulder et al., 2017). Thus, populations with their epigenetic repertoire increase the adaptive behavior and phenotypic plasticity of their individuals, allowing an organism's structural coupling with its environment (Maturana-Romesín and Mpodozis, 2000). All this is thanks to the development of distinct epigenotypes helped by ncRNAs and without concomitant mutations to the underlying genes. Under this novel epigenetic understanding of gene expression and phenotypic variation, we find an explanation for the current phenotypic variation and biodiversity on our planet, without resorting to mutation as the only source of evolution.

WHERE DOES NATURAL SELECTION ACT?

Transgenerational epigenetic inheritance tells us that natural selection acts on the epigenome of the organism (Ruden et al., 2015), specifically on ncRNAs, which correspond to the interface between the genotype and the environment, capturing environmental signals. This contradicts one of the fundamental ideas of population genetics, which establishes that natural selection acts on the genotypes of the individuals in the population.

Making an analogy between an organism and a building: If a catastrophic event occurs, it acts directly on the building and not on the blueprints. The resistance of the building to the catastrophe will depend on the quality of the materials used in construction. The genome corresponds to the blueprints of the building, while the epigenome is the construction company and the workers who make the building (viruses, transposable elements, ncARNs). Biotic and abiotic environmental factors are fundamental during the development process, and that will depend on the capabilities that the organism has for overcoming the negative aspects of natural selection (Furrow, 2014; Burggren et al., 2016).

HOW DO NEW SPECIES ORIGINATE? THROUGH MUTATIONS OR THROUGH EPIMUTATIONS?

With the advent of epigenetics, and transgenerational inheritance, it is now possible to propose as a hypothesis that the very epigenetic mechanisms that regulate ontogenetic gene expression and cell differentiation also intervene in the origin of new species in a phylogenetic dimension. In other words, the organisms' behaviors in response to environmental pressures leave its epigenetic marks, via similar epigenetic paths (ncRNAs) both during individual's life as well as transgenerationally, through its progeny. NcRNAs are complementary to the role of proteins in the model proposed by Jacob and Monod, which refers to the mechanisms of regulation of gene expression during development (Gann, 2010). Both processes integrally contribute to an understanding of the mechanisms of organic development and evolution (EvoDevo) and the genome-epigenome circuit. For instance, the differences of structural genes in chimpanzees and humans is only about 4% (Varki and Altheide, 2005). However, the phenotypic differences between them are significantly higher and are probably due to differences in the epigenome of these species. Under current ncRNA evidence, speciation should be considered a process where the epigenomic changes are caused by the pressures of the environment. The landscape of ncRNAs in an organism not only allows cellular differentiation and development in eukaryotes, but also relief from the negative effects of stress and natural selection, as has been demonstrated in model system organisms as well as in our own species.

Epigenetic changes involving ncRNAs that produce phenotype variability (epimutation, splicing, and RNA editing) may have an adaptive value for individuals who are carriers of these variations (Steele et al., 1998). However, they do not follow the Mendelian principles of heredity and are closer to the model proposed by Lamarck on the inheritance of acquired characteristics, foundations now denominated Neo-Lamarckism (Jablonka et al., 2005; Jablonka and Raz, 2009). Based on current findings, ncRNAs arise as active vehicles for epigenetic variation, phenotypic plasticity and heredity, revisiting classic concepts, and contributing with mechanistic explanatory power to a

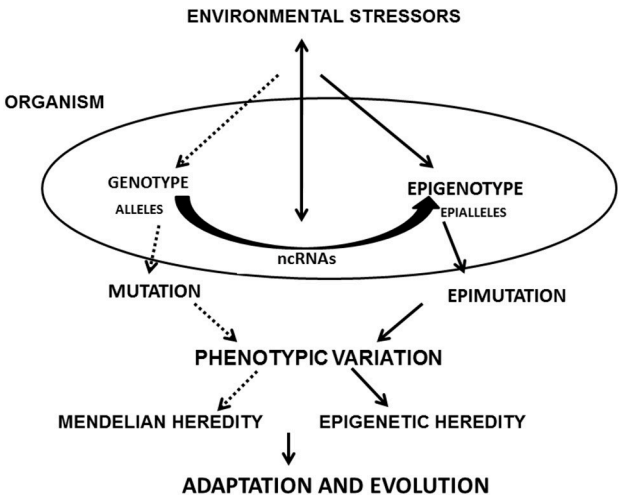


FIGURE 1 | Flow Diagram showing the paths where ncRNAs are involved in the development of phentopic variations and evolution: environmental stressors act on the epigenotype and genotype. ncRNAs receive and respond to these stimuli. These epigenetic influences, together with the effect of mutations and Mendelian inheritance originate new adaptations and evolutionary noveties would arise.

non-reductionist view of modern biology and the evolution of species (**Figure 1**).

NcRNAs AND THEIR IMPORTANCE IN MOLECULAR COADAPTATION AND EVOLUTION

A genome's molecular structure, both in the animal and plant kingdom, demonstrates that ncRNAs are scattered among the species that constitute the three domains of the tree of life. These ncRNAs act as co-adapted endosymbiotic molecules with the genome and epigenome of their hosts and are the product of molecular coevolution from the origins of the first cells. With the exception of ribozyme, the most relictual molecules in organic evolution (as proposed by Gilbert, 1986), all the others ncRNAs require interaction with different protein molecules to exert their regulatory epigenetic function on genetic expression. Therefore, a primary stage in the evolutionary process that gave rise to the first cells, and the subsequent diversification of living forms, consisted of a molecular coevolution forming dynamic co-adapted molecular complexes. Without this molecular co-adaptation, organic evolution would not have been possible. The increasing number and diversity of these small and long ncRNAS in relation to the complexity and adaptability of living beings, explains that they have been paramount in complex biological processes and are not an evolutionary paradox.

The fact that miRNAs can be mobilized by the fluids of plants and animals, allows them to act at different distances to where they were transcribed, much like hormones or pheromones do. In addition, they can respond to environmental stimuli, favoring the adaptation of organisms through the modification of epigenetic marks and also a transgenerational inheredity and the evolution of species as part of a Neo-Lamarckian model.

AUTHOR CONTRIBUTIONS

Main hypothesis developed by DF-L. Blibliographic review and secondary writing by CV.

ACKNOWLEDGMENTS

Financed by Research Direction of Metropolitan University of Educational Sciences (DIUMCE).

REFERENCES

Adam, M., Murali, B., Glenn, N. O., and Potter, S. S. (2008). Epigenetic inheritance based evolution of antibiotic resistance in bacteria. *BMC Evol. Biol.* 8:52. doi: 10.1186/1471-2148-8-52

Agrawal, N., Dasaradhi, P. V. N., Mohmmed, A., Malhotra, P., Bhatnagar, R. K., and Mukherjee, S. K. (2003). RNA Interference: biology, mechanism, and applications. *Microbiol. Mol. Biol. Rev.* 67, 657–685. doi: 10.1128/MMBR.67.4.657-685.2003

Akam, M. (1989). Hox and HOM: homologous gene clusters in insects and vertebrates. *Cell* 57, 347–349. doi: 10.1016/0092-8674(89)90909-4

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437, 1149–1152. doi: 10.1038/nature04107

Asgari, S. (2011). Role of MicroRNAs in insect host–microorganism interactions. *Front. Physiol.* 2:48. doi: 10.3389/fphys.2011.00048

Asgari, S. (2013). MicroRNA functions in insects. *Insect Biochem. Mol. Biol.* 43, 388–397. doi: 10.1016/j.ibmb.2012.10.005

Ashe, A., Sapetschnig, A., Weick, E. M., Mitchell, J., Bagijn, M. P., Cording, A. C., et al. (2012). PiRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans. Cell* 150, 88–99. doi: 10.1016/j.cell.2012.06.018

Baldwin, J. M. (1896). A new factor in evolution. *Am. Nat.* 30, 441–451.

Baldwin, J. M. (1897). Organic selection. *Science* 5, 634–636.

Bannister, A. J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res.* 21, 381–395. doi: 10.1038/cr.2011.22

Benne, R., Van Den Burg, J., Brakenhoff, J. P. J., Sloof, P., Van Boom, J. H., and Tromp, M. C. (1986). Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46, 819–826. doi: 10.1016/0092-8674(86)90063-2

Berk, A. J. (2016). Discovery of RNA splicing and genes in pieces. *Proc. Natl. Acad. Sci. U.S.A.* 113, 801–805. doi: 10.1073/pnas.1525084113

Berk, A. J., and Sharp, P. A. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* 12, 721–732. doi: 10.1016/0092-8674(77)90272-0

Bidarimath, M., Khalaj, K., Wessels, J. M., and Tayade, C. (2014). MicroRNAs, immune cells and pregnancy. *Cell. Mol. Immunol.* 11, 538–547. doi: 10.1038/cmi.2014.45

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21. doi: 10.1101/gad.947102

Bird, A. (2007). Perceptions of epigenetics. *Nature* 447, 396–398. doi: 10.1038/nature05913

Black, D. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336. doi: 10.1146/annurev.biochem.72.121801.161720

Blanc, V., and Davidson, N. (2003). C-to-U RNA editing: mechanisms leading to genetic diversity. *J. Biol. Chem.* 278, 1395–1398. doi: 10.1074/jbc.R200024200

Blencowe, B. J. (2006). Alternative splicing: new insights from global analyses. *Cell* 126, 37–47. doi: 10.1016/j.cell.2006.06.023

Blignaut, M. (2012). Review of Non-coding RNAs and the epigenetic regulation of gene expression: a book edited by Kevin Morris. *Epigenetics* 7, 664–666. doi: 10.4161/epi.20170

Bongiorni, S., Cintio, O., and Prantera, G. (1999). The relationship between DNA methylation and chromosome imprinting in the Coccid *Planococcus citri*. *Genetics* 151, 1471–1478.

Brink, R., Style, E., and Axtell, J. (1968). Paramutation directed genetic change. Paramutation occurs in somatic cells and heritable alters the functional state of a locus. *Science* 159, 161–170.

Brown, S. (1966). Heterochromatin. *Science* 28, 417–425.

Brown, S. W., and Nur, U. (1964). Heterochromatic chromosomes in the coccids. *Science* 145, 130–136.

Burggren, W., O'callaghan, C., Finne, J., and Torday, J. S. (2016). Epigenetic inheritance and its role in evolutionary biology: re-evaluation and new perspectives. *Biology* 4:22. doi: 10.3390/biology5020024

Bush, S. J., Chen, L., Tovar-Corona, J. M., and Urrutia, A. O. (2017). Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372, 1–7. doi: 10.1098/rstb.2015.0474

Cao, J. (2014). The functional role of long non-coding RNAs and epigenetics. *Biol. Proced. Online* 16:11. doi: 10.1186/1480-9222-16-11

Castel, S., and Martienssen, R. (2013). RNA interference in the nucleous:roles,:roles for smalls RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.* 14, 100–112. doi: 10.1038/nrg3355

Cavalli, G., and Paro, R. (1999). Epigenetic inheritance of active chromatin after removal of the main transactivator. *Science* 286, 955–958. doi: 10.1126/science.286.5441.955

Chandler, V. L. (2007). Paramutation: from maize to mice. *Cell* 128, 641–645. doi: 10.1016/j.cell.2007.02.007

Chisholm, K. M., Wan, Y., Li, R., Montgomery, K. D., Chang, H. Y., and West, R. B. (2012). Detection of Long Non-Coding RNA in archival tissue: correlation

with polycombprotein expression in primary and metastatic breast Carcinoma. *PLoS ONE* 7:e47998. doi: 10.1371/journal.pone.0047998

Choudhuri, S. (2010). Small noncoding RNAs: biogenesis, function, and emerging significance in toxicology. *J. Biochem. Mol. Toxicol.* 24, 195–216. doi: 10.1002/jbt.20325

Chow, L. T., Roberts, J. M., Lewis, J. B., and Broker, T. R. (1977). A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell* 11, 819–836. doi: 10.1016/0092-8674(77)90294-X

Coe, E. H. (1966). The properties, origin, and mechanism of conversion-type inheritance at the B locus in maize. *Genetics* 53, 1035–1063.

Creevey, C. J., and McInerney, J. O. (2003). CRANN: detecting adaptive evolution in protein-coding DNA sequences. *Bioinformatics* 19:1726. doi: 10.1093/bioinformatics/btg225

Crick, F. (1958). "On protein synthesis," in *The Symposia of the Society for Experimental Biology, No. XII: Biological Replication Macromolecules,* ed F. K. Sanders (Cambridge, UK: Cambridge University Press), 138–163.

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563.

Crouse, H. V. (1960). The controlling element in sex chromosome behavior in Sciara. *Genetics* 45, 1429–1443.

Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *Plant Cell* 23, 431–442. doi: 10.1105/tpc.110.082784

D'Urso, A., and Brickner, J. H. (2014). Mechanisms of epigenetic memory. *Trends Genet.* 30, 230–236. doi: 10.1016/j.tig.2014.04.004

De la Peña, M., and Garcia-Robles, I. (2010a). Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep.* 11, 711–716. doi: 10.1038/embor.2010.100

De la Peña, M., and Garcia-Robles, I. (2010b). Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. *RNA* 16, 1943–1950. doi: 10.1261/rna.2130310

De Lucia, F., and Dean, C. (2011). Long non-coding RNAs and chromatin regulation. *Curr. Opin. Plant Biol.* 14, 168–173. doi: 10.1016/j.pbi.2010.11.006

Denhardt, D. T. (2017). Effect of stress on human biology: Epigenetics, adaptation, inheritance, and social significance. *J. Cell. Physiol.* 233, 1975–1984. doi: 10.1002/jcp.25837

Dietrich, M. R. (2000). From hopeful monsters to homeotic effects: Richard Goldschmidt's integration of development, evolution, and genetics. *Am. Zool.* 40, 738–747. doi: 10.1093/icb/40.5.738

Dietrich, M. R. (2003). Richard Goldschmidt: hopeful monsters and other "heresies." *Nat. Rev. Genet.* 4, 68–74. doi: 10.1038/nrg979

Dobzhansky, T. (1940). Catastrophism versus evolutionism. *Science* 98, 356–358.

Dulcis, D., Lippi, G., Stark, C. J., Do, L. H., Berg, D. K., and Spitzer, N. C. (2017). Neurotransmitter switching regulated by miRNAs controls changes in social preference. *Neuron* 95, 1–15. doi: 10.1016/j.neuron.2017.08.023

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA World. *Nat. Rev. Genet.* 2, 919–929. doi: 10.1038/35103511

Feagin, J. E., Abraham, J. M., and Stuart, K. (1988). Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. *Cell* 53, 413–422.

Feinberg, A. (2000). "DNA methylation, genomic imprinting and cancer," in *Current Topics in Microbiology and Immunology*, eds P. Jones and P. Vog (Berlín: Springer-Verlag), 87–99.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391, 806–811. doi: 10.1038/35888

Flanagan, J. M., and Wild, L. (2007). An epigenetic role for noncoding RNAs and intragenic DNA methylation. *Genome Biol.* 8:307. doi: 10.1186/gb-2007-8-6-307

Frías, D. (2010). Omissions in the synthetic theory of evolution. *Biol. Res.* 43, 299–306. doi: 10.4067/S0716-97602010000300006

Frías-Lasserre, D. (2012). Non coding RNAs and viruses in the framework of the phylogeny of the genes, epigenesis and heredity. *Int. J. Mol. Sci.* 13, 477–490. doi:10.3390/ijms13010477

Fried, C., Prohaska, S. J., and Stadler P. F. (2004). Exclusion of repetitive DNA elements from gnathostome hox clusters. *J. Exp. Zool. Mol. Dev. Evol.* 302B, 165–173. doi: 10.1002/jez.b.20007

Furrow, R. E. (2014). Epigenetic inheritance, epimutation, and the response to selection. *PLoS ONE* 9:e101559. doi: 10.1371/journal.pone.0101559

Gaiti, F., Calcino, A. D., Tanurdzic, M., and Degnan, B. M. (2016). Origin and evolution of the metazoan non-coding regulatory genome. *Dev. Biol.* 427, 193–202. doi: 10.1016/j.ydbio.2016.11.013

Gann, A. (2010). Jacob and Monod: from operons to EvoDevo. *Curr. Biol.* 20, 718–723. doi: 10.1016/j.cub.2010.06.027

Gapp, K., Jawaid, A., Sarkies, P., Bohacek, J., Pelczar, P., Prados, J., et al. (2014). Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nat. Neurosci.* 17, 667–669. doi: 10.1038/nn.3695

Ghildiyal, M., and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* 10, 94–108. doi: 10.1038/nrg2504

Gilbert, W. (1978). Why genes in pieces? *Nature* 271:501. doi: 10.1038/271501a0

Gilbert, W. (1986). Origin of life: the RNA world. *Nature* 319:618.

Gillings, M. R., and Westoby, M. (2014). DNA technology and evolution of the Central Dogma. *Trends Ecol. Evol.* 29, 1–2. doi: 10.1016/j.tree.2013.10.001

Goldschmidt, R. (1940). *The Material Basis of Evolution*. New Haven, CT: Yale University Press.

Goldschmidt, R. (1945a). Evolution of mouth part in Diptera a counter critique. *Pan. Pac. Entomol.* 21, 41–47.

Goldschmidt, R. (1945b). Podoptera a homeotic mutant in Drosophila and the origin of the insects wing. *Science* 11, 389–380.

Gommans, W. M., Mullen, S. P., and Maas, S. (2009). RNA editing: a driving force for adaptive evolution? *Bioessays* 31, 1137–1145. doi: 10.1002/bies.200900045

Gott, J. M., and Emeson, R. B. (2000). Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34, 499–531. doi: 10.1146/annurev.genet.34.1.499

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107. doi: 10.1016/S0168-9525(00)02176-4

Grewal, S. I. S., and Klar, A. J. S. (1996). Chromosomal inheritance of epigenetic states in fission yeast during mitosis and meiosis. *Cell* 86, 95–101. doi: 10.1016/S0092-8674(00)80080-X

Gu, T., Gatti, D. M., Srivastava, A., Snyder, E. M., Raghupathy, N., Simecek, P., et al. (2016). Genetic architectures of quantitative variation in RNA editing pathways. *Genetics* 202, 787–798. doi: 10.1534/genetics.115.179481

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature07672

Hanson, M. A., and Skinner, M. K. (2016). Developmental origins of epigenetic transgenerational inheritance. *Env. Epigenetic* 2, 1–21. doi: 10.1093/eep/dvw002.Developmental

Harjanto, D., Papamarkou, T., Oates, C. J., Rayon-Estrada, V., Papavasiliou, F. N., and Papavasiliou, A. (2016). RNA editing generates cellular subsets with diverse sequence within populations. *Nat. Commun.* 7:12145. doi: 10.1038/ncomms12145

Hauser, M. T., Aufsatz, W., Jonak, C., and Luschnig, C. (2011). Transgenerational epigenetic inheritance in plants. *Biochim. Biophys. Acta* 1809, 459–468. doi: 10.1016/j.bbagrm.2011.03.007

Herbert, A., and Rich, A. (1999). RNA processing and the evolution of eukaryotes. *Nat. Genet.* 21, 265–269. doi: 10.1038/6780

Hollick, J. (2010). Paramutation and development. *Annu. Rev. Cell Dev. Biol.* 26, 557–579. doi: 10.1146/annurev.cellbio.042308.113400

House, A. E., and Lynch, K. W. (2008). Regulation of alternative splicing: more than just the ABCs. *J. Biol. Chem.* 283, 1217–1221. doi: 10.1074/jbc.R700031200

House, M., and Lukens, L. (2014). "The role of germinally inherited epialleles in plant breeding," in *Epigenetics in Plants of Agronomic Importance: Fundamentals and Applications*, eds R. Alvarez-Venegas, C. De la Peña, and J. Casas-Mollano (London: Springer), 1–11.

Jablonka, E., and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol.* 84, 131–176. doi: 10.1007/s13398-014-0173-7.2

Jablonka, E., Lamb, M. J., and Zeligowski, A. (2005). Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life. *J. Clin. Invest.* 115:2961. doi: 10.1172/JCI27017

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356. doi: 10.1016/S0022-2836(61)80072-7

Jacob, F., and Monod, J. (1963). "Genetic repression, allosteric inhibition, and cellular differentiation," in *Cytodifferentiation and Macromolecular Synthesis*, ed M. Locke (New York, NY; London: Academic Press), 30–64.

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33(Suppl.), 245–254. doi: 10.1038/ng1089

Joh, R. I., Palmieri, C. M., Hill, I. T., and Motamedi, M. (2014). Regulation of histone methylation by noncoding RNAs. *Biochim. Biophys. Acta* 1839, 1385–1394. doi: 10.1016/j.bbagrm.2014.06.006

Kawasaki, H., and Taira, K. (2004). Induction of DNA methylation and gene silencing by short interfering RNAs in human cells. *Nature* 431, 211–216. doi: 10.1038/nature02889

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science* 303, 1626–1632. doi: 10.1126/science.1089670

Keller, C., and Bühler, M. (2013). Chromatin-associated ncRNA activities. *Chromosom. Res.* 21, 627–641. doi: 10.1007/s10577-013-9390-8

Khosla, S., Kantheti, P., Brahmachari, V., and Chandra, H. S. (1996). A male-specific nuclease-resistant chromatin fraction in the mealybug *Planococcus lilacinus*. *Chromosoma* 104, 386–392. doi: 10.1007/s004120050130

Khraiwesh, B., Zhu, J. K., and Zhu, J. (2012). Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim. Biophys. Acta* 1819, 137–148. doi: 10.1016/j.bbagrm.2011.05.001

Kim, V. (2006). Small RNAs just got bigger, Piwi-interacting RNAs (piRNAs) in mammalian testes. *Gene Dev.* 20, 1993–1997. doi: 10.1101/gad.1456106

Klose, R. J., and Bird, A. P. (2006). Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* 31, 89–97. doi: 10.1016/j.tibs.2005.12.008

Knowles, D. G., and McLysaght, A. (2009). Recent *de novo* origin of human protein-coding genes. *Genome Res.* 19, 1752–1759. doi: 10.1101/gr.095026.109

Koerner, M. V., Pauler, F. M., Huang, R., and Barlow, D. P. (2009). The function of non-coding RNAs in genomic imprinting. *Development* 136, 1771–1783. doi: 10.1242/dev.030403

Kosten, T., and Nielsen, D. (2014). "Maternal epigenetic inheritance and stress during gestation: focus on brain and behavioral disorders," in *Transgenerational Epigenetics, Evidence and Debate*, ed T. Tollefsbol (New York, NY: Elsevier AP), 197–214.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., et al. (2012). Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 493, 231–235. doi: 10.1038/nature11661

Kulis, M., Queirós, A. C., Beekman, R., and Martín-Subero, J. I. (2013). Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim. Biophys. Acta* 1829, 1161–1174. doi: 10.1016/j.bbagrm.2013.08.001

Kung, J. T. Y., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics* 193, 651–669. doi: 10.1534/genetics.112.146704

Kurokawa, R. (2015). "Long noncoding RNAs," in *Structures and Functions*, ed R. Kurokawa (Tokyo: Springer). doi: 10.1007/978-4-431-55576-6

Lake, J., de la Cruz, V. F., Ferreira, P. C., Morel, C., and Simpson, L. (1988). Evolution of parasitism: kinetoplastid protozoan history reconstructed from mitochondrial rRNA gene sequences. *Proc. Natl. Acad. Sci. U.S.A.* 85, 4779–4783. doi: 10.1073/pnas.88.6.2612a

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Landweber, L. F., and Gilbert, W. (1993). RNA editing as a source of genetic variation. *Nature* 363, 179–182. doi: 10.1038/363179a0

Larriba, E., and del Mazo, J. (2016). Role of non-coding RNAs in the transgenerational epigenetic transmission of the effects of reprotoxicants. *Int. J. Mol. Sci.* 17:452. doi: 10.3390/ijms17040452

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843–854. doi: 10.1016/0092-8674(93)90529-Y

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., and Zhou, R. (2016). Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol.* 8, 562–577. doi: 10.1093/gbe/evw025

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature* 366, 362–365. doi: 10.1038/366362a0

Ling, L., and Wurtele, E. S. (2014). The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. *Plant Biotechnol. J.* 13, 177–187. doi: 10.1111/pbi.12238

Li, Y., Zhang, Y., Li, S., Lu, J., Chen, J., Wang, Y., et al. (2015). Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer. *Sci. Rep.* 5:8790. doi: 10.1038/srep08790

Liang, H., and Landweber, L. F. (2007). Hypothesis: RNA editing of microRNA target sites in humans? *RNA* 13, 463–467. doi: 10.1261/rna.296407

Lim, L. P., Lim, L. P., Lau, N. C., Lau, N. C., Weinstein, E. G., Weinstein, E. G., et al. (2003). The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* 17, 991–1008. doi: 10.1101/gad.1074403

Liu, F., Peng, W., Li, Z., Li, W., Li, L., Pan, J., et al. (2012). Next-generation small RNA sequencing for microRNAs profiling in *Apis mellifera*: comparison between nurses and foragers. *Insect Mol. Biol.* 21, 297–303. doi: 10.1111/j.1365-2583.2012.01135.x

Liu, S., Sun, K., Jiang, T., and Feng, J. (2015). Natural epigenetic variation in bats and its role in evolution. *J. Exp. Biol.* 218, 100–106. doi: 10.1242/jeb.107243

Louro, R., El-Jundi, T., Nakaya, H. I., Reis, E. M., and Verjovski-Almeida, S. (2008). Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. *Genomics* 92, 18–25. doi: 10.1016/j.ygeno.2008.03.013

Luciano, D. J., Mirsky, H., Vendetti, N. J., and Maas, S. (2004). RNA editing of a miRNA precursor. *RNA* 10, 1174–1177. doi: 10.1261/rna.7350304

Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. *Cell* 144, 16–26. doi: 10.1016/j.cell.2010.11.056

Luco, R. F., and Misteli, T. (2011). More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr. Opin. Genet. Dev.* 21, 366–372. doi: 10.1016/j.gde.2011.03.004

Lunter, G., Ponting, C. P., and Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* 2:e5. doi: 10.1371/journal.pcbi.0020005

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., and Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol.* 8:506. doi: 10.1371/journal.pbio.1000506

MacDonald, W. A. (2012). Epigenetic mechanisms of genomic imprinting: common themes in the regulation of imprinted regions in mammals, plants, and insects. *Genet. Res. Int.* 2012:585024. doi: 10.1155/2012/585024

Mahfouz, M. M. (2010). RNA-directed DNA methylation: mechanisms and functions. *Plant Signal. Behav.* 5, 806–816. doi: 10.4161/psb.5.7.11695

Mannoor, K., Liao, J., and Jiang, F. (2012). Small nucleolar RNAs in cancer. *Biochim. Biophys. Acta Rev. Cancer* 1826, 121–128. doi: 10.1016/j.bbcan.2012.03.005

Martick, M., Horan, L. H., Noller, H. F., and Scott, W. G. (2008). A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. *Nature* 454, 899–902. doi: 10.1038/nature07117

Mashoodh, R., and Champagne, F. (2014). "Paternal epigenetic inheritance," in *Transgenerational Epigenetics, Evidence and Debate*, ed T. Tollefsbol (New York, NY: Elsevier AP), 221–232.

Matlin, A. J., Clark, F., and Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* 6, 386–398. doi: 10.1038/nrm1645

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* 2, 986–991. doi: 10.1093/embo-reports/kve230

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS Genet.* 5:e1000459. doi: 10.1371/journal.pgen.1000459

Maturana-Romesín, H., and Mpodozis, J. (2000). The origin of species by means of natural drift. *Rev. Chilena Hist. Nat.* 73, 203–310. doi: 10.4067/S0716-078X2000000200005

Matylla-Kulinska, K., Tafer, H., Weiss, A., and Schroeder, R. (2014). Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs. *Wiley Interdiscip. Rev. RNA* 5, 591–600. doi: 10.1002/wrna.1243

Mayr, E. (1949). "Speciation and systematic," in *Genetics, Plaelontology and Evolution,* eds G. L. Jepsen, G. G. Simpson, and E. Mayr (Columbia, NY: Princeton University Press), 281–298.

Mayr, E. (1985). Weismann and evolution. *J. Hist. Biol.* 18, 295–329. doi: 10.1007/BF00138928

McClintock, B. (1950). The origin and behavior of Mutable Loci in Maize. *Genetics* 36, 344–355. doi: 10.1073/pnas.36.6.344

McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and axial patterning. *Cell* 68, 283–302. doi: 10.1016/0092-8674(92)90471-N

McGrath, J., and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell* 37, 179–183. doi: 10.1016/0092-8674(84)90313-1

Mehler, M. F., and Mattick, J. S. (2007). Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. *Physiol. Rev.* 87, 799–823. doi: 10.1152/physrev.00036.2006

Mendell, J. T., and Olson, E. N. (2012). MicroRNAs in stress signaling and human disease. *Cell* 148, 1172–1187. doi: 10.1016/j.cell.2012.02.005

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi: 10.1038/nrg2521

Mondal, T., Rasmussen, M., Pandey, G. K., Isaksson, A., and Kanduri, C. (2010). Characterization of the RNA content of chromatin. *Genome Res.* 20, 899–907. doi: 10.1101/gr.103473.109

Moss, L. (2001). "Deconstructing the gene and reconstructing molecular developmental systems," in *Cycles of Contingency: Developmental Systems and Evolution*, eds S. Oyama, P. E. Griffiths, and R. D. Gray (Cambridge: MIT Press), 85–97.

Mouillet, J. F., Ouyang, Y., Bayer, A., Coyne, C. B., and Sadovsky, Y. (2014). The role of trophoblastic microRNAs in placental viral infection. *Int. J. Dev. Biol.* 58, 281–289. doi: 10.1387/ijdb.130349ys

Mulder, R. H., Rijlaarsdam, J., and Van IJzendoorn, M. H. (2017). "DNA methylation: a mediator between parenting stress and adverse child development?," in *Parental Stress and Early Child Development Adaptive and Maladaptive Outcomes*, eds K. Deater-Deckard and R. Panneton (Berlin: Springer International Publishing), 157–180.

Muniz, L., Egloff, S., and Kiss, T. (2013). RNA elements directing *in vivo* assembly of the 7SK/MePCE/Larp7 transcriptional regulatory snRNP. *Nucleic Acids Res.* 41, 4686–4698. doi: 10.1093/nar/gkt159

Nei, M. (2013). *Mutation-Driven Evolution*. Oxford: Oxford University Press.

Nilsen, T. W., and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463, 457–463. doi: 10.1038/nature 08909

Ouyang, Y., Mouillet, J. F., Coyne, C. B., and, Sadovsky, Y. (2013). Review: placentaspecific microRNAs in exosomes e Good things come in nano-packages. *Placenta* 35, 1–5. doi: 10.1016/j.placenta.2013. 11.002

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Penn, A. C., Balik, A., and Greger, I. H. (2013). Reciprocal regulation of A-to-I RNA editing and the vertebrate nervous system. *Front. Neurosci.* 7:61. doi: 10.3389/fnins.2013.00061

Petruk, S., Sedkov, Y., Riley, K. M., Hodgson, J., Schweisguth, F., Hirose, S., et al. (2006). Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. *Cell* 127, 1209–1221. doi: 10.1016/j.cell.2006.10.039

Pluskota, W. E., Martínez-Andújar, C., Martin, R. C., and Nonogaki, H. (2011). "MicroRNA function in seed biology," in *Non Coding RNAs in Plants. RNA Technologies*, eds V. Erdmann and J. Barciszewski (Berlin; Heidelberg: Springer), 339–357. doi: 10.1007/978-3-642-19454-2_21

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641. doi: 10.1016/j.cell.2009.02.006

Przybilski, R., Gräf, S., Lescoute, A., Nellen, W., Westhof, E., Steger, G., et al. (2005). Functional hammerhead ribozymes naturally encoded in the genome of *Arabidopsis thaliana*. *Plant Cell* 17, 1877–1885. doi: 10.1105/tpc.105. 032730

Pulukuri, S. M. K., Knost, J., Estes, N., and Rao, J. S. (2009). Small interfering RNA-directed knockdown of uracil DNA glycosylase induces apoptosis and sensitizes human prostate cancer cells to genotoxic stress. *Mol. Cancer Res.* 7, 1285–1293. doi: 10.1158/1541-7786.MCR-08-0508

Qu, Z., and Adelson, D. L. (2012). Evolutionary conservation and functional roles of ncRNA. *Front. Genet.* 3:205. doi: 10.3389/fgene.2012.00205

Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I., and Whitelaw, E. (2002). Metastable epialleles in mammals. *Trends Genet.* 18, 348–351. doi: 10.1016/S0168-9525(02)02709-9

Rangwala, S. H., and Richards, E. J. (2004). The value-added genome: building and maintaining genomic cytosine methylation landscapes. *Curr. Opin. Genet. Dev.* 14, 686–691. doi: 10.1016/j.gde.2004.09.009

Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., and Cuzin, F. (2006). RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* 441, 469–474. doi: 10.1038/nature04674

Richards, E. J., and Elgin, S. C. R. (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* 108, 489–500. doi: 10.1016/S0092-8674(02)00644-X

Riddle, N. (2014). "Heritable generational epigenetic effects through RNA," in *Transgenerational Epigenetics. Evidence and Debate,* ed T. Tollefsbol (New York, NY: Elsevier AP), 105–119.

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX Loci by noncoding RNAs. *Cell* 129, 1311–1323.doi: 10.1016/j.cell.2007.05.022

Rivera, C., Gurard-Levin, Z. A., Almouzni, G., and Loyola, A. (2014). Histone lysine methylation and chromatin replication. *Biochim. Biophys. Acta* 1839, 1433–1439. doi: 10.1016/j.bbagrm.2014.03.009

Rubio, M. A. T., Pastar, I., Gaston, K. W., Ragone, F. L., Janzen, C. J., Cross, G. A. M., et al. (2007). An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7821–7826. doi: 10.1073/pnas.0702394104

Ruden, D. M., Cingolani, P. E., Sen, A., Qu, W., Wang, L., Senut, M. C., et al. (2015). Epigenetics as an answer to Darwin's "special difficulty," part 2: Natural selection of metastable epialleles in honeybee castes. *Front. Genet.* 5:60. doi: 10.3389/fgene.2015.00060

Sabin, L. R., Delás, M. J., and Hannon, G. J. (2013). Dogma derailed: the many influences of RNA on the genome. *Mol. Cell* 49, 783–794. doi: 10.1016/j.molcel.2013.02.010

Sakaguchi, K. (1990). Invertrons, a class of structurally and functionally related genetic elements that includes linear DNA plasmids, transposable elements, and genomes of adeno-type viruses. *Microbiol. Rev.* 54, 66–74.

Salta, E., and De Strooper, B. (2012). Non-coding RNAs with essential roles in neurodegenerative disorders. *Lancet Neurol.* 11, 189–200. doi: 10.1016/S1474-4422(11)70286-1

Schreiber, S. L. (2005). Small molecules: the missing link in the central dogma. *Nat. Chem. Biol.* 1, 64–66. doi: 10.1038/nchembio0705-64

Schubert, F. R., Nieselt-Struwe, K., and Gruss, P. (1993). The Antennapedia-type homeobox genes have evolved from three precursors separated early in metazoan evolution. *Proc. Natl. Acad. Sci. U.S.A.* 90, 143–147. doi: 10.1073/pnas.90.1.143

Scott, P., Tamkun, J., and Hartzell, G. (1989). The structure and function of the homeodomain. *Biochim. Biophys. Acta* 989, 25–48.

Seehafer, C., Kalweit, A., Steger, G., Gräf, S., and Hammann, C. (2011). From alpaca to zebrafish: hammerhead ribozymes wherever you look. *RNA* 17, 21–26. doi: 10.1261/rna.2429911

Serganov, A., and Dinshaw, J. P. (2007). Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.* 8, 776–790. doi: 10.1038/nrg2172

Shapiro, J. A. (2009). Revisiting the central dogma in the 21st century. *Ann. N.Y. Acad. Sci.* 1178, 6–28. doi: 10.1111/j.1749-6632.2009.04990.x

Sharp, P. A. (2005). The discovery of split genes and RNA splicing. *Trends Biochem. Sci.* 30, 279–281. doi: 10.1016/j.tibs.2005.04.002

Siegfried, Z., and Simon, I. (2010). DNA methylation and gene expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 362–371. doi: 10.1002/wsbm.64

Singh, M. (2013). Dysregulated A to I RNA editing and non-coding rnas in neurodegeneration. *Front. Genet.* 3:326. doi: 10.3389/fgene.2012.00326

Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* 12, 246–258. doi: 10.1038/nrm3089

Skinner, M. K. (2011). Environmental epigenetic transgenerational inheritance and somatic epigenetic mitotic stability. *Epigenetics* 6, 838–842. doi: 10.4161/epi.6.7.16537

Slack, J. (2002). Conrad Hal Waddington: the last Renaissance biologist? *Nat. Rev. Genet.* 3, 889–895. doi: 10.1038/nrg933

Steele, E. J., Lindley, R. A., and Blanden, R. V. (1998). *Lamarck's Signature: How Retrogenes are Changing Darwin's Natural Selection Paradigm*. Frontiers of Science: Series eds P. Davies (Sydney, NSW: Allen ans Unwin).

Storz, G. (2002). An expanding universe of noncoding of RNAs. *Science* 296, 1260–1263. doi: 10.1126/science.1072249

Sunkar, R., Chinnusamy, V., Zhu, J., and Zhu, J. K. (2007). Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends Plant Sci.* 12, 301–309. doi: 10.1016/j.tplants.2007.05.001

Suzuki, M. M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9, 465–476. doi: 10.1038/nrg2341

Swati, D. (2017). "Riboswitches: regulatory ncRNAs in Archaea," in *Biocommunication of Archaea*, ed G. Witzany (Berlin; Heidelberg: Springer International Publishing AG), 277–303.

Sweatt, J. D., and Tamminga, C. A. (2016). An epigenomics approach to individual differences and its translation to neuropsychiatric conditions. *Dialogues Clin. Neurosci.* 18, 289–298.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299. doi: 10.1002/bies.20544

Thoday, J. (1953). Component of fitness. *Symp. Soc. Exp. Biol.* 7:96.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., et al. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616–1625. doi: 10.1101/gr.134445.111

Tollefsbol, T. O. (2011). "Epigenetics: the new science of genetics," in *Handbook of Epigenetics*, ed T. Tollefsbol (New York, NY: Academic Press), 1–6. doi: 10.1016/B978-0-12-375709-8.00001-0

Tollefsbol, T. O. (2014). "Transgenerational epigenetics," in *Transgenerational Epigenetics, Evidence and Devate*, ed T. Tollefsbol (New York, NY: Elsevier AP), 1–8.

Van Otterdijk, S. D., and Michels, K. B. (2016). Transgenerational epigenetic inheritance in mammals: how good is the evidence? *FASEB J.* 30, 2457–2465. doi: 10.1096/fj.201500083

Varki, A., and Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res.* 15, 1746–1758. doi: 10.1101/gr.3737405

Vastenhouw, N. L., Brunschwig, K., Okihara, K. L., Müller, F., Tijsterman, M., and Plasterk, R. (2006). Gene expression: long-term gene silemcing by RNAi. *Nature* 442:882. doi: 10.1038/442882a

Vella, M. C., and Slack, F. J. (2005). "*C. elegans* microRNAs," in *WormBook*, ed T. Blumenthal (The *C. elegans* Research Community), 1–9. Available online at: http://www.wormbook.org

Verhoeven, K. J. F., Jansen, J. J., van Dijk, P. J., and Biere, A. (2010). Stress-induced DNA methylation changes and their heritability in asexual dandelions. *New Phytol.* 185, 1108–1118. doi: 10.1111/j.1469-8137.2009.03121.x

Waddington, C. H. (2012). The epigenotype. 1942. *Int. J. Epidemiol.* 41, 10–13. doi: 10.1093/ije/dyr184

Wahl, M. C., Will, C. L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell* 136, 701–718. doi: 10.1016/j.cell.2009.02.009

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Wang, W., Kwon, E. J., and Tsai, L.-H. (2012). MicroRNAs in learning, memory, and neurological diseases. *Learn. Mem.* 19, 359–368. doi: 10.1101/lm.026492.112

Watson, M., Hawkes, E., and Meyer, P. (2014). Transmission of epi-alleles with MET1-dependent dense methylation in *Arabidopsis thaliana*. *PLoS ONE* 9:e105338. doi: 10.1371/journal.pone.0105338

Weber, M. (2006). The Central Dogma as a thesis of causal specificity. *Hist. Philos. Life Sci.* 28, 595–609. Available online at: https://philarchive.org/archive/WEBTCD

Weigel, D., and Colot, V. (2012). Epialleles in plant evolution. *Genome Biol.* 13:249. doi: 10.1186/gb-2012-13-10-249

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi: 10.1101/gad.1800909

Woltereck, R. (1909). Weitere experimentelle Untersuchungen {ü}ber Artver{ä}nderung, speziel {ü}ber das Wesen quantitativer Artunterschiede bei Daphniden. *Verhandlungen der Dtsch. Zool. Gesellschaft* 19, 110–173.

Wright, M. W., and Bruford, E. (2011). Naming "junk": human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics* 5, 90–98. doi: 10.1186/1479-7364-5-2-90

Yan, W. (2014). Potential roles of noncoding RNAs in environmental epigenetic transgenerational inheritance. *Mol. Cell. Endocrinol.* 398, 24–30. doi: 10.1016/j.mce.2014.09.008

Yang, L. (2015). Splicing noncoding RNAs from the inside out. *Wiley Interdiscip. Rev. RNA* 6, 651–660. doi: 10.1002/wrna.1307.

Yekta, S., Shih, I. H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304, 594–6. doi: 10.1126/science.1113329

Yohn, N. L., Marisa, S., Bartolomei, M. S., and Blendy, J. A. (2015). Multigenerational and transgenerational inheritance of drug exposure: the effects of alcohol, opiates, cocaine, marijuana, and nicotine. *Prog. Biophys. Mol. Biol.* 118, 21–33. doi: 10.1016/j.pbiomolbio.2015.03.002

Yu, C., Xue, J., Zhu, W., Jiao, Y., Zhang, S., and Cao, J. (2014). Warburg meets non-coding RNAs: the emerging role of ncRNA in regulating the glucose metabolism of cancer cells. *Tumor Biol.* 36, 81–94. doi: 10.1007/s13277-014-2875-z

Yuan, S., Oliver, D., Andrew Schuster, A., Zheng, H., and Yan, W. (2015). Breeding scheme and maternal small RNAs affect the efficiency of transgenerational inheritance of a paramutation in mice. *Sci. Rep.* 5:9266. doi: 10.1038/srep09266

Zhao, Y., Sun, H., and Wang, H. (2016). Long noncoding RNAs in DNA methylation: new players stepping into the old game. *Cell Biosci.* 6:45. doi: 10.1186/s13578-016-0109-3

Zhou, J., Li, H., Li, S., Zaia, J., and Rossi, J. J. (2008). Novel dual inhibitory function aptamer–siRNA delivery system for HIV-1 Therapy. *Mol. Ther.* 16, 1481–1489. doi: 10.1038/mt.2008.92

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

PERSPECTIVE
published: 22 December 2017
doi: 10.3389/fmicb.2017.02625

Experimental Aspects Suggesting a "Fluxus" of Information in the Virions of Herpes Simplex Virus Populations

*Luis A. Scolaro[1], Julieta S. Roldan[1,2], Clara Theaux[1], Elsa B. Damonte[1,2] and Maria J. Carlucci[1,2]**

[1] Laboratorio de Virología, Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina, [2] Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad de Buenos Aires, Buenos Aires, Argentina

Our perspective on nature has changed throughout history and at the same time has affected directly or indirectly our perception of biological processes. In that sense, the "fluxus" of information in a viral population arises a result of a much more complex process than the encoding of a protein by a gene, but as the consequence of the interaction between all the components of the genome and its products: DNA, RNA, and proteins and its modulation by the environment. Even modest "agents of life" like viruses display an intricate way to express their information. This conclusion can be withdrawn from the huge quantity of data furnished by new and potent technologies available now to analyze viral populations. Based on this premise, evolutive processes for viruses are now interpreted as a simultaneous and coordinated phenomenon that leads to global (i.e., not gradual or 'random') remodeling of the population. Our system of study involves the modulation of herpes simplex virus populations through the selective pressure exerted by carrageenans, natural compounds that interfere with virion attachment to cells. On this line, we demonstrated that the passaging of virus in the presence of carrageenans leads to the appearance of progeny virus phenotipically different from the parental seed, particularly, the emergence of syncytial (syn) variants. This event precedes the emergence of mutations in the population which can be readily detected five passages after from the moment of the appearance of syn virus. This observation can be explained taking into consideration that the onset of phenotypic changes may be triggered by "environmental-sensitive" glycoproteins. These "environmental-sensitive" glycoproteins may act by themselves or may transmit the stimulus to "adapter" proteins, particularly, proteins of the tegument, which eventually modulate the expression of genomic products in the "virocell." The modulation of the RNA network is a common strategy of the virocell to respond to environmental changes. This "fast" adaptive mechanism is followed eventually by the appearance of mutations in the viral genome. In this paper, we interpret these findings from a philosophical and scientific point of view

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria

Reviewed by:
Pengwei Zhang,
Yale University, United States
Takashi Irie,
Hiroshima University, Japan

***Correspondence:**
Maria J. Carlucci
majoc@qb.fcen.uba.ar

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *28 July 2017*
Accepted: *15 December 2017*
Published: *22 December 2017*

Citation:
Scolaro LA, Roldan JS, Theaux C, Damonte EB and Carlucci MJ (2017) Experimental Aspects Suggesting a "Fluxus" of Information in the Virions of Herpes Simplex Virus Populations. Front. Microbiol. 8:2625. doi: 10.3389/fmicb.2017.02625

interconnecting epigenetic action, exerted by carragenans from early RNA network–DNA interaction to late DNA mutation. The complexity of HSV virion structure is an adequate platform to envision new studies on this topic that may be complemented in a near future through the analysis of the genetic dynamics of HSV populations.

Keywords: herpes simplex virus, virus–host interactions, microRNAs, non-coding RNAs, regulatory networks, epigenetic, viral population, carrageenans

BACKGROUND

Currently, the study of biological processes, not in the understanding of the process as a whole but in the fragmented analysis increasingly smaller and dazzled by the new technologies, allow us to have a large amount of data that has generated a crisis due to excessive information. Metagenomics studies are a good example of this fact. Such information is lacking in organization and meaningful understanding within the conceptual paradigms of biological phenomena and, therefore, in the interpretation of the data available to us within a general context (Sandín, 2004). A problem that has its origin in the lack of consistency of the theoretical base of biology, this means, in the explanation of the phenomena of life. As explain Oltvai and Barabási (2002) at present, it is widely accepted that DNA is not the only container of biological complexity. The genome, transcriptome, proteome, and metabolome represent distinct levels of organization at which information can be stored and processed. Also, various cellular programs reside at these levels. Thus, although the genome almost exclusively stores long-term information, the proteome is essential for storing information in the short term and the recovery of this information is controlled by transcription factors strongly influenced by the metabolome (Bray, 1995). These different levels of organization and cellular functionality constitute groups of heterogeneous components that would act all interconnected in large networks (Oltvai and Barabási, 2002). Thus, the integration of complex systems would imply that the complexity of life phenomena derives from a great initial complexity of their constituent units (i.e., not only key agents of DNA replication, etc.) and that the properties of the systems that make up life (cells, organs, organisms, ecosystems) are a consequence of the properties of its components (on the other hand, with extremely conserved processes). Populations of viruses are also modeled in this way by processes that take place in the "virocell," an infected cell whose aim is to produce virions. In this line, viruses also contribute to the diversity of processes within the virocell providing new information that might eventually become part of the cell genome (Forterre, 2010). In this respect, non-coding RNAs may represent a suitable target for viral modulation in view of their viral origin and the variety of cellular processes they control (Witzany, 2009). In order to analyze a process of viral population variability influenced by the environment we worked on a system consisting of herpes simplex virus (HSV) and cell cultures in an environment containing sulfated polysaccharides known as carrageenans (CGNs). Cell heparan sulfate-like chemical structures in the CGNs are known to be very active and selective compounds against HSV (Carlucci et al., 1999). Their mechanism of action mainly affects viral adsorption stage, interacting with the surface glycoproteins, thus blocking interaction with cell receptors. Multiplication of HSV in the presence of CGN leads to the emergence of syn variants with phenotypic characteristics quite different from parental virus (Mateu et al., 2011; Artuso et al., 2016).

EXPERIMENTAL MODEL

Isolation of viral variants was performed after successive passages of HSV in Vero cells subjected to increasing doses of CGN. For this purpose we monitored the changes of viral yield for each passage with or without CGN, using virus passages in the presence of acyclovir (ACV), the antiviral currently in use for herpetic infections, as controls. Also, the resistance pattern generated by the CGN, compared with ACV, has been evaluated. For CGN and ACV the initial doses were below the inhibitory concentration 50% (IC_{50}) and were increased slightly in next passages (Carlucci et al., 2002). Titers of virus without CGN ranged from 10^7 to 5×10^8 PFU/ml throughout the 20 passages analyzed. Titers of virus in the presence of the polysaccharide was similar to the untreated control except for passages #1, 8, 14, and 19 where a 1.5 to 2 log drop in virus titer was detected. In the case of ACV titers were similar to untreated controls during the six passages analyzed after which recovered virus proved to be resistant to the antiviral. IC_{50} increased very rapidly in the first passages and the relative resistance also increased significantly from passage #4 onward, with a value of 46.6 μg/ml reaching 60.0 μg/ml in passage #6. In accordance to previous reports, selection of resistant virus to ACV was detected after few passages in the presence of the drug (Mateu et al., 2017). In the case of CGN, from passage 11 onward, the traditional type of cytopathic effect (CPE) of HSV, characterized by cell rounding and clumping that appeared as small focuses on the monolayer and eventually spread over the entire culture changed to the appearance of multinucleated cells (syncytia) due to the fusion of adjacent infected cells (**Figure 1**). Also, in this passage, a marked change was detected in the size of the viral plaques, coexisting small viral plaques (1 mm diameter) (similar to the parental strain) and large viral plaques of 1.5–2 mm diameter, until passage 16, when only large plaques could be observed. The augment in plaque size precedes the formation of syncytium. These changes in CPE were not detected after sequential passaging of HSV in the absence of CGN. IC_{50} also showed to be variable with a relative resistance (RR is the ratio between IC_{50} for each syn variant and IC_{50} for the F parental strain) between 1.5 and 6.6 for passages with CGN, while for viral controls without CGN the RR ranged between 1.6 and 3.4 (**Table 1**) (Carlucci et al., 2002). From passages 11

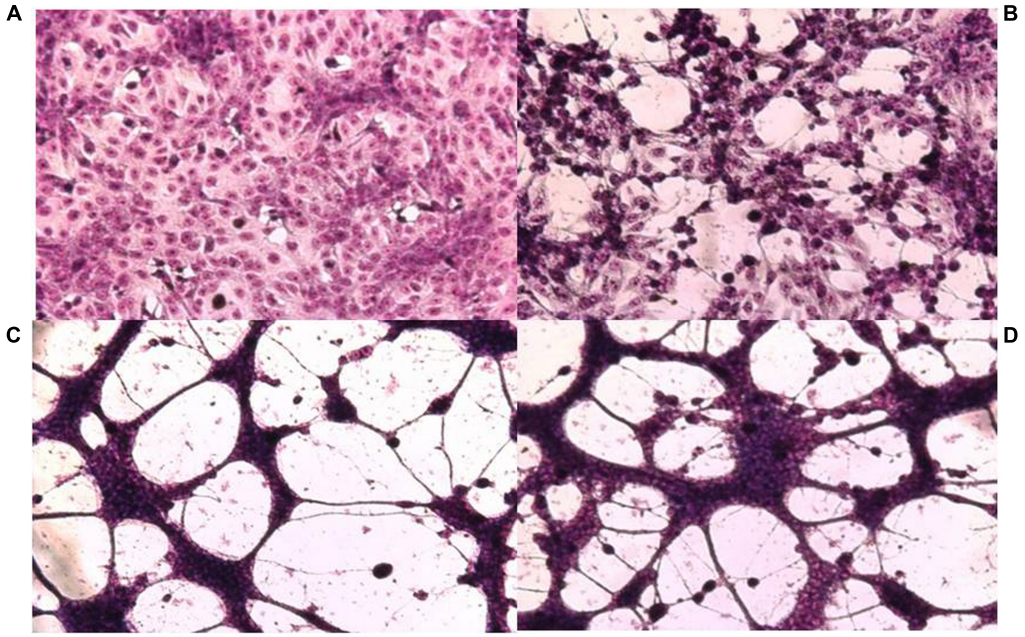


FIGURE 1 | Cytopathic effect on Vero cells of HSV-1 F and its syncytial variants 48 h post-infection. **(A)** Uninfected control cells, **(B)** parental strain F, **(C)** passage 14, **(D)** passage 17. Cell monolayers were infected with the different viral variants at m.o.i: 0.1. The cells were fixed with methanol and stained with Giemsa 48 h p.i. (x100).

TABLE 1 | Characteristics of viral cytophatic effect and drug susceptibility arising after different passages with CGN.

N° passage	IC_{50} (RR[a])	Non syn (%)	Syn (%)
0	1.1 (1.0)	100	0
11	1.6 (1.5)	58.8	41.2 (rev.)*
12	4.3 (3.9)	89.2	10.8 (rev.)
13	4.1 (3.7)	72.2	27.3 (rev.)
14	6.8 (6.2)	70.0	30.0 (rev.)
15	7.3 (6.6)	71.7	28.3 (rev.)
16	3.1 (2.8)	0	100 (irrev.)

**rev.: reversible, irrev.: irreversible. a = Relative Resistance.*

to 15 phenotypic changes (RR and CPE) in virus collected from supernatants showed a marked variability and, when passaged in the absence of CGN, recovered the phenotype of parental virus.

From passage 11 onward, the modification of the CPE accompanied by the increase of RR observed in the next passages suggest a successful adjustment of the viral population to the environment (CGN). It is tempting to speculate that treatment with CGN increases the "fluxus" of information in the viral population leading to the onset of a "temporary memory" as a useful tool for the virus to cope with environmental changes.

CARRAGEENAN AND HSV

Evidence published in Meckes and Wills (2008) showed that like viruses can modulate cell signaling pathways when their receptors bind to the plasma membrane (Marsh and Helenius, 2006), also the signals can be transmitted in reverse through the envelope into any virus after receptor binding (Rein et al., 1994; Aguilar et al., 2003; Murakami et al., 2004; Wyma et al., 2004; Meckes and Wills, 2008). Interaction of HSV glycoproteins with its initial cell attachment protein (i.e., heparin or CGNs, a surrogate for heparan sulfate) triggers a rapid and highly efficient change in the structure of proteins of the tegument, a region between the viral membrane and the DNA-containing capsid. This phenomenum, has been described during studies of UL16. This protein associates with cytoplasmic capsids while on the other hand, interacts with a membrane-bound tegument protein, UL11.

The initial binding of HSV occurs through the interaction of glycoprotein gC and the cell receptor. As gC has a short cytoplasmic tail preventing signal transmission within the virion, probably, the interaction with other glycoprotein such as gB and gD may be necessary. On this respect, Cocchi et al. (2004) proposed a tripartite structure for the complex formed by gD together with gB and gH-L and its receptor and, one or various of the fusion glycoproteins. This complex would play an important role in recruiting/activating the fusion glycoproteins, activating them and promoting fusion of the viral envelope with cell membrane (Cocchi et al., 2004). But if this activation/deactivation of the glycoproteins are not coordinated with an eventual cellular fusion, as would be the case of the interaction CGN-virus, a "temporary memory" may be generated in the viral population. This memory would be crucial, particularly for DNA viruses, whose genomes are not prone to accumulate mutations in a fast manner as may be the case for RNA viruses, and may account for

the RR and syn phenotype of virus collected during passages #11 to 16.

In view of the facts commented above, we may hypothesize that the binding of HSV through the glycoproteins gC, gB, gD to the CGN, leads to a structural change in the tegument proteins UL11 and UL16. Both proteins would interact with the viral capsid modulating the expression of immediate early genes (IE) (alpha) (Meckes and Wills, 2008; Svobodova et al., 2011) and, in consequence also modulating the remaining genetic blocks of herpes (beta and gamma).

Another point to address related to proteins of the viral tegument is linked to the fact that the variants manifested their syn CPE at a shorter time (16 h p.i.) than the parental virus (24 h p.i.). This observation may be related to the organization of the microtubules. Stable microtubules (MT) formation would be reduced in cells infected with syn variants by the viral Ser/Thr kinase, Us3 (Purves et al., 1987; Ryckman and Roller, 2004). Many viruses are dependent on MTs for their intracellular movement. During the early steps of infection, HSV is able to disrupt the centrosome, impairing MT organization. On the other hand, as infection goes further, HSV-1 induces the formation of stable structures formed stable MT subsets through inactivation of glycogen synthase kinase 3 beta by the viral Ser/Thr kinase, Us3 (Naghavi et al., 2013).

RNA network, particularly microRNAs (miRNAs), would be also modulated by the tegument proteins. miRNAs are small non-coding RNAs that interact with highly conserved proteins and are important in rapid gene regulation. miRNAs encoded by viruses exploit RNA silencing for regulation of their own genes, host genes, or both (Sullivan and Ganem, 2005). Also, viral miRNAs modulate biological processes of paramount importance: latent and lytic infection, evasion from the immune system, modulation of apoptosis, synthesis of viral macromolecules, etc. (Sullivan and Ganem, 2005; Boss and Renne, 2010). miR-H6 affects negatively the expression of ICP4. This protein is necessary for an efficient transcription of viral genes and regulates the onset of the characteristic CPE of HSV (Taylor et al., 2002; Umbach et al., 2008; Duan et al., 2012). On this line, miR-H2 targets ICP0 protein, an IE gene that has a major role in lytic infection and entrance of HSV into cells (Piedade and Azevedo-Pereira, 2016). On the other hand, miR-92944 is involved in the growth of virus and variants lacking miR92944 exhibited significant reductions in viral titers and fourfold reduction in plaque size (Munson and Burch, 2012). miR-23a and miR-146a are miRNAs of cellular origin that are involved in HSV replication because they interfere with the innate immune response diminishing the levels of interferon and activating pro-inflamatory cytokines (Ru et al., 2014). On the other side, HSV-1 induces the pro-inflammatory miR-146a. This molecule targets complement factor H and induces key elements of the arachidonic acid cascade (Hill et al., 2009). Also, it is an NF-κB-dependent gene, which in turn actively participates in the onset of the innate immune system (Taganov et al., 2006; Baltimore et al., 2008; Hill et al., 2009). Although these miRNAs are of cellular origin it cannot be ruled out that they are incorporated within the viral structure, providing the virus with valuable information for the next multiplication cycle in the presence of CGN.

In view of the facts exposed above, we hypothesize that the appearance of the syn variants during the early passages in the presence of CGN might be a consequence of an alteration of the tegument proteins which in turn modulate microtubules physiology and functioning of the RNA network in the virocell.

CONCLUSION AND PERSPECTIVES

The first inkling that herpesviruses modify cellular membranes was based on the observations that mutants differ wild type strains with respect to their effects on cells (Ejercito et al., 1968). These observations led the prediction that herpesviruses alter the structure and antigenicity of cellular membranes, a prediction fulfilled by (a) the demonstration of alterated structure and antigenic specificity and (b) the presence of viral glycoproteins in the cytoplasmic and plasma membranes of infected cells (Roizman and Sears, 1991). It's know that the presence of gD in the plasma membrane of infected cells precludes reinfection of cells with the progeny virus released from that cell (Campadelli-Fiume et al., 1988). In our system the CGNs would act to interfere with viral glycoproteins of both virus and those exposed at the level of the cell membrane. The first phenotypic effect observed by the constant action of CGN with the virus is the increase in plaque size and the subsequent appearance of syn effects and variability in RR. We can assume that these glycoproteins could perceive the presence of the CGN in the environment and transmit this information both, inside the virus and the cell. In this sense, it has been shown the CGNs do not possess virucidal activity and have no action by pretreatment of the infected cell and do not penetrate into the cells (Carlucci et al., 1997, 1999, 2002; Yermak et al., 2012).

Likewise, herpesviruses are examples of dynamic and complex systems based on the interactions of multiple cellular and viral factors, leading to lifelong viral infections. These interactions control the expression of cellular proteins that may modulate the infection. In this work we modified the networks of target transcripts in the virocell by action of the CGN and verified a rapid viral adaptation to the presence of the polysaccharide before the manifestation of genetic modifications. We hypothesized that glycoproteins would fulfill, in addition to the functions already known, a fundamental function primarily as antennas of environmental perception. Host miRNA modulates viral infections by influencing antiviral responses, promoting several phases of the viral life cycle, or participating in cellular tropism. Also, as cellular miRNAs participate in multiple processes, their sequestration by the virus may cause a desregulation in the expresion of different cellular mRNAs, which might eventually lead to an aberrant process of protein translation. It is believed that one of the multiple parameters that cooperate to viral adaptation arises as a consequence of altered host miRNA–mRNA interactions, thus favoring the cellular environment for viral persistence or chronicity (Bruscella et al., 2017). Because viral miRNAs generally have a surprising lack of evolutionary conservation, it could be hypothesized that they are sites of rapid evolution, even as a driver of speciation (Kincaid and Sullivan, 2012). In agreement with Li et al. (2014), . . . "*viral RNAs could act*

as sponges that can sequester endogenous miRNAs within infected cells and thus impact the stability and translational efficiency of host mRNAs with shared miRNA response elements"...(Li et al., 2014). Also, the use of miRNA as elements of shared responses between viral RNAs and host mRNA form complex networks during infection which affect replication, pathogenesis, and viral persistence. In this way the field of action of RNAs and viral mRNA would not only be limited to the level of viral protein synthesis or as PAMPs in innate immunity but would have multiple ways of working (Li et al., 2014). Finally, an important feature not to forget is that viral populations are plastic and in constant change. On this line, we are analyzing the virus recovered during the different passages with CGNs by High Throughput Sequencing in order to determine the relative abundance of viruses that exhibit differences with the parental strain at the genomic level.

We live neither in an arbitrary world of pure chance nor in a deterministic world without novelty and creativity. Life and Nature interplay in a never-ending process of evolution. Nature and humanity are interwoven creatively in this process, recognizing the "sensitive intelligence" of the viral population with the environment would be part of our learning.

AUTHOR CONTRIBUTIONS

All authors contributed to planning, writing, and revision of the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

This study was supported by the Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET) PIP 201 00338. JR is postdoctoral fellow from CONICET. ED, MC, and LS are members of the Research Career from CONICET.

REFERENCES

Aguilar, H. C., Anderson, W. F., and Cannon, P. M. (2003). Cytoplasmic tail of Moloney murine leukemia virus envelope protein influences the conformation of the extracellular domain: implications for mechanism of action of the R peptide. *J. Virol.* 77, 1281–1291. doi: 10.1128/JVI.77.2.1281-1291.2003

Artuso, M. C., Roldán, J. S., Scolaro, L. A., and Carlucci, M. J. (2016). Viruses: as mediators in "elan vital" ' of the "creative" evolution. *Infect. Genet. Evol.* 46, 78–84. doi: 10.1016/j.meegid.2016.10.028

Baltimore, D., Boldin, M. P., O'Connell, R. M., Rao, D. S., and Taganov, K. D. (2008). MicroRNAs: new regulators of immune cell development and function. *Nat. Immunol.* 9, 839–845. doi: 10.1038/ni.f.209

Boss, I. W., and Renne, R. (2010). Viral miRNAs: tools for immune evasion. *Curr. Opin. Microbiol.* 13, 540–545. doi: 10.1016/j.mib.2010.05.017

Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature* 376, 307–312. doi: 10.1038/376307a0

Bruscella, P., Bottini, S., Baudesson, C., Pawlotsky, J. M., Feray, C., and Trabucchi, M. (2017). Viruses and miRNAs: more friends than foes. *Front. Microbiol.* 8:824. doi: 10.3389/fmicb.2017.00824

Campadelli-Fiume, G., Arsenakis, M., Farafegoli, F., and Roizman, B. (1988). Entry of herpes simplex virus 1 in BJ cells that constitutively express viral glycoprotein D is by endocytosis and results in the degradation of the virus. *J. Virol.* 62, 159–167.

Carlucci, M. J., Ciencia, M., Matulewicz, M. C., Cerezo, A. S., and Damonte, E. B. (1999). Antiherpetic activity and mode of action of natural carrageenans of diverse structural types. *Antiv. Res.* 43, 93–102. doi: 10.1016/S0166-3542(99)00038-8

Carlucci, M. J., Scolaro, L. A., and Damonte, E. B. (2002). Herpes simples virus type 1 variants arising after selection with an antiviral carrageenan: lack of correlation between drug susceptibility and syn phenotype. *J. Med. Virol.* 68, 92–98. doi: 10.1002/jmv.10174

Carlucci, M. J., Scolaro, L. A., Matulewicz, M. C., and Damonte, E. B. (1997). Antiviral activity of natural sulphated galactans on herpes virus multiplication in cell culture. *Planta Med.* 63, 429–432. doi: 10.1055/s-2006-957727

Cocchi, F., Fusco, D., Menotti, L., Gianni, T., Eisenberg, R. J., Cohen, G. H., et al. (2004). The soluble ectodomain of herpes simplex virus gD contains a membrane-proximal pro-fusion domain and suffices to mediate virus entry. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7445–7450. doi: 10.1073/pnas.0401883101

Duan, F., Liao, J., Huang, Q., Nie, Y., and Wu, K. (2012). HSV-1 miR-H6 inhibits HSV-1 replication and IL-6 expression in human corneal epithelial cells in vitro. *Clin. Dev. Immunol.* 2010:192791. doi: 10.1155/2012/192791

Ejercito, P. M., Kieff, E. D., and Roizman, B. (1968). Characterization of herpes simplex virus strains differing in their effect on social behavior on infected cells. *J. Gen. Virol.* 2, 357–364. doi: 10.1099/0022-1317-2-3-357

Forterre, P. (2010). Manipulation of cellular syntheses and the nature of viruses: the virocell concept. *C. R. Chim.* 14, 392–399. doi: 10.1016/j.crci.2010.06.007

Hill, J. M., Zhao, Y., Clement, C., Neumann, D. M., and Lukiw, W. J. (2009). HSV-1 infection of human brain cells induces miRNA-146a and Alzheimer-type inflammatory signaling. *Neuroreport* 20, 1500–1505. doi: 10.1097/WNR.0b013e3283329c05

Kincaid, R. P., and Sullivan, C. S. (2012). Virus-encoded microRNAs: an overview and a look to the future. *PLOS Pathog.* 8:e1003018. doi: 10.1371/journal.ppat.1003018

Li, C., Hu, J., Hao, J., Zhao, B., Wu, B., Sun, L., et al. (2014). Competitive virus and host RNAs: the interplay of a hidden virus and host interaction. *Prot. Cell* 5, 348–356. doi: 10.1007/s13238-014-0039-y

Marsh, M., and Helenius, A. (2006). Virus entry: open sesame. *Cell* 124, 729–740. doi: 10.1016/j.cell.2006.02.007

Mateu, C., Perez Recalde, M., Artuso, C., Hermida, G., Linero, F., Scolaro, L., et al. (2011). Emergence of HSV-1 syncytial variants with altered virulence for mice after selection with a natural carrageenan. *Sex Transm. Dis.* 38, 555–561.

Mateu, C. G., Artuso, M. C., Pujol, C. A., Linero, F. N., Scolaro, L. A., and Carlucci, M. J. (2017). *In vitro* isolation of variant of herpes simplex virus attenuated with altered thymidine kinase and DNA polymerase genes using carrageenans as selection agents. *Symbiosis* 72, 23. doi: 10.1007/s13199-016-0437-4

Meckes, D. G., and Wills, J. W. (2008). Structural rearrangement within an enveloped virus upon binding to the host cell. *J. Virol.* 82, 10429–10435. doi: 10.1128/JVI.01223-08

Munson, D. J., and Burch, A. D. (2012). A novel miRNA produced during lytic HSV-1 infection is important for efficient replication in tissue culture. *Arch. Virol.* 157, 1677–1688. doi: 10.1007/s00705-012-1345-4

Murakami, T., Ablan, S., Freed, E. O., and Tanaka, Y. (2004). Regulation of human immunodeficiency virus type 1 Env-mediated membrane fusion by viral protease activity. *J. Virol.* 78, 1026–1031. doi: 10.1128/JVI.78.2.1026-1031.2004

Naghavi, M. H., Gundersenb, G. G., and Walsh, D. (2013). Plus-end tracking proteins, CLASPs, and a viral Akt mimic regulate herpesvirus-induced stable microtubule formation and virus spread. *Proc. Natl. Acad. Sci. U.S.A.* 110, 18268–18273. doi: 10.1073/pnas.1310760110

Oltvai, Z. N., and Barabási, A. L. (2002). Life's complexity pyramid. *Science* 298, 763–764. doi: 10.1126/science.1078563

Piedade, D., and Azevedo-Pereira, J. M. (2016). The role of microRNAs in the pathogenesis of herpesvirus infection. *Viruses* 156, 1–32. doi: 10.3390/v8060156

Purves, F. C., Longnecker, R. M., Leader, D. P., and Roizman, B. (1987). Herpes simplex virus 1 protein kinase is encoded by open reading frame US3 which is not essential for virus growth in cell culture. *J. Virol.* 61, 2896–2901.

Rein, A. J., Mirro, J. G., Haynes, S. M., Ernst, S. M., and Nagashima, K. (1994). Function of the cytoplasmic domain of a retroviral transmembrane protein: p15E-p2E cleavage activates the membrane fusion capability of the murine leukemia virus Env protein. *J. Virol.* 68, 1773–1781.

Roizman, B., and Sears, A. E. (1991). "Herpes simplex viruses and their replication," in *Fundamental Virology*, 2nd Edn, eds B. N. Fields and D. M. Knipe (New York, NY: Raven Press, Ltd).

Ru, J., Sun, H., Fan, H., Wang, C., Li, Y., Liu, M., et al. (2014). MiR-23a facilitates the replication of HSV-1 through the suppression of interferon regulatory factor 1. *PLOS ONE* 9:e114021. doi: 10.1371/journal.pone.0114021

Ryckman, B. J., and Roller, R. J. (2004). Herpes simplex virus type 1 primary envelopment: UL34 protein modification and the US3-UL34 catalytic relationship. *J. Virol.* 78, 399–412. doi: 10.1128/JVI.78.1.399-412.2004

Sandín, M. (2004). *Sucesos Excepcionales de la Evolución*. Available at: http://www.redcientifica.com/doc/doc200311130001.html

Sullivan, C. S., and Ganem, D. (2005). MicroRNAs and viral infection. *Mol. Cell.* 20, 3–7. doi: 10.1016/j.molcel.2005.09.012

Svobodova, S., Bell, S., and Crump, C. (2011). Analysis of the interaction between the essential herpes simples virus 1 tegument proteins VP16 and VP1/2. *J. Virol.* 86, 473–483. doi: 10.1128/JVI.05981-11

Taganov, K. D., Boldin, M. P., Chang, K. J., and Baltimore, D. (2006). NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12481–12486. doi: 10.1073/pnas.0605298103

Taylor, T. J., Brockman, M. A., McNamee, E., and Knipe, D. M. (2002). Herpes simplex virus. *Front. Biosci.* 7:752–764. doi: 10.2741/taylor

Umbach, J. L., Kramer, M. F., Jurak, I., Karnowski, H. W., Coen, D. M., and Cullen, B. R. (2008). MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature* 454, 780–783. doi: 10.1038/nature07103

Witzany, G. (2009). Noncoding RNAs: persistent viral agents as modular tools for cellular needs. *Ann. N. Y. Acad.. Sci.* 1178, 244–267. doi: 10.1111/j.1749-6632.2009.04989.x

Wyma, D. J., Jiang, J., Shi, J., Zhou, J., Lineberger, J. E., Miller, M. D., et al. (2004). Coupling of human immunodeficiency virus type 1 fusion to virion maturation: a novel role of the gp41 cytoplasmic tail. *J. Virol.* 78, 3429–3435. doi: 10.1128/JVI.78.7.3429-3429-3435.2004

Yermak, I. M., Barabanova, A. O., Aminin, D. L., Davydova, V. N., Sokolova, E. V., Soloveva, T. F., et al. (2012). Effects of structural peculiarities of carrageenans on their immunomodulatory and anticoagulant activities. *Carbohydr. Polym.* 87, 713–720. doi: 10.1016/J.carbpol.2011.08.053

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 15 November 2017
doi: 10.3389/fmicb.2017.02252

A Multilayered Control of the Human *Survival Motor Neuron* Gene Expression by Alu Elements

*Eric W. Ottesen, Joonbae Seo, Natalia N. Singh and Ravindra N. Singh**

Department of Biomedical Sciences, Iowa State University, Ames, IA, United States

Humans carry two nearly identical copies of *Survival Motor Neuron* gene: *SMN1* and *SMN2.* Mutations or deletions of *SMN1,* which codes for SMN, cause spinal muscular atrophy (SMA), a leading genetic disease associated with infant mortality. Aberrant expression or localization of SMN has been also implicated in other pathological conditions, including male infertility, inclusion body myositis, amyotrophic lateral sclerosis and osteoarthritis. *SMN2* fails to compensate for the loss of *SMN1* due to skipping of exon 7, leading to the production of SMNΔ7, an unstable protein. In addition, SMNΔ7 is less functional due to the lack of a critical C-terminus of the full-length SMN, a multifunctional protein. Alu elements are specific to primates and are generally found within protein coding genes. About 41% of the human *SMN* gene including promoter region is occupied by more than 60 Alu-like sequences. Here we discuss how such an abundance of Alu-like sequences may contribute toward SMA pathogenesis. We describe the likely impact of Alu elements on expression of SMN. We have recently identified a novel exon 6B, created by exonization of an Alu-element located within *SMN* intron 6. Irrespective of the exon 7 inclusion or skipping, transcripts harboring exon 6B code for the same SMN6B protein that has altered C-terminus compared to the full-length SMN. We have demonstrated that SMN6B is more stable than SMNΔ7 and likely functions similarly to the full-length SMN. We discuss the possible mechanism(s) of regulation of *SMN* exon 6B splicing and potential consequences of the generation of exon 6B-containing transcripts.

Keywords: spinal muscular atrophy, SMA, survival motor neuron, SMN, SMN6B, Alu, exonization, transposable elements

OPEN ACCESS

Edited by:
Guenther Witzany, Telos-Philosophische Praxis, Austria

Reviewed by:
Emanuele Buratti, International Centre for Genetic Engineering and Biotechnology, Italy
Girish C. Shukla, Cleveland State University, United States

****Correspondence:***
Ravindra N. Singh
singhr@iastate.edu

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *27 September 2017*
Accepted: *31 October 2017*
Published: *15 November 2017*

Citation:
Ottesen EW, Seo J, Singh NN and Singh RN (2017) A Multilayered Control of the Human Survival Motor Neuron Gene Expression by Alu Elements. Front. Microbiol. 8:2252. doi: 10.3389/fmicb.2017.02252

INTRODUCTION

Transposable elements (TEs) including long and short interspersed elements (LINES and SINES) occupy ∼45% of human genome (Lander et al., 2001; Smit et al., 2015). The primate-specific Alu elements are the most abundant SINES totaling >1 million copies and accounting for ∼11% of the human genome (Lander et al., 2001; Hedges and Batzer, 2005; Deininger, 2011). Alu elements are ∼300 bp bipartite motifs derived from the 7SL RNA, which is one of the components of the protein signal recognition complex (Deininger et al., 2003). The spread of Alu elements started with the radiation of primates ∼65 million years ago (Mya) and peaked ∼40 Mya. Alu elements

are broadly classified into three subfamilies J, S, and Y, with S and Y being the youngest and the only active subfamilies (Deininger, 2011). Insertion of Alu elements has played a significant role in primate evolution due to their drastic effect on chromatin remodeling and transcription, and the generation of novel exons (Gu et al., 2009; Antonaki et al., 2011; Cui et al., 2011; Su et al., 2014; Attig et al., 2016; Bouttier et al., 2016). TEs, including Alu elements, promote non-allelic homologous recombination (NAHR) and have caused and continue to contribute toward genomic instability (White et al., 2015; Wang et al., 2017). A recent genome-wide association study (GWAS) links Alu insertions to the high risk for many human diseases (Payer et al., 2017).

Alu-derived sequences affect various posttranscriptional steps including pre-mRNA splicing, mRNA stability, and translation (Lev-Maor et al., 2003; Aktaş et al., 2017; Elbarbary and Maquat, 2017). As per one estimate, ∼5% of alternative exons in humans are derived from Alu-like sequences (Sorek et al., 2002). Given the fact that transcripts carrying Alu exons harboring premature termination codon may skip detection due to nonsense-mediated decay (NMD) (Attig et al., 2016), this number could be an underestimation. Insertion of Alu-derived exons is generally suppressed by hnRNP C, which blocks recognition of the 3′ splice site (3′ss) by competing with the splicing factor U2AF65 (Zarnack et al., 2013). Inverted Alu repeats facilitate production of circular RNAs (circRNAs) due to their ability to loop-out sequences via stable double-stranded RNA structures (Liang and Wilusz, 2014; Wilusz, 2015). Depletion of DHX9, an RNA helicase that resolves the double-stranded RNA structures, was recently shown to enhance Alu-induced RNA processing defects including aberrant pre-mRNA splicing and circRNA production from transcripts harboring Alu repeats (Aktaş et al., 2017). In some instances, the stability of the Alu-derived transcripts is regulated by Adenosine Deaminase Acting on RNAs (ADARs) and the unmodified Alu-containing transcripts are degraded by Staufen-mediated RNA decay (SMD) (Elbarbary and Maquat, 2017). Consistently, a recent report suggests that the accelerated nuclear export of ADARs under stress-associated conditions leads to an enhanced stabilization of critical mRNAs harboring Alu repeats (Sakurai et al., 2017).

Chromosome 5 is one of the largest human chromosomes and harbors at least ten clusters of intrachromosomal repeats (Schmutz et al., 2004). One such intrachromosomal repeat at the 5q13.3 locus resulted in the generation of two nearly identical copies of *Survival Motor Neuron* gene: *SMN1* and *SMN2* (Lefebvre et al., 1995; Schmutz et al., 2004). Other duplicate genes at this locus include *SERF1*, *NAIP* (*BIRC1*), and *LOC647859* (*psi.OCLN*) (Schmutz et al., 2004; **Figure 1A**). Alu elements occupy ∼28% of the sequence at the 5q13.3 locus and account for a whopping 39% of the sequence in the *SMN* genes (**Figure 1A**). However, very limited attention has been paid toward understanding the consequences of such a high abundance of Alu elements in the *SMN* genes. Both *SMN* genes contain nine exons and code for SMN, an essential protein involved in various processes including snRNP biogenesis, transcription, translation, selenoprotein synthesis, stress granule formation, signal recognition particle biogenesis, signal transduction, vesicular transport, and motor neuron trafficking (Singh et al., 2017c). While the coding region of *SMN* is conserved between human and rodents, there are substantial differences in the promoter, intronic, the 5′ and 3′ untranslated regions (UTRs) primarily due to insertion of TEs including Alu-like sequences (**Figure 1B**). *SMN1* and *SMN2* differ in how the last coding exon, exon 7, is spliced (Singh, 2007; Singh and Singh, 2011; Singh et al., 2015a, 2017c). In the case of *SMN1*, all nine exons are included to produce the full-length transcript coding for the full-length SMN. In the case of *SMN2*, the majority of transcripts lack exon 7 due to a critical C-to-T mutation at the 6th position (C6U) of exon 7 (Lorson et al., 1999; Monani et al., 1999a). Transcripts lacking exon 7 code for SMNΔ7, an unstable protein, which is only partially functional (Lorson et al., 1998; Cho and Dreyfuss, 2010). Loss of *SMN1* leads to an SMN deficit, resulting in spinal muscular atrophy (SMA), a devastating genetic disease of children and infants (Ahmad et al., 2016). Among various options for SMA therapy, correction of *SMN2* exon 7 splicing has shown high promise (Seo et al., 2013; Howell et al., 2014). The recently approved drug Spinraza™ (nusinersen) for SMA is an antisense oligonucleotide (ASO) that fully corrects *SMN2* exon 7 splicing upon sequestering intronic splicing silencer N1 (ISS-N1) located within intron 7 (**Figure 1B**; Singh et al., 2006, 2017b; Ottesen, 2017). Small ASOs targeting a GC-rich sequence overlapping ISS-N1 also promote *SMN2* exon 7 inclusion and provide therapeutic benefits in mouse models of SMA (Singh et al., 2009, 2015b; Sivanesan et al., 2013; Kiel et al., 2014). We have recently shown that ISS-N1 sequesters a cryptic 5′ss, activation of which carries therapeutic implications for patients who cannot be treated by Spinraza™ or any other ASO targeting ISS-N1 (Singh et al., 2017a). In addition to SMA, SMN has been found to play an important role in male reproductive organ development and male fertility in mammals (Ottesen et al., 2016). Aberrant expression and/or localization of SMN have also been associated with other human diseases, including amyotrophic lateral sclerosis, inclusion body myositis and osteoarthritis (Singh et al., 2017c). Considering that Alu elements affect multiple steps of gene expression, understanding their potential role in the regulation of expression of the disease-linked *SMN* gene has broad implications.

Given the high abundance of Alu elements within the introns of both *SMN* genes, one would expect exonization of one or several of Alu elements. However, until recently, an exonized Alu element of *SMN* evaded detection due in part to the lack of an appropriate assay. We optimized a multi-exon-skipping-detection assay (MESDA) that determines the relative abundance of all *SMN* splice-isoforms in a single reaction (Singh et al., 2012; Seo et al., 2016a,b). Using MESDA, we detected a novel exon, exon 6B, generated by exonization of an Alu element within *SMN* intron 6 (Seo et al., 2016a). Others have independently validated/identified the exon 6B-containing transcripts in various human tissues and cell lines (Yoshimoto et al., 2016; Sutherland et al., 2017). In this brief review, we describe the likely impact of Alu elements on expression of SMN. We also discuss the possible mechanism(s) of regulation of *SMN* exon 6B splicing and potential consequences of the generation of the exon 6B-containing transcripts.

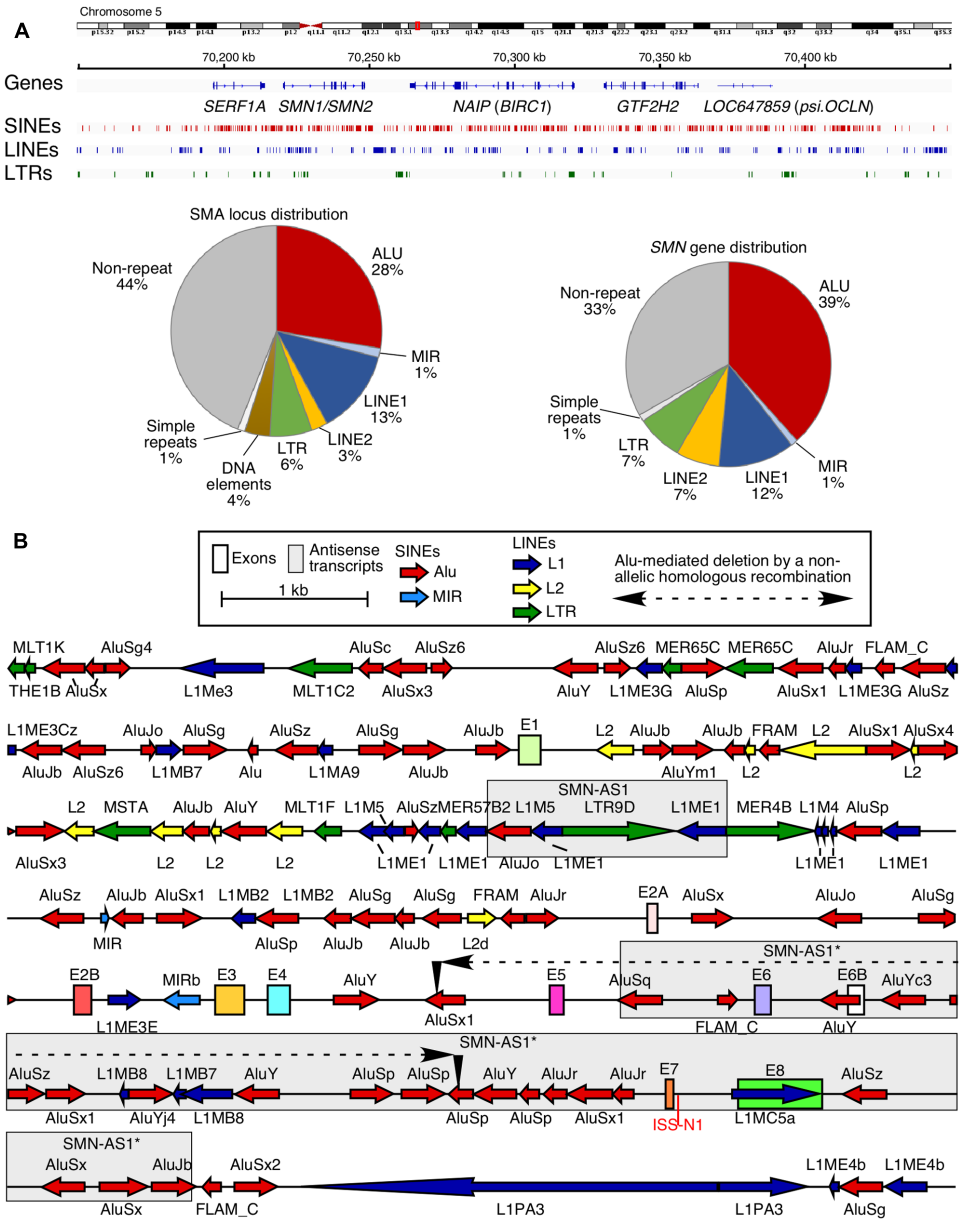


FIGURE 1 | High prevalence of Alu-derived repeats in *SMN* locus. **(A)** Genomic overview of the duplicated genomic region in chromosome 5 encompassing the *SMN* genes. Upper panel indicates the location of genes and the three most prevalent types of repeats: SINEs, LINEs, and long terminal repeats (LTRs). Lower panel pie charts indicate the total percentage of sequence occupied by different types of repeats in either the whole SMN locus including other duplicated genes (left) or in the *SMN* gene itself (right). **(B)** Detailed view of the *SMN* gene and nearby surrounding sequences. *SMN* exons are indicated with colored boxes. Repeat sequences are indicated as colored arrows, with the direction of the arrow indicating the orientation of the repeat-derived sequence. An Alu-mediated recombination event which resulted in a deletion in an SMA patient (Wirth et al., 1999) is indicated. Two boxed regions indicate the location of known antisense transcripts derived from the *SMN* locus (d'Ydewalle et al., 2017; Woo et al., 2017).

ALU ELEMENTS AND PATHOGENESIS OF SMA

Transposable elements, including Alu elements, occupy more than 65% of the human *SMN* (*SMN1* or *SMN2*) gene that spans ∼44 kb sequence including a ∼10 kb promoter region (**Figure 1**). Such a dense distribution of intrachromosomal repetitive Alu elements is often associated with NAHR to repair double strand breaks. Alu elements also serve as hotspots for non-homologous end joining (NHEJ)-based DNA repairs. Both NAHR and NHEJ that involve intrachromosomal Alu repeats potentially result in deletion or duplication of sequences ranging in size from 300 bases to tens of kilobases (Sen et al., 2006). A vast majority of SMA cases arise from deletion of a short genomic sequence

encompassing exons 7 and 8 of *SMN1* (Lefebvre et al., 1995). Although the information of the exact breakpoints of these deletions is not publicly available, they appear to include the Alu-rich intron 6 and the Alu-rich intergenic region downstream of exon 8. Interestingly, an AluSx1 and an AluSz are located immediately upstream of exon 7 and downstream of exon 8, respectively. These two Alu elements are known to be involved in the deletion in *MLL* gene associated with leukemia and cell-based experiments confirm that both NAHR- and NHEJ-based DNA repair mechanisms are the potential mechanisms of DNA deletion (Morales et al., 2015). Hence, it is likely that the pathogenic deletion of exons 7 and 8 of *SMN1* also happens through both mechanisms. Other SMA cases involve Alu/Alu-mediated deletion of sequences from intron 4 through intron 6 (**Figure 1B**; Wirth et al., 1999). The breakpoint of this Alu/Alu-mediated deletion occurred in the first 100 bp of the Alu elements. Such breakpoints are common characteristics of NAHR-mediated deletion as recently confirmed by a novel cell-based reporter system (Morales et al., 2015). While most SMA cases arise from inheritance from unaffected carriers, ∼2% of patients acquire *de novo* mutations (Wirth et al., 1997). It is likely that Alu elements have contributed toward *de novo* mutations in *SMN1* through NAHR- and/or NHEJ-based repair mechanisms in germline or in progenitors of germline cells.

ALU ELEMENTS AND *SMN* TRANSCRIPTION

Alu elements have drastically impacted the epigenetic landscape of the human genome and have consequently contributed toward the unique regulation of transcription of human genes (Daniel et al., 2014; Tajaddod et al., 2016). When it comes to *SMN* genes, several lines of evidence support a strong effect of Alu-derived motifs on their transcription. For instance, a deletion of ∼1.1 kb sequence within the *SMN* promoter region containing several Alu elements produced more than fivefold increase in the transcription activity in a cell-based reporter assay (Echaniz-Laguna et al., 1999). Further, an AluJb located immediately upstream of the most frequently used transcription start site (TSS1a) harbors several transcription regulatory motifs including a fetal transcription start site, TSS2 (Germain-Desprez et al., 2001; **Figure 2A**). Another transcription start site, TSS3, located 134 bp upstream of TSS1a also falls within the AluJb sequence (Monani et al., 1999b). *SMN* transcripts generated from either TSS2 or TSS3 possess longer 5′UTR with significance to unique regulation of transport, stability, and translation of these mRNAs. Transcription regulatory motifs located within the Alu elements include AP2alpha, ARNT, CREB, E2F, EN-1, Ets, HNF-3beta, interferon-stimulated responsive element (ISRE), MZF 1-4, PAX-2, SP-1, and SRY (**Figure 2A**). However, the significance of these motifs remains to be investigated. Interestingly, promoters harboring Alu elements are subject to regulation by long-noncoding RNAs (lncRNAs). For instance, *ANRIL*, a lncRNA identified in a GWAS as a risk factor in coronary artery disease, has been suggested to regulate expression of a network of genes through Alu elements located in their promoters (Holdt et al., 2013). This lncRNA potentially recruits PRC2, the chromatin remodeling complex (Holdt et al., 2013). Interestingly, *SMN* locus has been shown to generate two lncRNAs through transcription in the antisense direction. One of these lncRNAs termed *SMN-AS1* is a ∼1.6 kb long transcript that maps to the Alu-rich intron 1 (d'Ydewalle et al., 2017; **Figure 2B**). The other lncRNA termed *SMN-AS1** is a ∼10 kb long transcript, which maps to the Alu-rich regions that contains a portion of intron 5, intron 6 and the intergenic region downstream of exon 8 (Woo et al., 2017; **Figure 2B**). Depletion of *SMN-AS1* or *SMN-AS1** has been found to enhance transcription of *SMN2* (d'Ydewalle et al., 2017; Woo et al., 2017). Interestingly, *SMN-AS1* maps to a chromatin region rich in acetylated and/or methylated histone H3 in embryonic stem cells, suggesting a tissue-specific regulation of transcription by this lncRNA (**Figure 2B**). In contrast, *SMN-AS1** is expressed from a region which is not so rich in histone H3 modifications (**Figure 2B**). It has been proposed that both *SMN-AS1* and *SMN-AS1** modulate rate of transcription elongation through recruitment of the PRC2 complex (d'Ydewalle et al., 2017; Woo et al., 2017).

ALU ELEMENTS AND *SMN* SPLICING

Currently there is no study on the impact of Alu elements on splicing of various *SMN* exons. Alu elements can affect pre-mRNA splicing depending upon their sequence, orientation, location, and abundance. When present as inverted repeats, Alu elements form long double-stranded structures looping out intronic and/or exonic sequences. In cases where intra-intronic sequences are looped out, the 5′ and the 3′ ss are brought into close proximity, favoring intron removal (**Figure 3**). However, when an exon is looped out, its skipping is likely to be favored due to sequestration of the splice sites in the loop and increased competition from upstream and downstream splice sites, which are now in closer proximity to each other (**Figure 3**). The high abundance of Alu elements in *SMN* introns serve as the potential source for the intragenic base pairing among Alu sequences with opposite orientations. It is not known if some of these Alu-associated intra-intronic structures of *SMN* pre-mRNA are stabilized by ADARs. Several *SMN* exons are susceptible to skipping under conditions of oxidative stress (Singh et al., 2012; Seo et al., 2016b). It is likely that the ATP deficit caused by oxidative stress reduces the efficiency of RNA helicases such as DHX9, which unwinds the Alu-associated secondary structures within *SMN* pre-mRNA.

In addition to secondary-structure-associated regulation of splicing, Alu sequences can also recruit splicing factors on pre-mRNAs by interacting with complementary Alu sequences in lncRNAs. One such interaction has recently been proposed for 5S-OT, a lncRNA transcribed from 5S ribosomal RNA gene (Hu et al., 2016). The 3′-end of 5S-OT contains an Alu-derived 152 nt sequence that is complementary to the 3′ region of the sense Alu elements within introns 1, 2b, 4, and 6 of *SMN*. A polypyrimidine tract (Py) in the middle of the 5S-OT recruits the splicing factor U2AF65. Bioinformatics analysis revealed that 5S-OT regulates splicing of several exons, for which the

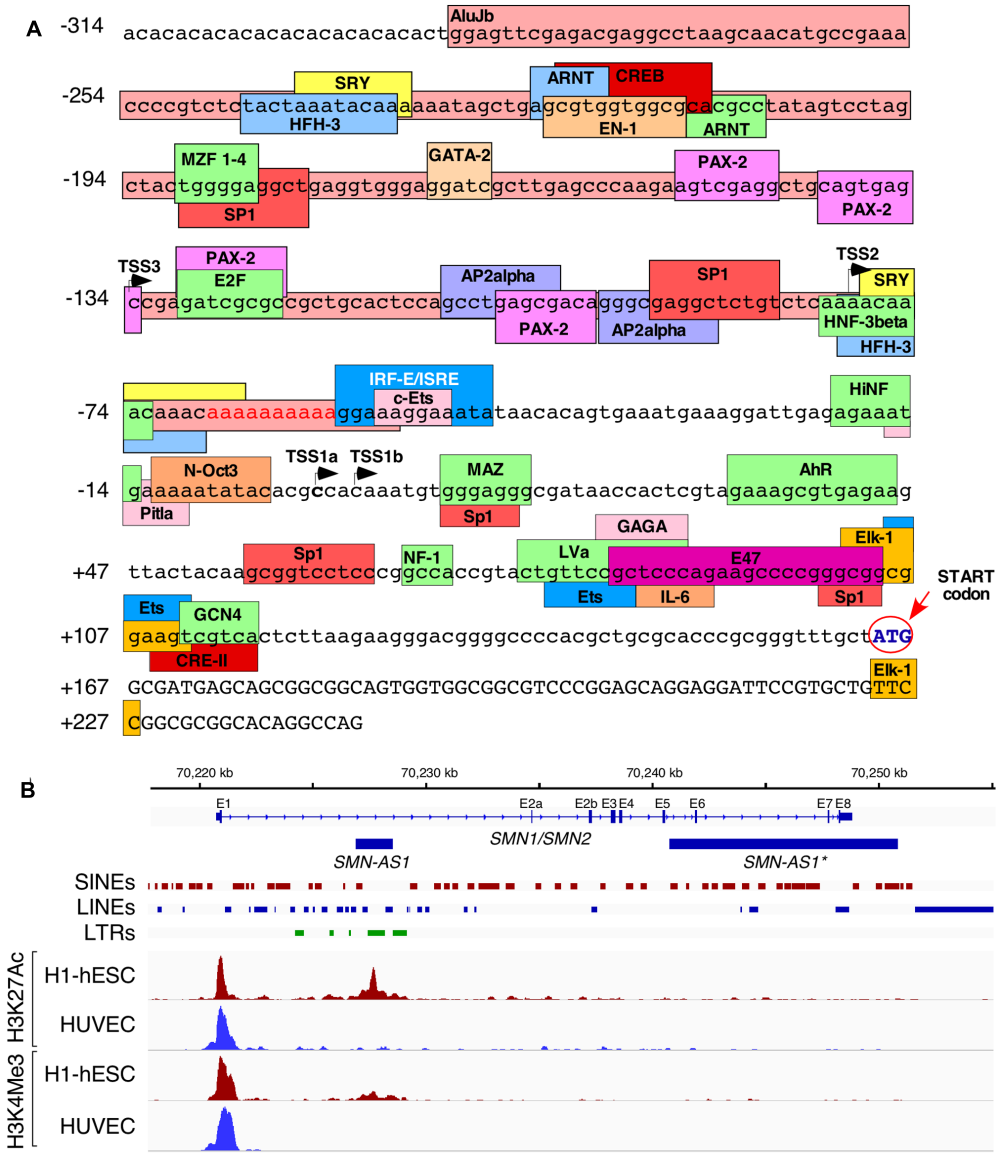


FIGURE 2 | Alu repeats reshape transcriptional regulation of *SMN*. **(A)** Overview of the promoter region of *SMN*. Lowercase letters indicate untranscribed or untranslated regions, uppercase letters indicate coding sequences. A partial AluJb element is inserted from 289 to 54 bases upstream of the canonical TSS (TSS1a) and is indicated with a dark pink colored box. Other colored boxes indicate the locations of putative promoter elements and/or transcription factor binding sites. TSSs are indicated with black arrows. The start codon is written in blue and is indicated with a red arrow and circle. Promoter elements were either described previously (Singh et al., 2016) or were computationally predicted by ConSite (consite.genereg.net). **(B)** Locations of Polycomb-associated antisense transcripts derived from *SMN*. A genomic view of the *SMN* gene is shown. Location of *SMN* exons (E1–E8) and antisense transcripts are indicated with blue boxes, introns are indicated by lines with arrows indicating the direction of transcription. The locations of SINEs, LINEs, and LTRs are indicated. Four ChIP-Seq outputs are given below. The red peaks indicate the read depth from ChIP-Seq of human embryonic stem cells using antibodies against acetylated H3K27 (upper) or tri-methylated H3K4 (lower). The blue peaks indicate the same readout from umbilical cord-derived HUVEC cells. ChIP-Seq data was obtained from ENCODE and previously described by d'Ydewalle et al. (2017).

Alu elements are located in the sense direction within the 2-kb upstream and/or downstream (Hu et al., 2016). In particular, knockdown of 5S-OT promoted inclusion and skipping of exons that had Alu elements located downstream and upstream of these exon, respectively. We speculate that splicing of *SMN* exons are regulated by 5S-OT, *SMN-AS1, SMN-AS1**, and possibly several other Alu-containing lncRNAs.

GENERATION OF CIRCULAR RNAS BY *SMN* GENES

Circular RNAs (circRNAs) are generated by back splicing in which the 5′ss of an exon joins the 3′ss of an upstream exon. In agreement with the potential link between Alu elements and back splicing, Alu elements are highly enriched upstream and

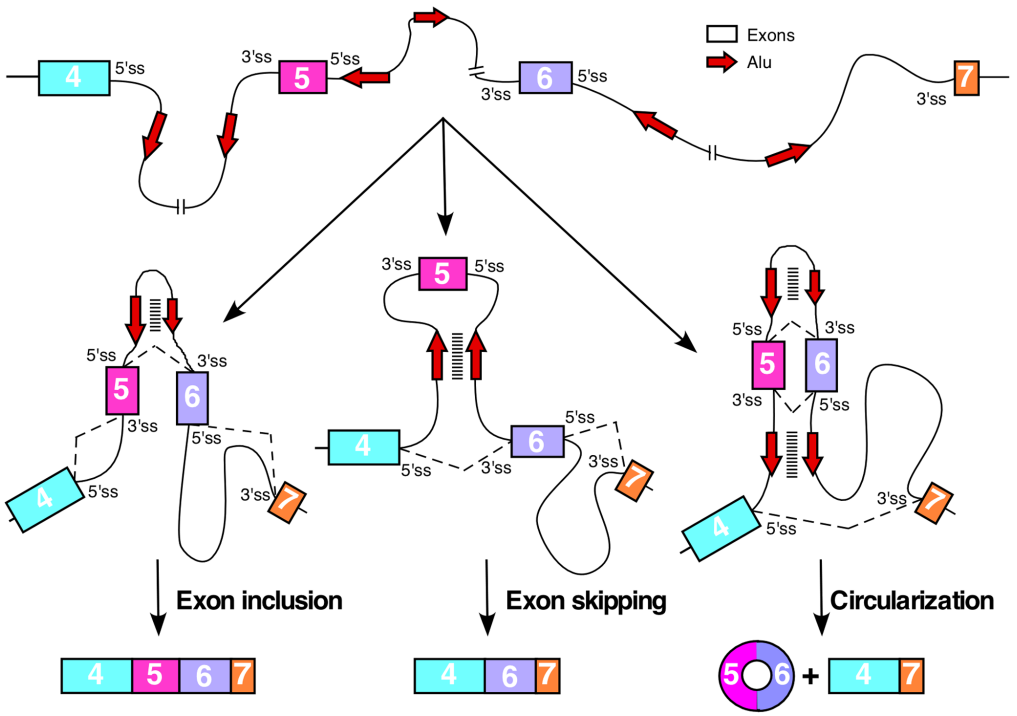


FIGURE 3 | Consequences of Alu–Alu base pairing within *SMN* pre-mRNA. Diagrammatic representation of the partial *SMN* pre-mRNA (not to the scale) showing inclusion, skipping and circularization of exon 5. Exons are shown in colored boxes, whereas, red arrows indicated Alu elements. For simplicity, only two Alu elements per intron are shown. Base pairing between complementary Alu sequences are shown by stacked lines. Various hypothetical scenarios of splicing reactions involving different combinations of splice site pairings are shown by broken lines. Base pairing between Alu sequences within intron 5 promotes inclusion of exon 5, whereas, base pairing between intronic Alu sequences flanking exon 5 promotes exon 5 skipping. Base pairing between Alu elements within intron 5 combined with the base pairing between intronic Alu sequences upstream of exon 5 and downstream of exon 6 promotes circularization event.

downstream of splice sites associated with circRNA formation (Jeck et al., 2013). Based on the online repository circBase, at least two circRNAs are generated by *SMN* (Glazar et al., 2014). One of these circRNAs is generated by back splicing of the 5′ss of exon 4 with the 3′ss of exon 2B, whereas, the other circRNA is generated by back splicing of the 5′ss of exon 6 with the 3′ss of the exon 5 (**Figure 3**). Based on the high density of intronic Alu elements, we expect generation of additional circRNAs by *SMN* genes. Owing to their extreme stability, even small levels of circRNAs may affect cellular metabolism by sequestering miRNAs and regulatory RNA-binding proteins (Hansen et al., 2013; Memczak et al., 2013). We expect that various *SMN* circRNAs are differentially expressed in different tissues. Future studies will determine what circRNAs are generated by *SMN* genes and how they impact the formation of linear *SMN* transcripts and affects cellular metabolism in different tissues. It will be also important to know if differential splicing of exon 7 leads to distinct circRNA patterns of *SMN1* and *SMN2*.

EXONIZATION OF AN INTRONIC ALU ELEMENT

We recently reported a novel exon, exon 6B, generated by exonization of a 109-nt long sequence located within *SMN* intron 6 (Seo et al., 2016a; **Figure 4A**). Thus far, exon 6B is the only known exon to be derived from an Alu element within *SMN*. Exon 6B maps to the left arm of the antisense sequence of an Alu element and appears to be conserved in all members of the Hominidae family (Seo et al., 2016a; **Figure 4A**). Considering most Alu-derived exons originate from the right arm of the antisense sequence of an Alu element (Sorek et al., 2002), generation of exon 6B is an example of a rare event. The relative abundance of exon 6B-containing transcripts was found to be low compared to transcripts lacking exon 6B. This is in part due to degradation of exon 6B-containing transcripts by NMD, a translation dependent process. Consistently, inhibition of translation by cycloheximide elevated the levels of exon 6B-containing transcripts (Seo et al., 2016a). Depletion of UPF1, an essential component of NMD pathway, was also found to upregulate exon 6B-containing transcripts. Degradation of exon 6B-containing transcripts could also be facilitated by SMD, an UPF1-dependent process triggered by base pairing of Alu sequences in mRNAs with the Alu-containing lncRNAs (Park and Maquat, 2013).

Consistent with the low abundance of exon 6B-containing transcripts, the predicted strengths of splice sites of exon 6B were significantly lower than that for the neighboring exons 6 and 7 (Seo et al., 2016a). However, we observed a dense map of overlapping enhancer motifs within exon 6B and its

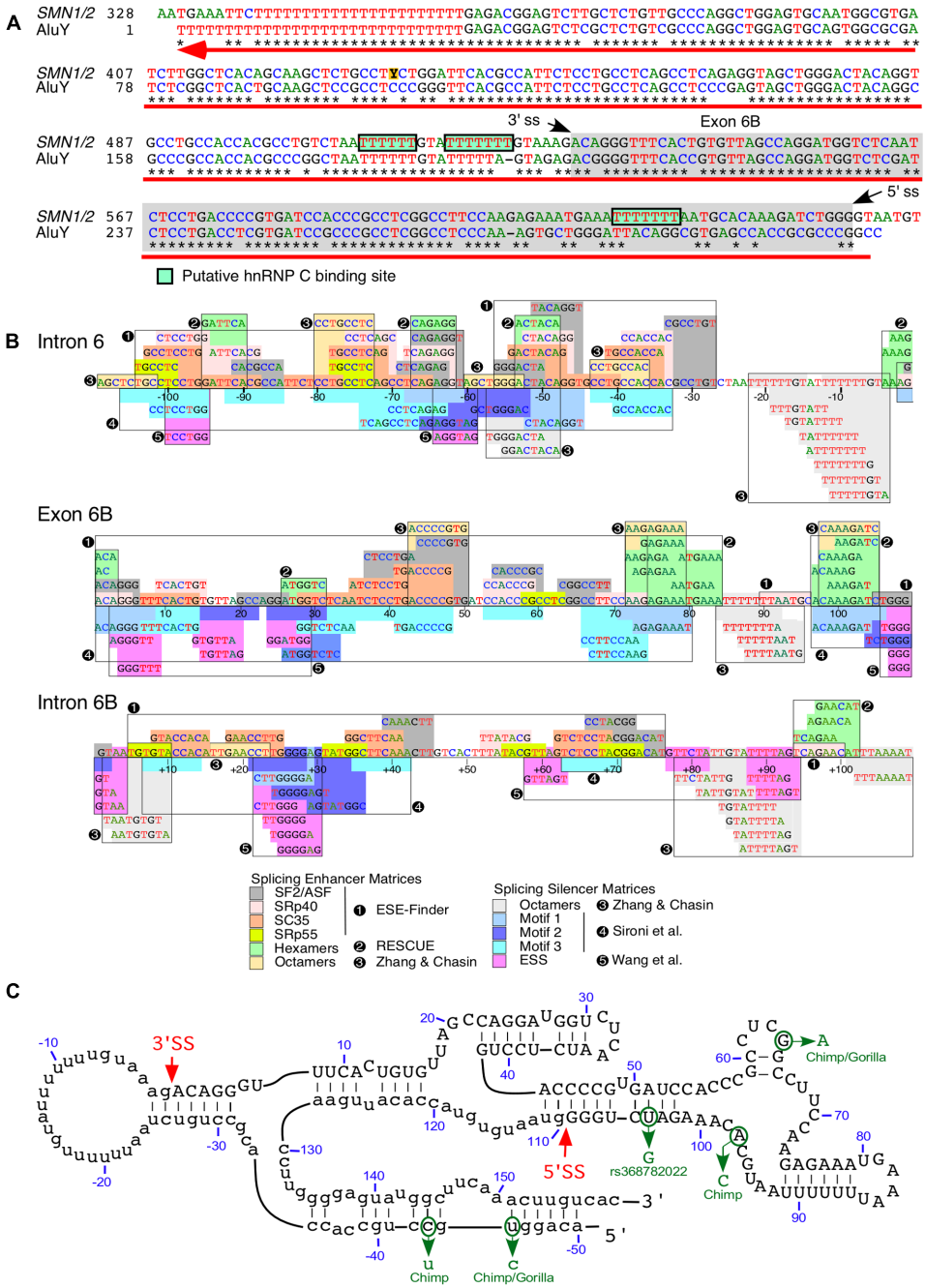


FIGURE 4 | Exon 6B is derived from an intronic Alu element. **(A)** Alignment of SMN intron 6 region spanning exon 6B. Numbering starts from the beginning of intron 6. Stars signify sequence identity. Hyphens designate the positions where gaps were introduced to maximize sequence identity. The gray box indicates exon 6B sequences and the green boxes indicate putative binding sites for hnRNP C. The black arrows indicate splice site (ss) positions of exon 6B. The red arrow indicates position and direction of AluY insertion (reverse and complement) which are obtained from Dfam (Accession: DF0000002). **(B)** Predicted splicing cis-elements. The exon 6B and 109 nt of upstream and downstream intronic sequences are shown. Colored boxes indicate potential regulatory elements identified by Human Splicing Finder (Desmet et al., 2009). Potential splicing enhancers are indicated above the *SMN* sequence and splicing silencers are below. Color code is explained in the bottom panel, where numbers indicate the software tool used for identification or publications in which motifs were originally described. Exonic splicing enhancer (ESE) finder is described in (Cartegni et al., 2003). RESCUE refers to an algorithm that predicts ESEs (Fairbrother et al., 2002). Octamer motifs are described in (Zhang and Chasin, 2004). Motifs 1-3 are described in (Sironi et al., 2004). Silencer motifs highlighted in pink are described in (Wang et al., 2004). **(C)** Secondary structure of *SMN* exon 6B. Numbering starts from the beginning of exon 6B. Exon 6B sequences are shown in capital letters, while adjacent intronic sequences are shown in lower-case letters. The red arrows indicate ss positions of exon 6B. The green arrows indicate sequence differences of exon 6B between human and primates. The secondary structure was predicted using mfold algorithm (Zuker, 2003).

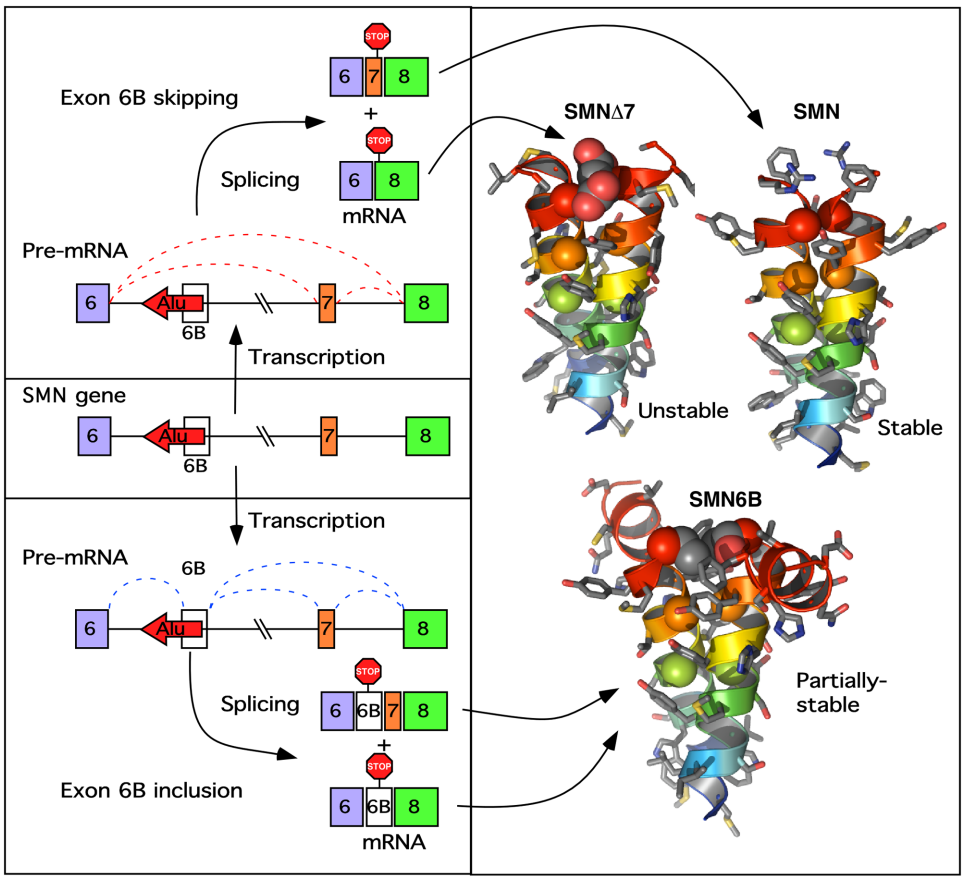


FIGURE 5 | A model of exon 6B action. **(Left)** Describes the transcription of the *SMN* gene and pre-mRNA splicing producing either the 6B-skipped (upper) or 6B-included (lower mRNA). Exons are indicated as colored boxes, the Alu element from which exon 6B is derived is indicated as a red arrow, introns are shown as lines. Potential splicing events are shown as red (exon 6B-skipped) or blue (exon 6B-included) dotted lines. Locations of stop codons generated by each potential transcript are indicated. **(Right)** Shows the computationally predicted glycine zipper dimers formed by the YG boxes at the C termini of each of the SMN protein isoforms. Both SMNΔ7 and SMN6B have altered YG boxes resulting in an increase in the inter-helical distances of the coiled-coil interaction, potentially reducing oligomerization. In SMNΔ7 this results in an unstable degron (Cho and Dreyfuss, 2010), whereas in SMN6B the destabilization is less pronounced (Seo et al., 2016a).

flanking intronic sequences (**Figure 4B**). Of note, nucleotide differences between exon 6B and AluY are predicted to create several enhancer elements toward the 3′ end of this exon (**Figure 4B**). These elements might contribute to the regulation of exon 6B splicing. As expected, the incorporation of exon 6B appeared to be suppressed by hnRNP C that is known to inhibit the exonization of intronic Alu elements (Seo et al., 2016a). It has been demonstrated that TIA1 regulates *SMN* exon 7 splicing through interaction with the intronic uridine-rich clusters downstream of exon 7 (Singh et al., 2011). Interestingly, similar uridine-rich clusters are present downstream of exon 6B, pointing to the potential involvement of TIA1 in splicing of exon 6B. However, analysis of the publicly available transcriptome data showed no effect of TIA1 depletion on splicing of exon 6B (Seo et al., 2016a). Consistent with these results, we did not detect changes in the splicing of *SMN2* exon 6B in a SMA mouse model in which *Tia1* was deleted (Howell et al., 2017b).

We have previously shown that RNA secondary structures that sequester the splice sites affect inclusion of *SMN* exon 7 (Singh et al., 2004a,b,c, 2007, 2013). Interestingly, the predicted secondary structure of exon 6B and its flanking intronic sequences puts both the 3′ss and the 5′ss in stems (**Figure 4C**). It is likely that these stems play a negative role in inclusion of exon 6B by suppressing the splice site recognition. We also observed that the most of the exon 6B sequence is sequestered within the predicted terminal and internal stems (**Figure 4C**). It remains to be seen if these intra-exonic structures play any role in exon 6B splicing regulation. Critical role of an intra-intronic structure formed by a unique long-distance interaction has been demonstrated for regulation of *SMN* exon 7 splicing (Singh et al., 2010, 2013; Howell et al., 2017a). Interestingly, the predicted secondary structure of sequences downstream of exon 6B reveal long-distance interactions formed between sense and antisense Alu elements (not shown). It is likely that these structures play some role in regulation of exon 6B splicing.

POTENTIAL FUNCTIONS OF SMN6B

Exon 6B-containg transcripts are expressed in all tissues and code for SMN6B, which contains an identical number of amino acids as SMN (Seo et al., 2016a). However, SMN differs from SMN6B by sixteen C-terminal amino acids (**Figure 5**). In particular, in the transcripts containing exon 7 but lacking exon 6B, the last sixteen C-terminal amino acids are coded by exon 7. It is known that the amino acids coded by exon 7 play an important role in stabilization, self-oligomerization and protein-protein interactions (Singh et al., 2017c). Hence, the loss of these amino acids is the primary cause of the poor stability as well as the suboptimal functions of SMNΔ7 (Cho and Dreyfuss, 2010). A side-by-side comparison showed that SMN6B is less stable than SMN (Seo et al., 2016a). At the same time, the stability of SMN6B was found to be greater than SMNΔ7 (Seo et al., 2016a). Similar to SMN, SMN6B localizes to both the nucleus and the cytosol. Further, SMN6B interacts with Gemin2, which is associated with most SMN functions. Hence, it is likely that SMN and SMN6B share most of the cellular functions including snRNP biogenesis, transcription, translation, macromolecular trafficking, telomerase biogenesis, selenoproteins biosynthesis, signal transduction and stress granule formation. High copy numbers of *SMN2* are associated with low severity of SMA, likely due to the expected high levels of SMN6B. However, levels of SMN6B produced in SMA patients remain unknown. Factors that regulate expression of SMN6B are expected to modulate the severity of SMA. Future studies will determine if SMN6B has a tissue-specific function. Of note, production of SMN6B will confer unparalleled therapeutic benefits in SMA patients carrying deletions of genomic sequences downstream of exon 6B. A proper understanding of inhibitory cis-elements that regulate exon 6B splicing will provide novel targets for the stimulation of exon 6B inclusion leading to the production of SMN6B.

CONCLUDING REMARKS

Given the importance of SMN in cellular metabolism and its association with various pathological conditions, there has been general interest in the mechanisms by which levels of SMN are regulated. Since the evolution of primates, the genomic landscape of the *SMN* locus has undergone massive changes, including duplication sometime before the divergence of human and chimpanzee lineages and the human-specific mutations characteristic of *SMN2* (Rochette et al., 2001). Among these changes, and perhaps a driving force for other ones, are the insertion of a large number of Alu elements into intronic and intergenic regions (**Figure 1**). Based on the deletions in the Alu-rich promoter region as well as the recent discoveries of the Alu-containing lncRNAs, we propose that the regulation of transcription of the *SMN* genes is distinct in primates. Similarly, given the preponderance of Alu elements in most *SMN* introns, we anticipate that splicing of *SMN* exons is uniquely regulated in primates. There is a strong likelihood that human *SMN* genes are subjected to a unique transcription-coupled splicing regulation primarily due to the abundance of Alu elements within *SMN* genes. However, mechanism of such regulation remains to be investigated.

The finding of Alu-derived exon 6B adds an additional regulatory step in the expression of *SMN* genes. Our results suggest that inclusion of exon 6B inhibits skipping of *SMN2* exon 7 (Seo et al., 2016a). However, irrespective of exon 7 inclusion or skipping, transcripts harboring exon 6B code for the same SMN6B protein, which displays higher stability than SMNΔ7 (Seo et al., 2016a). Our findings suggest that an enhanced expression of SMN6B may confer therapeutic benefits when SMN is absent or expressed at very low levels. A handful of genes code for proteins with C-terminal sequences similar to those coded by exon 6B (Seo et al., 2016a). Hence, it will be interesting to know if these proteins possess some common properties. Mice carry B1 elements that share several properties with Alu elements. There is also evidence to suggest that several of the functions of Alu elements are carried by B1 elements in mice (Aktaş et al., 2017). However, due to a size difference between Alu and B1 elements, it is expected that B1 elements cannot perform all Alu-associated functions. In particular, it is highly unlikely that circRNAs induced by Alu elements are also generated in non-primates. Further, human *SMN* genes are unique in producing lncRNAs harboring Alu elements. Future studies will determine how insertion of Alu elements have impacted the regulation and regulatory activities of *SMN* genes, which are linked to various pathological conditions in humans.

AUTHOR CONTRIBUTIONS

EO, JS, and RS analyzed literature and/or publically available data. EO, JS, NS, and RS designed diagrams and wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (R01 NS055925 and R21 NS101312), Iowa Center for Advanced Neurotoxicology (ICAN), and Salsbury Endowment (Iowa State University, Ames, IA, United States) to RS.

REFERENCES

Ahmad, S., Bhatia, K., Kannan, A., and Gangwani, L. (2016). Molecular mechanisms of neurodegeneration in spinal muscular atrophy. *J. Exp. Neurosci.* 10, 39–49. doi: 10.4137/jen.s33122

Aktaş, T., Ilik, I. A., Maticzka, D., Bhardwaj, V., Rodrigues, C. P., Mittler, G., et al. (2017). DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* 544, 115–119. doi: 10.1038/nature21715

Antonaki, A., Demetriades, C., Polyzos, A., Banos, A., Vatsellas, G., Lavigne, M. D., et al. (2011). Genomic analysis reveals a novel nuclear factor-kappa B (NF-kappa B)-binding Site in Alu-repetitive elements. *J. Biol. Chem.* 286, 38768–38782. doi: 10.1074/jbc.M111.234161

Attig, J., Mozos, I., Haberman, N., Wang, Z., Emmett, W., Zarnack, K., et al. (2016). Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *Elife.* 5:e19545. doi: 10.7554/eLife.19545

Bouttier, M., Laperriere, D., Memari, B., Mangiapane, J., Fiore, A., Mitchell, E., et al. (2016). Alu repeats as transcriptional regulatory platforms in macrophage responses to M-tuberculosis infection. *Nucleic Acids Res.* 44, 10571–10587. doi: 10.1093/nar/gkw782

Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., and Krainer, A. R. (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* 31, 3568–3571. doi: 10.1093/nar/gkg616

Cho, S. C., and Dreyfuss, G. (2010). A degron created by SMN2 exon 7 skipping is a principal contributor to spinal muscular atrophy severity. *Genes Dev.* 24, 438–442. doi: 10.1101/gad.1884910

Cui, F., Sirotin, M. V., and Zhurkin, V. B. (2011). Impact of Alu repeats on the evolution of human p53 binding sites. *Biol. Direct.* 6:2. doi: 10.1186/1745-6150-6-2

Daniel, C., Silberberg, G., Behm, M., and Ohman, M. (2014). Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biol.* 15:R28. doi: 10.1186/gb-2014-15-2-r28

Deininger, P. (2011). Alu elements: know the SINEs. *Genome Biol.* 12:236. doi: 10.1186/gb-2011-12-12-236

Deininger, P. L., Moran, J. V., Batzer, M. A., and Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13, 651–658. doi: 10.1016/j.gde.2003.10.013

Desmet, F. O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37:e67. doi: 10.1093/nar/gkp215

d'Ydewalle, C., Ramos, D. M., Pyles, N. J., Ng, S. Y., Gorz, M., Pilato, C. M., et al. (2017). The antisense transcript SMN-AS1 regulates SMN expression and is a novel therapeutic target for spinal muscular atrophy. *Neuron* 93, 66–79. doi: 10.1016/j.neuron.2016.11.033

Echaniz-Laguna, A., Miniou, P., Bartholdi, D., and Melki, J. (1999). The promoters of the survival motor neuron gene (SMN) and its copy (SMNc) share common regulatory elements. *Am. J. Hum. Genet.* 64, 1365–1370. doi: 10.1086/302372

Elbarbary, R. A., and Maquat, L. E. (2017). Distinct mechanisms obviate the potentially toxic effects of inverted-repeat Alu elements on cellular RNA metabolism. *Nat. Struct. Mol. Biol.* 24, 496–498. doi: 10.1038/nsmb.3416

Fairbrother, W. T., Yeh, R. F., Sharp, P. A., and Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* 292, 1007–1013. doi: 10.1126/science.1073774

Germain-Desprez, D., Brun, T., Rochette, C., Semionov, A., Rouget, R., and Simard, L. R. (2001). The SMN genes are subject to transcriptional regulation during cellular differentiation. *Gene* 279, 109–117. doi: 10.1016/s0378-1119(01)00758-2

Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *RNA* 20, 1666–1670. doi: 10.1261/rna.043687.113

Gu, T. J., Yi, X., Zhao, X. W., Zhao, Y., and Yin, J. Q. (2009). Alu-directed transcriptional regulation of some novel miRNAs. *BMC Genomics* 10:563. doi: 10.1186/1471-2164-10-563

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Hedges, D. J., and Batzer, M. A. (2005). From the margins of the genome: mobile elements shape primate evolution. *Bioessays* 27, 785–794. doi: 10.1002/bies.20268

Holdt, L. M., Hoffmann, S., Sass, K., Langenberger, D., Scholz, M., Krohn, K., et al. (2013). Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLOS Genet.* 9:e1003588. doi: 10.1371/journal.pgen.1003588

Howell, M. D., Ottesen, E. W., Singh, N. N., Anderson, R. L., and Singh, R. N. (2017a). Gender-specific amelioration of SMA phenotype upon disruption of a deep intronic structure by an oligonucleotide. *Mol. Ther.* 25, 1328–1341. doi: 10.1016/j.ymthe.2017.03.036

Howell, M. D., Ottesen, E. W., Singh, N. N., Anderson, R. L., Seo, J., Sivanesan, S., et al. (2017b). TIA1 is a gender-specific disease modifier of a mild mouse model of spinal muscular atrophy. *Sci. Rep.* 7:18. doi: 10.1038/s41598-017-07468-2

Howell, M. D., Singh, N. N., and Singh, R. N. (2014). Advances in therapeutic development for spinal muscular atrophy. *Future Med. Chem.* 6, 1081–1099. doi: 10.4155/fmc.14.63

Hu, S. S., Wang, X. L., and Shan, G. (2016). Insertion of an Alu element in a lncRNA leads to primate-specific modulation of alternative splicing. *Nat. Struct. Mol. Biol.* 23, 1011–1019. doi: 10.1038/nsmb.3302

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J. Z., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi: 10.1261/rna.035667.112

Kiel, J. M., Seo, J., Howell, M. D., Hsu, W. H., Singh, R. N., and DiDonato, C. J. (2014). A short antisense oligonucleotide ameliorates symptoms of severe mouse models of spinal muscular atrophy. *Mol. Ther. Nucleic Acids* 3:e174. doi: 10.1038/mtna.2014.23

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Lefebvre, S., Bürglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., et al. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 80, 155–165. doi: 10.1016/0092-8674(95)90460-3

Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3 ' splice-site selection in Alu exons. *Science* 300, 1288–1291. doi: 10.1126/science.1082588

Liang, D. M., and Wilusz, J. E. (2014). Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* 28, 2233–2247. doi: 10.1101/gad.251926.114

Lorson, C. L., Hahnen, E., Androphy, E. J., and Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6307–6311. doi: 10.1073/pnas.96.11.6307

Lorson, C. L., Strasswimmer, J., Yao, J. M., Baleja, J. D., Hahnen, E., Wirth, B., et al. (1998). SMN oligomerization defect correlates with spinal muscular atrophy severity. *Nat. Genet.* 19, 63–66. doi: 10.1038/ng0598-63

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Monani, U. R., Lorson, C. L., Parsons, D. W., Prior, T. W., Androphy, E. J., Burghes, A. H. M., et al. (1999a). A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.* 8, 1177–1183. doi: 10.1093/hmg/8.7.1177

Monani, U. R., McPerson, J. D., and Burghes, A. H. M. (1999b). Promoter analysis of the human centromeric and telomeric survival motor neuron genes (SMNC and SMNT). *Biochim. Biophys. Acta* 1445, 330–336. doi: 10.1016/s0167-4781(99)00060-3

Morales, M. E., White, T. B., Streva, V. A., DeFreece, C. B., Hedges, D. J., and Deininger, P. L. (2015). The contribution of alu elements to mutagenic DNA double-strand break repair. *PLOS Genet.* 11:e1005016. doi: 10.1371/journal.pgen.1005016

Ottesen, E. W. (2017). ISS-N1 makes the first FDA-approved drug for spinal muscular atrophy. *Transl. Neurosci.* 8, 1–6. doi: 10.1515/tnsci-2017-0001

Ottesen, E. W., Howell, M. D., Singh, N. N., Seo, J., Whitley, E. M., and Singh, R. N. (2016). Severe impairment of male reproductive organ development in a low SMN expressing mouse model of spinal muscular atrophy. *Sci. Rep.* 6:20193. doi: 10.1038/srep20193

Park, E., and Maquat, L. E. (2013). Staufen-mediated mRNA decay. *Wiley Interdiscip. Rev. RNA* 4, 423–435. doi: 10.1002/wrna.1168

Payer, L. M., Steranka, J. P., Yang, W. R., Kryatova, M., Medabalimi, S., Ardeljan, D., et al. (2017). Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proc. Natl. Acad. Sci. U.S.A.* 114, E3984–E3992. doi: 10.1073/pnas.1704117114

Rochette, C. F., Gilbert, N., and Simard, L. R. (2001). SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Hum. Genet.* 108, 255–266. doi: 10.1007/s004390100473

Sakurai, M., Shiromoto, Y., Ota, H., Song, C. Z., Kossenkov, A. V., Wickramasinghe, J., et al. (2017). ADAR1 controls apoptosis of stressed cells by inhibiting Staufen1-mediated mRNA decay. *Nat. Struct. Mol. Biol.* 24, 534–543. doi: 10.1038/nsmb.3403

Schmutz, J., Martin, J., Terry, A., Couronne, O., Grimwood, J., Lowry, S., et al. (2004). The DNA sequence and comparative analysis of human chromosome 5. *Nature* 431, 268–274. doi: 10.1038/nature02919

Sen, S. K., Han, K. D., Wang, J. X., Lee, J., Wang, H., Callinan, P. A., et al. (2006). Human genomic deletions mediated by recombination between Alu elements. *Am. J. Hum. Genet.* 79, 41–53. doi: 10.1086/504600

Seo, J., Howell, M. D., Singh, N. N., and Singh, R. N. (2013). Spinal muscular atrophy: an update on therapeutic progress. *Biochim. Biophys. Acta* 1832, 2180–2190. doi: 10.1016/j.bbadis.2013.08.005

Seo, J., Singh, N. N., Ottesen, E. W., Lee, B. M., and Singh, R. N. (2016a). A novel human-specific splice isoform alters the critical C-terminus of survival motor neuron protein. *Sci. Rep.* 6:30778. doi: 10.1038/srep30778

Seo, J., Singh, N. N., Ottesen, E. W., Sivanesan, S., Shishimorova, M., and Singh, R. N. (2016b). Oxidative stress triggers body-wide skipping of multiple exons of the spinal muscular atrophy gene. *PLOS ONE* 11:e0154390. doi: 10.1371/journal.pone.0154390

Singh, N. K., Singh, N. N., Androphy, E. J., and Singh, R. N. (2006). Splicing of a critical exon of human survival motor neuron is regulated by a unique silencer element located in the last intron. *Mol. Cell. Biol.* 26, 1333–1346. doi: 10.1128/mcb.26.4.1333-1346.2006

Singh, N. N., Androphy, E. J., and Singh, R. N. (2004a). An extended inhibitory context causes skipping of exon 7 of SMN2 in spinal muscular atrophy. *Biochem. Biophys. Res. Commun.* 315, 381–388. doi: 10.1016/j.bbrc.2004.01.067

Singh, N. N., Androphy, E. J., and Singh, R. N. (2004b). In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. *RNA* 10, 1291–1305. doi: 10.1261/rna.7580704

Singh, N. N., Androphy, E. J., and Singh, R. N. (2004c). The regulation and regulatory activities of alternative splicing of the SMN gene. *Crit. Rev. Eukaryot. Gene Exp.* 14, 271–285. doi: 10.1615/CritRevEukaryotGeneExpr.v14.i4.30

Singh, N. N., Del Rio-Malewski, J. B., Luo, D., Ottesen, E. W., Howell, M. D., and Singh, R. N. (2017a). Activation of a cryptic 5' splice site reverses the impact of pathogenic splice site mutations in the spinal muscular atrophy gene. *Nucleic Acids Res.* doi: 10.1093/nar/gkx824 [Epub ahead of print].

Singh, N. N., Howell, M. D., Androphy, E. J., and Singh, R. N. (2017b). How the discovery of ISS-N1 led to the first medical therapy for spinal muscular atrophy. *Gene Ther.* 24, 520–526. doi: 10.1038/gt.2017.34

Singh, R. N., Howell, M. D., Ottesen, E. W., and Singh, N. N. (2017c). Diverse role of survival motor neuron protein. *Biochim. Biophys. Acta* 1860, 299–315. doi: 10.1016/j.bbagrm.2016.12.008

Singh, N. N., Hollinger, K., Bhattacharya, D., and Singh, R. N. (2010). An antisense microwalk reveals critical role of an intronic position linked to a unique long-distance interaction in pre-mRNA splicing. *RNA* 16, 1167–1181. doi: 10.1261/rna.2154310

Singh, N. N., Howell, M. D., and Singh, R. N. (2016). "Transcriptional and splicing regulation of spinal muscular atrophy genes," in *Spinal Muscular Atrophy: Disease Mechanisms and Therapy*, eds S. J. Charlotte, S. Paushkin, and C.-P. Ko (Amsterdam: Elsevier Inc).

Singh, N. N., Lawler, M. N., Ottesen, E. W., Upreti, D., Kaczynski, J. R., and Singh, R. N. (2013). An intronic structure enabled by a long-distance interaction serves as a novel target for splicing correction in spinal muscular atrophy. *Nucleic Acids Res.* 41, 8144–8165. doi: 10.1093/nar/gkt609

Singh, N. N., Lee, B. M., DiDonato, C. J., and Singh, R. N. (2015a). Mechanistic principles of antisense targets for the treatment of spinal muscular atrophy. *Fut. Med. Chem.* 7, 1793–1808. doi: 10.4155/fmc.15.101

Singh, N. N., Lee, B. M., and Singh, R. N. (2015b). Splicing regulation in spinal muscular atrophy by a RNA structure formed by long distance interactions. *Ann. N. Y. Acad. Sci.* 1341, 176–187. doi: 10.1111/nyas.12727

Singh, N. N., Seo, J., Rahn, S. J., and Singh, R. N. (2012). A multi-exon-skipping detection assay reveals surprising diversity of splice isoforms of spinal muscular atrophy genes. *PLOS ONE* 7:e49595. doi: 10.1371/journal.pone.0049595

Singh, N. N., Seo, J. B., Ottesen, E. W., Shishimorova, M., Bhattacharya, D., and Singh, R. N. (2011). TIA1 prevents skipping of a critical exon associated with spinal muscular atrophy. *Mol. Cell. Biol.* 31, 935–954. doi: 10.1128/mcb.00945-10

Singh, N. N., Shishimorova, M., Cao, L. C., Gangwani, L., and Singh, R. N. (2009). A short antisense oligonucleotide masking a unique intronic motif prevents skipping of a critical exon in spinal muscular atrophy. *RNA Biol.* 6, 341–350. doi: 10.4161/rna.6.3.8723

Singh, N. N., and Singh, R. N. (2011). Alternative splicing in spinal muscular atrophy underscores the role of an intron definition model. *RNA Biol.* 8, 600–606. doi: 10.4161/rna.8.4.16224

Singh, N. N., Singh, R. N., and Androphy, E. J. (2007). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Res.* 35, 371–389. doi: 10.1093/nar/gkl1050

Singh, R. N. (2007). Evolving concepts on human SMN Pre-mRNA splicing. *RNA Biol.* 4, 7–10. doi: 10.4161/rna.4.1.4535

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G. P., Bresolin, N., et al. (2004). Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res.* 32, 1783–1791. doi: 10.1093/nar/gkh341

Sivanesan, S., Howell, M. D., DiDonato, C. J., and Singh, R. N. (2013). Antisense oligonucleotide mediated therapy of spinal muscular atrophy. *Transl. Neurosci.* 4, 1–7. doi: 10.2478/s13380-013-0109-2

Smit, A. F. A., Hubley, R., and Green, P. (2015). *RepeatMasker Open-4.0*. Available at: www.repeatmasker.org. [accessed April 26, 2017].

Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067. doi: 10.1101/gr.229302

Su, M., Han, D. L., Boyd-Kirkup, J., Yu, X. M., and Han, J. D. J. (2014). Evolution of alu elements toward enhancers. *Cell Rep.* 7, 376–385. doi: 10.1016/j.celrep.2014.03.011

Sutherland, L. C., Thibault, P., Durand, M., Lapointe, E., Knee, J. M., Beauvais, A., et al. (2017). Splicing arrays reveal novel RBM10 targets, including SMN2 pre-mRNA. *BMC Mol. Biol.* 18:19. doi: 10.1186/s12867-017-0096-x

Tajaddod, M., Tanzer, A., Licht, K., Wolfinger, M. T., Badelt, S., Huber, F., et al. (2016). Transcriptome-wide effects of inverted SINEs on gene expression and their impact on RNA polymerase II activity. *Genome Biol.* 17, 220. doi: 10.1186/s13059-016-1083-0

Wang, L., Norris, E. T., and Jordan, I. K. (2017). Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. *Front. Microbiol.* 8:1418. doi: 10.3389/fmicb.2017.01418

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., and Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell* 119, 831–845. doi: 10.1016/j.cell.2004.11.010

White, T. B., Morales, M. E., and Deininger, P. L. (2015). Alu elements and DNA double-strand break repair. *Mob. Genet. Elements* 5, 81–85. doi: 10.1080/2159256X.2015.1093067

Wilusz, J. E. (2015). Repetitive elements regulate circular RNA biogenesis. *Mob. Genet. Elements* 5, 39–45. doi: 10.1080/2159256X.2015.1045682

Wirth, B., Herz, M., Wetter, A., Moskau, S., Hahnen, E., Rudnik-Schoneborn, S., et al. (1999). Quantitative analysis of survival motor neuron copies: identification of subtle SMN1 mutations in patients with spinal muscular atrophy, genotype-phenotype correlation, and implications for genetic counseling. *Am. J. Hum. Genet.* 64, 1340–1356. doi: 10.1086/302369

Wirth, B., Schmidt, T., Hahnen, E., RudnikSchoneborn, S., Krawczak, M., MullerMyhsok, B., et al. (1997). De novo rearrangements found in 2% of index patients with spinal muscular atrophy: mutational mechanisms, parental origin, mutation rate, and implications for genetic counseling. *Am. J. Hum. Genet.* 61, 1102–1111. doi: 10.1086/301608

Woo, C. J., Maier, V. K., Davey, R., Brennan, J., Li, G. D., Brothers, J., et al. (2017). Gene activation of SMN by selective disruption of lncRNA-mediated recruitment of PRC2 for the treatment of spinal muscular atrophy. *Proc. Natl. Acad. Sci. U.S.A.* 114, E1509–E1518. doi: 10.1073/pnas.1616521114

Yoshimoto, S., Harahap, N. I., Hamamura, Y., Ar Rochmah, M., Shima, A., Morisada, N., et al. (2016). Alternative splicing of a cryptic exon embedded in intron 6 of *SMN1* and *SMN2*. *Hum. Genome Var.* 3:16040. doi: 10.1038/hgv.2016.40

Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., et al. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of alu elements. *Cell* 152, 453–466. doi: 10.1016/j.cell.2012.12.023

Zhang, X. H., and Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* 18, 1241–1250. doi: 10.1101/gad.1195304

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415. doi: 10.1093/nar/gkg595

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 02 November 2017
doi: 10.3389/fmicb.2017.02151

Evolutionary Analysis of HIV-1 Pol Proteins Reveals Representative Residues for Viral Subtype Differentiation

Shohei Nagata[1,2], *Junnosuke Imai*[1,3], *Gakuto Makino*[1], *Masaru Tomita*[1,2,3] and *Akio Kanai*[1,2,3]*

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan, [2] Faculty of Environment and Information Studies, Keio University, Fujisawa, Japan, [3] Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa, Japan

RNA viruses have been used as model systems to understand the patterns and processes of molecular evolution because they have high mutation rates and are genetically diverse. *Human immunodeficiency virus 1* (HIV-1), the etiological agent of acquired immune deficiency syndrome, is highly genetically diverse, and is classified into several groups and subtypes. However, it has been difficult to use its diverse sequences to establish the overall phylogenetic relationships of different strains or the trends in sequence conservation with the construction of phylogenetic trees. Our aims were to systematically characterize HIV-1 subtype evolution and to identify the regions responsible for HIV-1 subtype differentiation at the amino acid level in the Pol protein, which is often used to classify the HIV-1 subtypes. In this study, we systematically characterized the mutation sites in 2,052 Pol proteins from HIV-1 group M (144 subtype A; 1,528 subtype B; 380 subtype C), using sequence similarity networks. We also used spectral clustering to group the sequences based on the network graph structures. A stepwise analysis of the cluster hierarchies allowed us to estimate a possible evolutionary pathway for the Pol proteins. The subtype A sequences also clustered according to when and where the viruses were isolated, whereas both the subtype B and C sequences remained as single clusters. Because the Pol protein has several functional domains, we identified the regions that are discriminative by comparing the structures of the domain-based networks. Our results suggest that sequence changes in the RNase H domain and the reverse transcriptase (RT) connection domain are responsible for the subtype classification. By analyzing the different amino acid compositions at each site in both domain sequences, we found that a few specific amino acid residues (i.e., M357 in the RT connection domain and Q480, Y483, and L491 in the RNase H domain) represent the differences among the subtypes. These residues were located on the surface of the RT structure and in the vicinity of the amino acid sites responsible for RT enzymatic activity or function.

Keywords: HIV-1, bioinformatics, pol protein, protein domain, network analysis, molecular evolution

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos - Philosophische Praxis, Austria

Reviewed by:
Hirotaka Ode,
Nagoya Medical Center (NHO), Japan
Masako Nomaguchi,
Tokushima University Graduate School of Medical Sciences, Japan

***Correspondence:**
Akio Kanai
akio@sfc.keio.ac.jp

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *08 August 2017*
Accepted: *20 October 2017*
Published: *02 November 2017*

Citation:
Nagata S, Imai J, Makino G, Tomita M and Kanai A (2017) Evolutionary Analysis of HIV-1 Pol Proteins Reveals Representative Residues for Viral Subtype Differentiation. Front. Microbiol. 8:2151. doi: 10.3389/fmicb.2017.02151

INTRODUCTION

Human Immunodeficiency Virus 1 (HIV-1) is a retrovirus, a specific type of RNA virus that has been widely used as a model system for studying the molecular evolution of life because it is highly adaptive and highly genetically diverse. HIV-1 has a single-stranded RNA genome and synthesizes double-stranded DNA based on its RNA genome using reverse transcriptase (RT), which is retained within the viral particle after it enters the target cell. The HIV-1 genome contains nine genes: *gag*, encoding the structural proteins involved in viral particle formation; *env*, encoding the envelope protein; *pol*, encoding the enzymes for replication (protease, RT, RNase H, integrase); *tat* and *rev*, involved in the regulation of gene expression; and *vif*, *vpr*, *vpu*, and *nef*, which are accessory genes required for optimal viral replication *in vivo*.

Molecular phylogenies have shown that HIV-1 arose in humans by cross-species infection from chimpanzees at the beginning of the twentieth century (Sharp and Hahn, 2010), and the infection has spread worldwide since the latter half of the twentieth century. This lineage, which is the predominant lineage throughout the world, is called group M and is classified into nine subtypes based on their phylogenetic relationships: subtypes A, B, C, D, F, G, H, J, and K. This genetic diversity is mainly attributed to an error-prone RT (Preston et al., 1988) and the genetic recombination mechanism of retroviruses (Hu and Temin, 1990). Recombination occurs frequently between the same subtypes or between different subtypes, and plays an important role in the diversification of HIV-1 (Rambaut et al., 2004).

The rates of disease progression and transmission differ according to the HIV-1 subtype involved, and it is thought that these differences contribute to differences in the prevalence and expansion of the subtypes. Several studies have reported that subtype D infections have a faster disease progression rate than subtype A infections (Kaleebu et al., 2002; Vasan et al., 2006; Baeten et al., 2007; Kiwanuka et al., 2008; Ng et al., 2013); the transmissibility of subtype C is greater than that of subtype A or D (Renjifo et al., 2004); the replication capacity of subtype C is lower than that of the other group M subtypes (Abraha et al., 2009; Kiguoya et al., 2017); and RT activity during replication differs between subtypes B and C (Armstrong et al., 2009; Iordanskiy et al., 2010). Several studies have also detected sequence differences in the HIV-1 proteases and the active N-terminal regions of RT and integrase (Gordon et al., 2003; Kantor et al., 2005; Rhee et al., 2006; Myers and Pillay, 2008). The Pol protein, which contains these regions, is thought to be associated with the differences in the replication capacity and disease progression of the different subtypes (Ng et al., 2013). However, the functional regions or amino acid residues in each viral protein that correspond to subtype differentiation have not been clarified.

Abbreviations: HIV-1, *Human immunodeficiency virus 1*; RT, reverse transcriptase; SSN, sequence similarity network; CRE, cumulative relative entropy.

The HIV-1 subtypes have usually been classified according to phylogenetic trees based on nucleotide or protein sequences of the viral core genes (*gag*, *pol*, and *env*; Castro-Nallar et al., 2012) and the clade relationships established (Robertson et al., 2000). Phylogenetic trees reflect the bifurcating phylogenetic relationships of sequences, but the construction of exact trees is difficult when the sequences contain intrasubtype recombinants, which occur frequently in HIV-1 (Posada et al., 2002; Arenas and Posada, 2010). Therefore, we constructed a sequence similarity network (SSN), a weighted undirected graph based on sequence similarities, to visualize the sequence space and observe the positional relationships among the subtypes in various regions of the HIV-1 genome.

Our aims were to systematically characterize the evolution of the HIV-1 subtypes, and to clarify the sequence regions that are responsible for the differentiation of the viral subtypes. In this study, we analyzed the mutation sites in 2,052 Pol proteins from HIV-1 group M using SSNs. Because the Pol protein is often used for group or subtype classification, we determined the overall positional relationships among the subtypes based on the Pol sequences. We then compared the structures of the domain-based networks to identify the regions that characterize the subtypes. The amino acid sites corresponding to the different subtypes were specified and mapped to the three-dimensional structure of the protein. We discuss the possible implications of these results in light of the kinds of regions that have changed during the adaptation of the virus in its spread throughout the world.

MATERIALS AND METHODS

Data Sources

The near-complete genome sequences and their attributions (sampling year, sampling region, and subtype/sub-subtype of the viral sequence) of HIV-1 group M subtypes A (which consists of sub-subtypes A1 and A2), B, and C were downloaded from the HIV Sequence Database at Los Alamos (http://www.hiv.lanl.gov, last accessed August 2014). We used these three subtypes because the genetic distances between their sequences are almost equivalent (Robertson et al., 2000) and the number of sequences registered in the database is large enough for our analysis. After the intersubtype recombinants and truncated sequences were excluded, 2,052 Pol protein sequences (144 subtype A; 1,528 subtype B; 330 subtype C) were obtained. The regional breakdown of the datasets is shown in **Table 1**.

Network Analysis Based on Sequence Similarities

The sequence similarity scores were calculated to construct a weighted undirected graph (SSN). The similarity scores (Basic Local Alignment Search Tool [BLAST] bit scores; Altschul et al., 1990) for all the HIV-1 Pol protein sequences were calculated with an all-against-all BLASTP (BLAST 2.2.31+) analysis (Altschul et al., 1997; Camacho et al., 2009), with a cut-off *E*-value of $\leq$ 1e−5. Using the BLAST bit scores,

TABLE 1 | Geographic regional breakdown of HIV-1 datasets used in this study.

Subtype	Number of sequences						Total
	Africa	Asia	Central and South America	Europe	North America	Oceania	
A	73 (1)	38 (1)	0	31	0	2	144 (2)
B	2	222	149	148	985	22	1,528
C	338	28	6	6	2	0	380

Subtype A can further be divided into sub-subtypes A1 and A2; the numbers in parentheses show the number of sub-subtype A2 sequences.

the sequence similarities were normalized to 0.0–1.0, with the following equation (Dufour et al., 2010; Matsui et al., 2013):

$$sim(x,y) = \frac{max\left(bit\ score\ (x,y),\ bit\ score\ (y,x)\right)}{max\left(bit\ score\ (x,x),\ bit\ score\ (y,y)\right)}$$

where sim(x,y) represents the normalized sequence similarity between two sequences x and y. If the score was 1.0, the pair was deemed to be identical. A weighted undirected graph was constructed based on the scores of all the pairs of sequences, and the edges were weighted with the scores. We set a threshold sequence identity value and connected the nodes when the sequence identity exceeded the threshold. The threshold to be used was determined by comparing the networks constructed with an incremental series of threshold values. We constructed SSNs of both the full-length Pol protein sequences and the functional domain sequences within the Pol protein. The constructed networks were visualized with Cytoscape 3.4.0 (Shannon et al., 2003), with a force-directed layout.

Clustering Based on the Network Structure

Spectral clustering (Paccanaro et al., 2006), a clustering method that divides data into clusters based on the structure of a network graph, was performed with SCPS 0.9.5 (Nepusz et al., 2010) for the networks constructed from the full-length Pol protein sequence. With this clustering algorithm, we analyzed the factors (sampling year, sampling region, and subtype) that affected the mutations in the Pol proteins by gradually changing the number of divisions.

Extraction of Functional Domain Sequences in the Pol Protein

Based on the HIV-1 group M subtype B reference strain HXB2 (GenBank accession: K03455), information on the functional domains was obtained from the Swiss-Prot database (http://www.uniprot.org/; The UniProt Consortium, 2017), a high-quality annotated protein sequence knowledgebase. The dataset sequences were then aligned to the HXB2 Pol sequence using MAFFT L-INS-i 7.245 (Katoh and Standley, 2013) to extract the functional domains. Note that the amino acid residues mentioned in this study are numbered according to the HXB2 sequence.

Construction of Phylogenetic Trees from Functional Domain Sequences of Pol Protein

We randomly selected 10 sequences from each subtype (subtype A, B, or C) to represent each functional domain sequence. A multiple-sequence alignment of each domain was created with MAFFT L-INS-i 7.245 (Katoh and Standley, 2013), and maximum likelihood phylogenetic trees were constructed with RAxML 8.2.9 (Stamatakis, 2014; GAMMA model with 1,000 bootstrap replicates). The calculated trees were visualized with FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/).

Calculation of Cumulative Relative Entropy (CRE)

To identify the sites that characterize the differences between each subtype of HIV-1, we calculated CRE (Hannenhalli and Russell, 2000) for each amino acid site in the Pol protein. We calculated the amino acid compositions of the three subtypes (A, B, and C) at each site in a multiple alignment. The *hmmbuild* program of HMMER 3.1b2 (Eddy, 1998) was used to build profile P of the alignment. The weighting method of Henikoff and Henikoff (1994) was used for the residue counts. The relative entropy (Shannon, 1996; Durbin et al., 1998) of position i for subtype $\bar{s}$ with respect to the entropy of that position for subtype was calculated. If RE_i^s is the relative entropy of P_i^s with respect to P_i:

$$RE_i^s = \sum_{for\ all\ x} P_{i,x}^s \log \frac{P_{i,x}^s}{P_{i,x}^{\bar{s}}}$$

Note that RE is greater than or equal to zero, and is exactly zero when the two distributions are identical (Durbin et al., 1998). To estimate the role of alignment position i in characterizing the HIV-1 subtypes, CRE_i was calculated as:

$$CRE_i = \sum_{for\ all\ subtypes\ s} RE_i^s$$

The CREs for all the positions were converted into Z-scores based on the distribution of the entropies within a sequence alignment. Let μ and σ be the means and standard deviations of the CREs of all positions, then the Z-score for position i is calculated as:

$$Z_i = \frac{CRE_i - \mu}{\sigma}$$

We expect a position with a high Z-score to be important in characterizing the subtypes. We calculated the CREs, and positions with Z-scores of CRE $\geq$ 3.0 were defined as "high-CRE" positions. All multiple sequence alignments used to calculate CRE were constructed with MAFFT L-INS-i 7.245 (Katoh and Standley, 2013). The generated alignments were visualized and sequence conservation was calculated with Jalview 2.9.0b2 (Waterhouse et al., 2009).

Protein Conformation Analysis of RT

The functional domains characterizing the subtypes and the high-CRE residues were mapped onto the structure of HIV-1 RT. The protein structural data (PDB ID: 1REV; 3KJV for RT complexed with the DNA duplex) were obtained from the Protein Data Bank (http://www.rcsb.org/pdb; Bernstein et al., 1977), a database of experimentally determined protein structures, and visualized with UCSF Chimera 1.10.2 (Pettersen et al., 2004).

RESULTS

Comparison of Thousands of HIV-1 Pol Sequences Based on a Network Analysis

The classification of and relationships between each HIV-1 subtype were determined by constructing networks based on the amino acid sequence similarities of the Pol polyprotein (Supplementary Figure 1 and **Figure 1**). The SSN is a graphical representation of the similarities between sequences. Each sequence is indicated by a point (node) and the similarity between the sequences is represented by the length of the line (edge) connecting the points. The smaller the distance between the nodes, the greater the degree of similarity between the sequences. We used subtypes A, B, and C from HIV-1 group M in the present analysis. Subtypes A, B, and C clearly form distinct groups when their sequence similarities are analyzed (Robertson et al., 2000) and more of these sequences are registered in the database than those of other subtypes, so we assumed that enough sequences were available for our purpose.

When constructing an SSN, the network structure changes according to the threshold value of the sequence identity used when connecting the edges (Fujishima et al., 2008). Therefore, we first constructed a series of networks by gradually changing the threshold value and compared their structures (Supplementary Figure 1). In networks in which the edges were connected with sequence identities $\geq$80%, the three subtypes of HIV-1 were not well-separated and formed one large network (Supplementary Figure 1A). When the sequence identity threshold was $\geq$92%, each of the three subtypes was properly separated. Because the nodes were still connected under this threshold, the relative positional relationships among subtypes were determined (Supplementary Figure 1B). However, as the threshold value became much stricter, the connections between the subtypes were broken, and with a sequence identity threshold of $\geq$96%, more than 130 graphs were generated, so that it was impossible to determine the exact positional relationships among the subtypes. Therefore, we adopted a threshold value for sequence identity of $\geq$92% for the subsequent analysis.

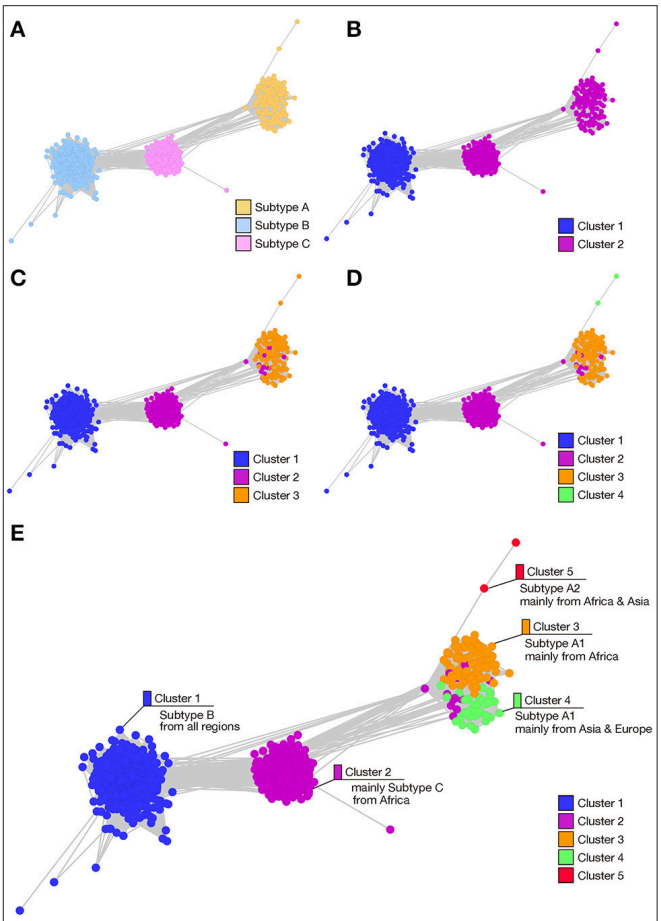


FIGURE 1 | Stepwise analysis of the SSNs of HIV-1 Pol protein. In total, 2,052 amino acid sequences of the HIV-1 Pol protein were used to construct networks based on sequence similarities (SSNs). Nodes (colored dots) represent the Pol protein sequences and the edge lengths represent the sequence similarities. **(A)** Nodes are colored according to the HIV-1 subtype. **(B–E)** To investigate the possible process of Pol protein evolution, a total of 2,052 Pol sequences, used to construct the network structure shown in panel A, were clustered into 2, 3, 4, or 5 groups with the spectral clustering method by changing the number of divisions in a sequential order, and are colored according to cluster. A stepwise analysis of the hierarchy of the clusters allowed the possible evolutionary pathways of the Pol proteins to be estimated. Panel **(E)** shows the attributes (subtype, region of sampling) of the sequences in each cluster.

In **Figure 1A**, the nodes are colored to represent the three subtypes shown in Supplementary Figure 1B, and provides an overall view of the sequence similarities between and/or within each subtype. To confirm the subtype groupings in the same SSN and to clarify the attributions obtained from the database (sampling year, sampling region, and viral sequence subtype), we used the spectral clustering method, which divides sequence data into clusters based on the structure of a network graph. This methodology enables the estimation of the process of subtype differentiation according to the order of cluster division

when dividing clusters step by step. Therefore, we changed the clustering division numbers to two clusters (**Figure 1B**), three clusters (**Figure 1C**), four clusters (**Figure 1D**), and five clusters (**Figure 1E**). At the two-cluster stage (**Figure 1B**), subtype B and subtypes A and C formed specific groups. Subtypes A and C were then differentiated into different groups in the three-cluster stage (**Figure 1C**). These evolutionary steps are consistent with the order of subtype differentiation reported in a previous study (Castro-Nallar et al., 2012), and the elements of the clusters and each subtype in the three-cluster stage showed a high coincidence ratio (99.4%), indicating that this network analysis is a suitable technique for classifying these subtypes (**Figures 1A,C**). When the number of divisions was set to 4, the sequence group corresponding to subtype A was further divided into two clusters, which exactly matched sub-subtypes A1 and A2, respectively (**Figure 1D**). With five clusters, subtype A1 was further divided into an additional two clusters, consisting mainly of the sequences from Africa or from Asia and Europe (**Figure 1E**).

Domain-Based Network Analysis Shows That RNase H Domain and RT Connection Domain Are Important for Subtype Differentiation

To analyze the regions in the HIV-1 Pol polyprotein that are responsible for HIV-1 subtype differentiation, we constructed SSNs based on each functional domain (**Figure 2**). **Figure 2A** shows the eight domains present in the HIV-1 Pol polyprotein (**a**, retroviral aspartyl protease, residues 61–153; **b**, RT (RNA-dependent DNA polymerase), residues 218–388; **c**, RT thumb domain, residues 396–458; **d**, RT connection domain, residues 473–573; **e**, RNase H, residues 591–711; **f**, integrase zinc-binding domain, residues 723–759; **g**, integrase core domain, residues 770–875; and **h**, integrase DNA binding domain, residues 936–982). When we compared the SSNs constructed for each domain, the nodes of the three subtypes were mixed in the networks of the RT (RNA-dependent DNA polymerase) region (**Figure 2Bb**) and the integrase DNA-binding domain

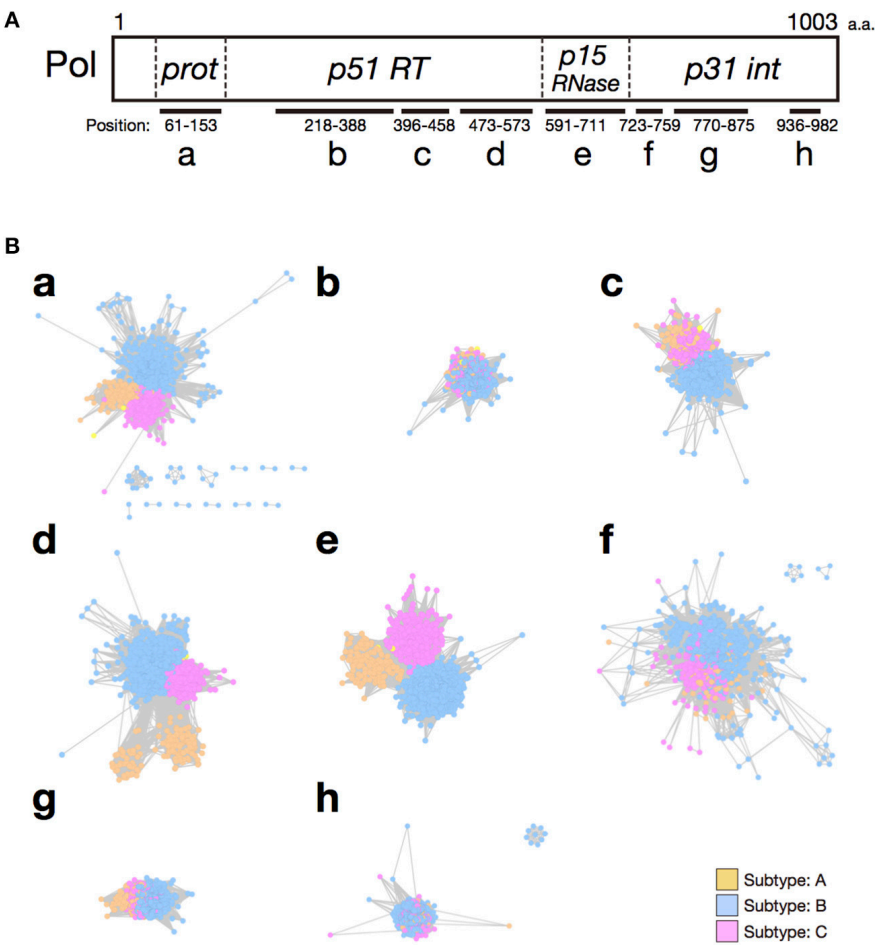


FIGURE 2 | Comparison of the SSNs of functional domains of the HIV-1 Pol protein. **(A)** Schematic representation of the positions and lengths of the functional domains of the HIV-1 Pol protein: **a**, retroviral aspartyl protease; **b**, RT (RNA-dependent DNA polymerase); **c**, RT thumb domain; **d**, RT connection domain; **e**, RNase H; **f**, integrase zinc-binding domain; **g**, integrase core domain; and **h**, integrase DNA-binding domain. **(B)** SSNs of each functional domain region of the HIV-1 Pol protein (2,052 sequences) were created and colored according to subtype. Nodes represent the sequences of each functional domain and the edge lengths represent the sequence similarities. Symbols **(a–h)** correspond to those in panel **(A)**.

(**Figure 2Bh**) because of the high sequence conservation in these regions. In the RT thumb domain (**Figure 2Bc**), the integrase zinc-binding domain (**Figure 2Bf**), and the integrase core domain (**Figure 2Bg**), the nodes for subtype A and C were mixed, although those for subtype B and subtypes A and C were separated. Therefore, we consider these regions unsuitable for distinguishing these subtypes. In contrast, the nodes of the three subtypes were well-separated in another three domains: the retroviral aspartyl protease region (**Figure 2Ba**), the RT connection domain (**Figure 2Bd**), and the RNase H domain (**Figure 2Be**). Among these, we consider the retroviral aspartyl protease region inappropriate for distinguishing the subtypes because the nodes are dispersed compared with those of the other two regions. Therefore, we conclude that the RT connection domain and the RNase H domain represent the differences among the HIV-1 subtypes. To observe effects of non-domain regions on subtype classification, we extracted five regions lying between the eight domains of the Pol protein (Supplementary Figures 2Ai–m) and constructed the corresponding SSNs (Supplementary Figure 2B). However, in these networks of non-domain regions, the boundaries of the three subtypes were indefinable. In particular, for the region shown in Supplementary Figure 2Ai, hundreds of graphs were generated, which did not provide clear indices for distinguishing the subtypes.

To confirm these results, we analyzed each selected domain phylogenetically, with the maximum likelihood method. The results supported our conclusions based on our domain-based SSNs (Supplementary Figure 3). In the phylogenetic trees for the RT connection domain (Supplementary Figure 3A) and the RNase H domain (Supplementary Figure 3B), branches of the three subtypes are clearly separated. However, in other domains, such as the RT (RNA-dependent DNA polymerase) domain (Supplementary Figure 3C) and integrase core domain (Supplementary Figure 3D), there are ambiguous boundaries among the nodes corresponding to each subtype. In particular, the nodes of the three subtypes are mixed in the clades for the integrase zinc-binding domain (Supplementary Figure 3E). Furthermore, to quantitatively identify the domains of the sequences that characterize the differences among subtypes, CRE was calculated for each site after the sequences were aligned. This index takes a large value when the difference between an amino acid distribution in one subtype and that in the other subtypes is large. The mean CRE value was calculated for each domain of the Pol protein. The regions with the highest average CRE values were the RT connection domain and the RNase H domain (**Figure 3**). This supports the results based on the network structures shown in **Figure 2B**. The mean CRE value for the RT connection domain was statistically significantly larger than those for the other regions.

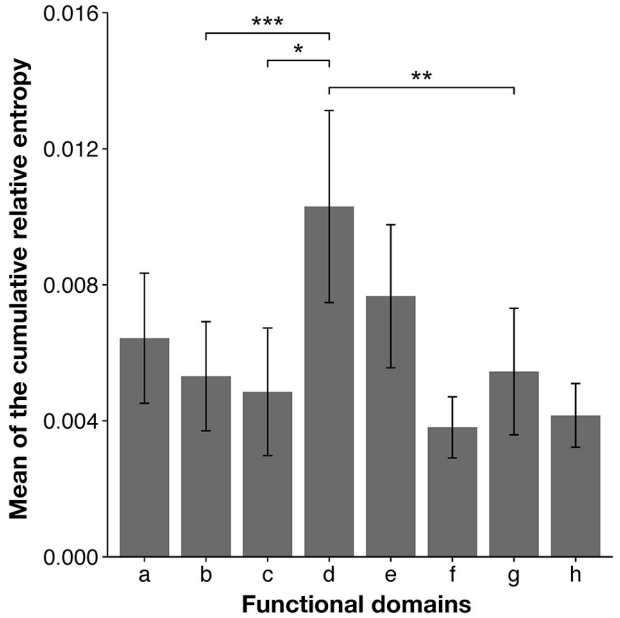


FIGURE 3 | Mean cumulative relative entropy (CRE) values for each functional domain of HIV-1 Pol protein. The x-axis represents each functional domain region, corresponding to those in **Figure 2A**: a, retroviral aspartyl protease; b, RT (RNA-dependent DNA polymerase); c, RT thumb domain; d, RT connection domain; e, RNase H; f, integrase zinc-binding domain; g, integrase core domain; and h, integrase DNA-binding domain. The y-axis represents the mean CRE value per site in each domain region. Groups were compared with the Mann–Whitney *U*-test (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). Error bars represent standard errors of the means and asterisks indicate the significance of differences between the RT connection domain or RNase H domain and each region.

Mapping the Amino Acid Residues That Are Crucial for Subtype Specification

To clarify the changes in the amino acid residues that distinguish the three subtypes, the CRE value was calculated for each site in both the RT connection domain and RNase H domain. The sites with CRE-derived Z-scores ≥ 3.0 (see section Materials and Methods) were then identified on an amino acid sequence alignments (**Figure 4**). High-CRE sites occurred at position 357 in the RT connection domain (**Figure 4A**) and at positions 480, 483, and 491 in the RNase H domain (**Figure 4B**). At position 357 (d1 site on **Figure 4A**) in the RT connection domain, lysine (K) occurs in 100% of subtype A sequences, methionine (M) in the majority (70%) of subtype B sequences, and methionine (M) and arginine (R) in about 50% each of the subtype C sequences, so the amino acid distribution at this site differs in the three subtypes. Meanwhile, in the RNase H domain, the amino acid compositions of subtypes B and C are similar at position 480 (e1), those of subtypes A and B are similar at position 483 (e2), and those of subtypes A and C are similar at position 491 (e3). In the RT connection domain, the subtypes can be distinguished at one site, whereas in the RNase H domain, the subtypes can be distinguished at three sites. These three sites are situated in close proximity in the RNase H domain. In terms of the sequence conservation in both domains, a region with low conservation is distributed throughout the domains, but these regions are not necessarily high-CRE regions. Therefore, we conclude that low sequence conservation does not generate the differences between the subtypes.

To localize the regions that characterize the subtypes on the tertiary structure of RT, the aforementioned two

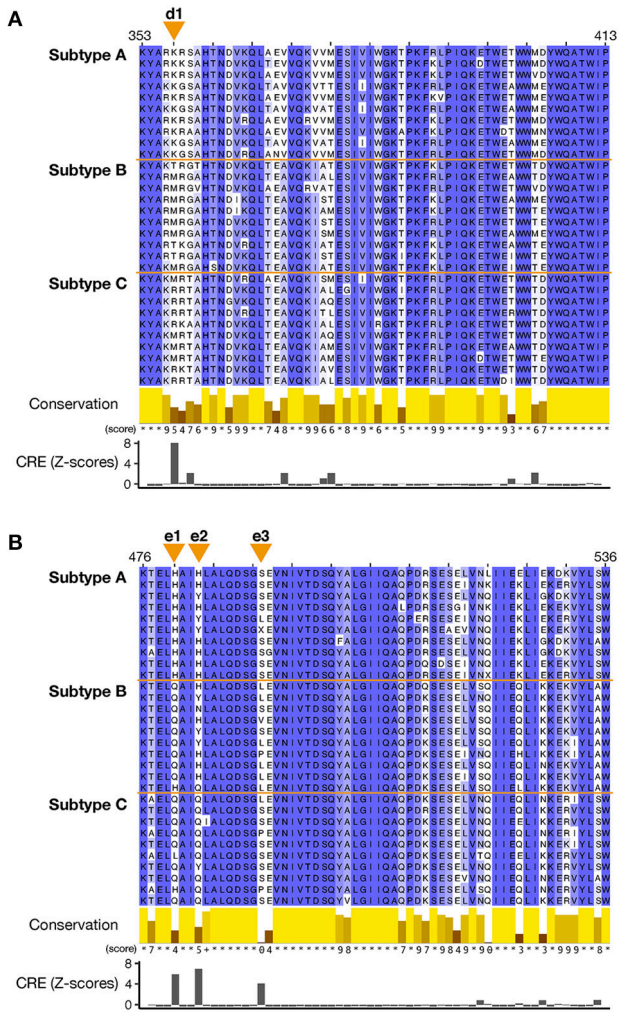


FIGURE 4 | Multiple amino acid sequence alignments of the RT connection domain and RNase H domain of the Pol protein. A total of 30 sequences, 10 sequences randomly selected from each subtypes (subtypes A, B, and C), were aligned with the MAFFT L-INS-i algorithm. **(A)** Sequence alignment around the sites in the RT connection domain that characterize the subtypes. **(B)** Sequence alignment around the sites in the RNase H domain that characterize the subtypes. For each amino acid position, the conservation scores according to Livingstone and Barton (1993) are shown as 12 ranks (from 0 to 11). Identical amino acids (rank 11) are indicated by asterisks, and partly conserved amino acids (rank 10) are indicated by a plus symbol. Orange triangles at the top of each alignment indicate the sites with CRE Z-scores ≥ 3.0 (designated d1 and e1-3). CRE was calculated based on all the sequences in the dataset. Sequence conservation scores and Z-scores of CRE for each site are shown together at the bottom of each alignment.

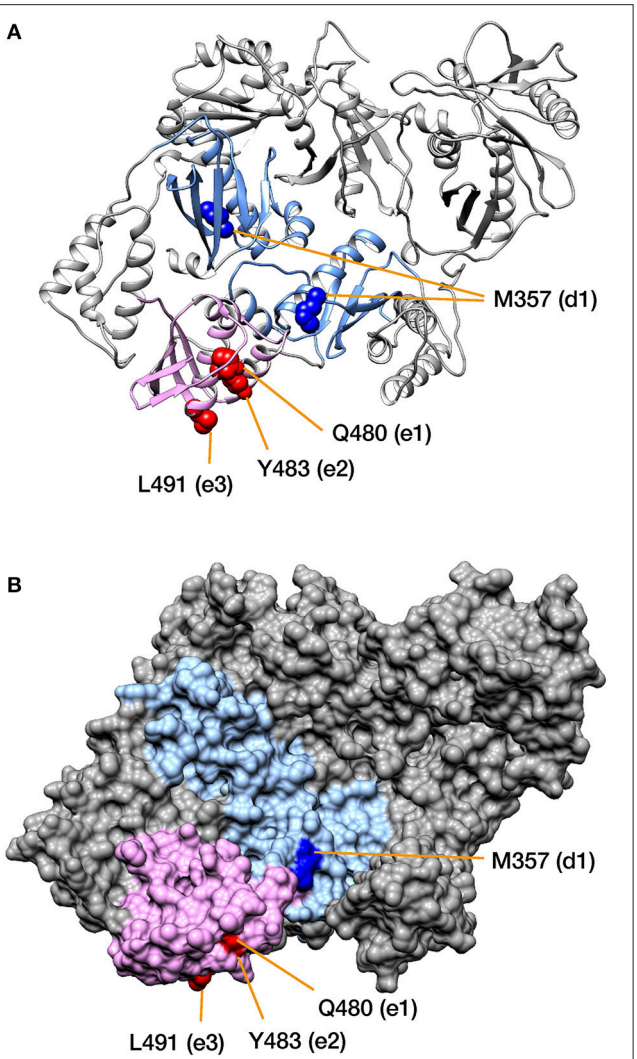


FIGURE 5 | Mapping the amino acid residues corresponding to the differences among HIV-1 subtypes onto RT. The structure of the HIV-1 RT p66/p51 heterodimer is shown in **(A)** as a ribbon diagram, and in **(B)** as a molecular surface representation. The RT connection domain (on p66 and p51) and the RNase H domain (on p66) are colored light blue and light pink, respectively. High-CRE positions d1 and e1-3 are colored blue and magenta, respectively (see also **Figure 4**). Note that the positions are represented on a space-filling model in panel **(A)**. PDB ID: 1REV.

domains and the high-CRE sites were mapped onto the three-dimensional conformation of HIV-1 RT (**Figure 5**). HIV-1 RT is a heterodimer, consisting of p66, which contains the RNase H domain, and p51 without the RNase H domain. The high-CRE sites M357 in the p66 chain of the RT connection domain and Q480, Y483, and L491 in the RNase H domain are physically close (**Figure 5A**). These amino acid residues were also located on the surface of RT in a representation of the molecular surface structure (**Figure 5B** and see section Discussion).

DISCUSSION

In this study, we successfully visualized the similarity relationships of thousands of sequences and summarized these data in SSNs. In the network constructed using the full-length amino acid sequence of the HIV-1 Pol protein, the sequence relationships of the three HIV-1 group M subtypes were visualized, and stepwise clustering divided the subtype A sequence into further clusters (**Figure 1**). This cluster division corresponded to the different regions in which the viruses were sampled. Therefore, although HIV-1 is prevalent throughout the world, sequence changes arise particularly frequently

in epidemic regions, and the network structure properly reflected these differences in their sequences. By comparing the network structures of each domain and the mean CRE values, we identified the RT connection domain and the RNase H domain of RT as characterizing the differences among the HIV-1 subtypes. By calculating the CREs for the amino acid sequences of the connection domain and the RNase H domain, we identified one amino acid residue in the connection domain and three residues in the RNase H domain as subtype-characterizing sites.

Using SSNs, we have previously determined the similarity relationships among a huge number of sequences with complex phylogenetic relationships and have discussed their evolution, including the bacterial CRP/FNR transcriptional regulator superfamily (Matsui et al., 2013) and the novel tRNA genes that have expanded in certain species of eukaryotes (Hamashima et al., 2015). In these reports, we estimated a possible evolutionary pathway by observing the phylogenies inferred from the stepwise analysis of cluster hierarchies. Although, those studies dealt mainly with the amino acid sequences of whole proteins or the nucleotide sequences of whole tRNA genes, in the present study, we constructed SSNs based on domain-level sequences for the first time. In **Figure 2Bd**, the RT connection domain of subtype A is divided further into two different groups. As described below, at least at the domain level, the difference in the RT activity corresponds to the difference between subtypes B and C. Thus, we speculated that there might be a difference in the RT enzymatic activity between these two groups of the RT connection domain of subtype A. Our results suggest that comparing the networks based on individual protein domains allows not only the detection of subtype differences, but also the functional divergence of the domains analyzed.

We excluded the retroviral aspartyl protease region from the analysis for the reasons described above (see section Domain-Based Network Analysis Shows That RNase H Domain and RT Connection Domain Are Important for Subtype Differentiation). However, comparisons of the network structures (**Figure 2**) and the mean CRE values (**Figure 3**) for each domain indicated that the retroviral aspartyl protease region is also a subtype-distinguishing region, in addition to the RT connection domain and RNase H domain. HIV-1 protease has been reported to have different activities and different target cleavage sites, predominantly in subtypes C and B (Velazquez-Campoy et al., 2001; de Oliveira et al., 2003). Together with the thumb domain, the RT connection domain forms a binding cleft and bridges both the N-terminal polymerase activity region and the C-terminal RNase H domain. RNase H is an enzyme that specifically degrades the RNA strand of DNA/RNA complexes during reverse transcription. It has been reported that differences in replication capacity between subtypes B and C are derived from the differences between the RT connection domain and the RNase H domain (Iordanskiy et al., 2010). This supported our current observations at least at the domain level. In addition, mutations in the connection domain and the RNase H domain are known to change the sensitivity of the virus to anti-HIV-1 drugs (RT inhibitors; Julias et al., 2003; Menéndez-Arias et al., 2011), and these are the same domains of the Pol protein that characterize the subtypes identified in this study.

At the amino acid level, the sites responsible for viral subtype differentiation do not perfectly match the drug-resistance mutations (Ehteshami and Götte, 2008; von Wyl et al., 2010; Menéndez-Arias et al., 2011), but are located very close to them in the RT structure (Supplementary Figures 4A,B). In particular, the drug-resistance mutations R356K, R358K, and A360V (Ehteshami and Götte, 2008; von Wyl et al., 2010; Menéndez-Arias et al., 2011) are located on the surface of the RT domain, close to M357, the position with the highest CRE in the connection domain. It is possible that mutations strongly associated with drug resistance are also closely related to the activity of RT, and the domain structure may be greatly altered and its activity reduced by a mutation that changes an amino acid to one with dissimilar biochemical properties. The three resistance mutations are conservative, including from arginine (R) to lysine (K) or from alanine (A) to valine (V). Therefore, we speculate that the HIV-1 subtypes were differentiated by the accumulation of mutations in the surface region where the sequence conservation is low, but at positions located very close to the critical amino acid residues required for enzymatic activity, changing the local structure and modulating the enzyme's activity. In contrast, RT is a heterodimer comprised of subunits p66 and p51. The larger p66 subunit contains the RNase H domain and the catalytic region with the main polymerase activities, whereas the smaller p51 subunit mainly plays a structural role (Telesnitsky and Goff, 1997). In this context, further analysis is required to identify the aspect (activity or structure) of the protein that most strongly affects subtype differentiation. Mutation Q509L, a drug-resistance mutation site in the RNase H domain (Ehteshami and Götte, 2008; Menéndez-Arias et al., 2011), is not physically close to any of the high-CRE sites (Supplementary Figure 4A). Instead, three amino acid residues critical for RNase H activity, D443, E478, and D498, are located together in the RNase H domain close to the high-CRE regions (Supplementary Figure 4C). Mutation E478, in particular, is located very close to Q480, one of three high-CRE sites. We also noted that in the structure of RT complexed with the DNA duplex, M357 is physically close to the DNA molecule but does not interact directly with it (Supplementary Figure S5). Again, these results suggest that the region that characterizes the subtypes is located in the vicinity of the amino acid site responsible for its enzyme activity or RT function.

By mapping the high-CRE sites onto sequence alignments, we found that locations with low sequence conservation do not necessarily characterize the differences in the subtypes (**Figure 4**). The amino acid residues at the high-CRE sites are conserved within each subtype, but differ between subtypes, so these sites are not fully conserved through all HIV-1 subtypes. This suggests that a region in which CRE is high can accommodate incoming mutations but has functional constraints that do not allow completely random mutations. We found that three HIV-1 subtypes can be distinguished by one amino acid residue (position 357) in the RT connection domain. However, each of the three amino acid residues (positions 480, 483, or 491) in the RNase H domain can be distinguished in only two of the three subtypes, indicating that all three amino acid

residues need to be considered to effectively classify the subtypes in this case. We cannot completely exclude the possibility that the accumulation of mutations in these subtype-characterizing regions of the HIV-1 Pol protein is caused by genetic drift. However, it is possible that these mutations are adaptions to the environment at places throughout the world in which HIV-1 is prevalent. The internal environments of various hosts are considered to differ among regions, based on race, the immune system, and the indigenous microbial flora. As seen in **Figure 1E**, HIV-1 even differs between regions in which the same subtype predominates in the populations. Therefore, we suggest that the virus does not mutate precisely in genomic regions encoding enzyme activities but in neighboring regions. This modulates the enzyme functions to allow the adaptation of the virus to geographic regional differences it encounters in areas of prevalence.

Many currently emerging viruses that cause pandemics throughout the world are RNA viruses characterized by high mutation rates, including *Ebola virus*, *Zika virus*, and others. Genome analyses have already shown that as viral infections spread, mutations accumulate in the viral genomic sequences, causing them to differ in different endemic areas (Simon-Loriere et al., 2015; Tong et al., 2015; Metsky et al., 2017). Our research provides a molecular basis for HIV-1 evolution and subtype differentiation, and should extend our understanding of the evolution and differentiation of other RNA viruses, including emerging viruses.

AUTHOR CONTRIBUTIONS

SN and AK conceived and designed the study, and SN wrote the manuscript. SN, JI, and GM performed the analyses and interpreted the data. AK and MT edited the manuscript. AK supervised the project. All the authors have read and approved the final manuscript.

FUNDING

This work was supported, in part, by research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan. The funding bodies played no role in the study design, data collection or analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Motomu Matsui, Dr. Haruo Suzuki, Dr. Yasuhiro Naito and Mr. Satoshi Tamaki for their constructive suggestions and discussions. We also thank all the members of the RNA Group at the Institute for Advanced Biosciences, Keio University, Japan, for their insightful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02151/full#supplementary-material

REFERENCES

Abraha, A., Nankya, I. L., Gibson, R., Demers, K., Tebit, D. M., Johnston, E., et al. (2009). CCR5- and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic. *J. Virol.* 83, 5592–5605. doi: 10.1128/JVI.02051-08

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Arenas, M., and Posada, D. (2010). The effect of recombination on the reconstruction of ancestral sequences. *Genetics* 184, 1133–1139. doi: 10.1534/genetics.109.113423

Armstrong, K. L., Lee, T.-H., and Essex, M. (2009). Replicative capacity differences of thymidine analog resistance mutations in subtype B and C human immunodeficiency virus type 1. *J. Virol.* 83, 4051–4059. doi: 10.1128/JVI.02645-08

Baeten, J. M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R. S., Certain, L., et al. (2007). HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. *J. Infect. Dis.* 195, 1177–1180. doi: 10.1086/512682

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E. Jr., Brice, M. D., Rodgers, J. R., et al. (1977). The protein data bank: a computer-based archival fole for macromolecular structures. *J. Mol. Biol.* 112, 535–542. doi: 10.1016/S0022-2836(77)80200-3

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Castro-Nallar, E., Pérez-Losada, M., Burton, G. F., and Crandall, K. A. (2012). The evolution of HIV: inferences using phylogenetics. *Mol. Phylogenet. Evol.* 62, 777–792. doi: 10.1016/j.ympev.2011.11.019

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucl. Acids Res.* 45, D158–D169. doi: 10.1093/nar/gkw1099

de Oliveira, T., Engelbrecht, S., Janse van Rensburg, E., Gordon, M., Bishop, K., Zur Megede, J., et al. (2003). Variability at human immunodeficiency virus type 1 subtype C protease cleavage sites: an indication of viral fitness? *J. Virol.* 77, 9422–9430. doi: 10.1128/JVI.77.17.9422-9430.2003

Dufour, Y. S., Kiley, P. J., and Donohue, T. J. (2010). Reconstruction of the core and extended regulons of global transcription factors. *PLoS Genet.* 6:e1001027. doi: 10.1371/journal.pgen.1001027

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Ehteshami, M., and Götte, M. (2008). Effects of mutations in the connection and RNase H domains of HIV-1 reverse transcriptase on drug susceptibility. *AIDS Rev.* 10, 224–235.

Fujishima, K., Sugahara, J., Tomita, M., and Kanai, A. (2008). Sequence evidence in the archaeal genomes that tRNAs emerged through the combination of ancestral genes as 5′ and 3′ tRNA halves. *PLoS ONE* 3:e1622. doi: 10.1371/journal.pone.0001622

Gordon, M., De Oliveira, T., Bishop, K., Coovadia, H. M., Madurai, L., Engelbrecht, S. , et al. (2003). Molecular characteristics of human immunodeficiency virus type 1 subtype C viruses from KwaZulu-Natal, South Africa: implications for vaccine and antiretroviral control strategies. *Society* 77, 2587–2599. doi: 10.1128/JVI.77.4.2587-2599.2003

Hamashima, K., Tomita, M., and Kanai, A. (2015). Expansion of noncanonical V-Arm-Containing tRNAs in eukaryotes. *Mol. Biol. Evol.* 33, 530–540. doi: 10.1093/molbev/msv253

Hannenhalli, S. S., and Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* 303, 61–76. doi: 10.1006/jmbi.2000.4036

Henikoff, S., and Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* 243, 574–578. doi: 10.1016/0022-2836(94)90032-9

Hu, W. S., and Temin, H. M. (1990). Retroviral recombination and reverse transcription. *Science* 250, 1227–1233. doi: 10.1126/science.1700865

Iordanskiy, S., Waltke, M., Feng, Y., and Wood, C. (2010). Subtype-associated differences in HIV-1 reverse transcription affect the viral replication. *Retrovirology* 7:85. doi: 10.1186/1742-4690-7-85

Julias, J. G., McWilliams, M. J., Sarafianos, S. G., Alvord, W. G., Arnold, E., and Hughes, S. H. (2003). Mutation of amino acids in the connection domain of human immunodeficiency virus type 1 reverse transcriptase that contact the template-primer affects RNase H activity. *J. Virol.* 77, 8548–8554. doi: 10.1128/JVI.77.15.8548-8554.2003

Kaleebu, P., French, N., Mahe, C., Yirrell, D., Watera, C., Lyagoba, F., et al. (2002). Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1–positive persons in uganda. *J. Infect. Dis.* 185, 1244–1250. doi: 10.1086/340130

Kantor, R., Katzenstein, D. A., Efron, B., Carvalho, A. P., Wynhoven, B., Cane, P., et al. (2005). Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. *PLoS Med.* 2:e112. doi: 10.1371/journal.pmed.0020112

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kiguoya, M. W., Mann, J. K., Chopera, D., Gounder, K., Lee, G. Q., Hunt, P. W., et al. (2017). Subtype-specific differences in gag-protease-driven replication capacity are consistent with inter-subtype differences in HIV-1 disease progression. *J. Virol.* 91:e00253-e17. doi: 10.1128/JVI.00253-17

Kiwanuka, N., Laeyendecker, O., Robb, M., Kigozi, G., Arroyo, M., McCutchan, F., et al. (2008). Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. *J. Infect. Dis.* 197, 707–713. doi: 10.1086/527416

Livingstone, C. D., and Barton, C. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of sequence conservation. *Cabios* 9, 745–756.

Matsui, M., Tomita, M., and Kanai, A. (2013). Comprehensive computational analysis of bacterial CRP/FNR superfamily and its target motifs reveals stepwise evolution of transcriptional networks. *Genome Biol. Evol.* 5, 267–282. doi: 10.1093/gbe/evt004

Menéndez-Arias, L., Betancor, G., and Matamoros, T. (2011). HIV-1 reverse transcriptase connection subdomain mutations involved in resistance to approved non-nucleoside inhibitors. *Antiviral Res.* 92, 139–149. doi: 10.1016/j.antiviral.2011.08.020

Metsky, H. C., Matranga, C. B., Wohl, S., Schaffner, S. F., Freije, C. A., Winnicki, S. M., et al. (2017). Zika virus evolution and spread in the Americas. *Nature* 546, 411–415. doi: 10.1038/nature22402

Myers, R. E., and Pillay, D. (2008). Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J. Virol.* 82, 9228–9235. doi: 10.1128/JVI.01535-07

Nepusz, T., Sasidharan, R., and Paccanaro, A. (2010). SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. *BMC Bioinformatics* 11:120. doi: 10.1186/1471-2105-11-120

Ng, O. T., Laeyendecker, O., Redd, A. D., Munshaw, S., Grabowski, M. K., Paquet, A. C., et al. (2013). HIV type 1 polymerase gene polymorphisms are associated with phenotypic differences in replication capacity and disease progression. *J. Infect. Dis.* 209, 66–73. doi: 10.1093/infdis/jit425

Paccanaro, A., Casbon, J. A., and Saqi, M. A. (2006). Spectral clustering of protein sequences. *Nucleic Acids Res.* 34, 1571–1580. doi: 10.1093/nar/gkj515

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera - a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Posada, D., Crandall, A. K., and Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* 54, 396–402. doi: 10.1007/s00239-001-0034-9

Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science* 242, 1168–1171. doi: 10.1126/science.2460924

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nat. Rev. Genet.* 5, 52–61. doi: 10.1038/nrg1246

Renjifo, B., Gilbert, P., Chaplin, B., Msamanga, G., Mwakagile, D., Fawzi, W., et al. (2004). Preferential in-utero transmission of HIV-1 subtype C as compared to HIV-1 subtype A or D. *AIDS* 18, 1629–1636. doi: 10.1097/01.aids.0000131392.68597.34

Rhee, S.-Y., Kantor, R., Katzenstein, D. A., Camacho, R., Morris, L., Sirivichayakul, S., et al. (2006). HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes. *AIDS* 20, 643–651. doi: 10.1097/01.aids.0000216363.36786.2b

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Shannon, C. E. (1996). The mathematical theory of communication. *MD Comput. Comput. Med. Pract.* 14, 306–317.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharp, P. M., and Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 2487–2494. doi: 10.1098/rstb.2010.0031

Simon-Loriere, E., Faye, O., Faye, O., Koivogui, L., Magassouba, N., Keita, S., et al. (2015). Distinct lineages of ebola virus in Guinea during the 2014 West African epidemic. *Nature* 524, 102–104. doi: 10.1038/nature14612

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Telesnitsky, A., and Goff, S. (1997). *Reverse Transcriptase and the Generation of Retroviral DNA*. New York, NY: Cold Spring Harbor Laboratory Press.

Tong, Y.-G., Shi, W.-F., Di, Liu, Qian, J., Liang, L., Bo, X.-C., et al. (2015). Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* 524, 93–96. doi: 10.1038/nature14490

Vasan, A., Renjifo, B., Hertzmark, E., Chaplin, B., Msamanga, G., Essex, M., et al. (2006). Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. *Clin. Infect. Dis.* 42, 843–852. doi: 10.1086/499952

Velazquez-Campoy, A., Todd, M. J., Vega, S., and Freire, E. (2001). Catalytic efficiency and vitality of HIV-1 proteases from African viral subtypes. *Proc. Natl. Acad. Sci. U.S.A.* 98, 6062–6067. doi: 10.1073/pnas.111152698

von Wyl, V., Ehteshami, M., Demeter, L. M., Bürgisser, P., Nijhuis, M., Symons, J., et al. (2010). HIV-1 reverse transcriptase connection domain mutations: dynamics of emergence and implications for success of combination antiretroviral therapy. *Clin. Infect. Dis.* 51, 620–628. doi: 10.1086/655764

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 March 2018
doi: 10.3389/fmicb.2018.00303

Integrated Lung and Tracheal mRNA-Seq and miRNA-Seq Analysis of Dogs With an Avian-Like H5N1 Canine Influenza Virus Infection

*Cheng Fu[1,2,3], Jie Luo[1,2,3], Shaotang Ye[1,2,3], Ziguo Yuan[1,2] and Shoujun Li[1,2,3]**

[1] College of Veterinary Medicine, South China Agricultural University, Guangzhou, China, [2] Guangdong Provincial Key Laboratory of Prevention and Control for Severe Clinical Animal Diseases, Guangzhou, China, [3] Guangdong Technological Engineering Research Center for Pet, Guangzhou, China

Avian-like H5N1 canine influenza virus (CIV) causes severe respiratory infections in dogs. However, the mechanism underlying H5N1 CIV infection in dogs is unknown. The present study aimed to identify differentially expressed miRNAs and mRNAs in the lungs and trachea in H5N1 CIV-infected dogs through a next-generation sequencing-based method. Eighteen 40-day-old beagles were inoculated intranasally with CIV, A/canine/01/Guangdong/2013 (H5N1) at a tissue culture infectious dose 50 ($TCID_{50}$) of 10^6, and lung and tracheal tissues were harvested at 3 and 7 d post-inoculation. The tissues were processed for miRNA and mRNA analysis. By means of miRNA-gene expression integrative negative analysis, we found miRNA–mRNA pairs. Lung and trachea tissues showed 138 and 135 negative miRNA–mRNA pairs, respectively. One hundred and twenty negative miRNA–mRNA pairs were found between the different tissues. In particular, pathways including the influenza A pathway, chemokine signaling pathways, and the PI3K-Akt signaling pathway were significantly enriched in all groups in responses to virus infection. Furthermore, dysregulation of miRNA and mRNA expression was observed in the respiratory tract of H5N1 CIV-infected dogs and notably, TLR4 (miR-146), NF-κB (miR-34c) and CCL5 (miR-335), CCL10 (miR-8908-5p), and GNGT2 (miR-122) were found to play important roles in regulating pathways that resist virus infection. To our knowledge, the present study is the first to analyze miRNA and mRNA expression in H5N1 CIV-infected dogs; furthermore, the present findings provide insights into the molecular mechanisms underlying influenza virus infection.

Keywords: canine influenza virus, H5N1, mRNA–miRNA integrate analysis, negative, KEGG

OPEN ACCESS

Edited by:
Akio Adachi,
Tokushima University, Japan

Reviewed by:
Luis Villarreal,
University of Califiornia, Irvine, United States
Meilin Jin,
Huazhong Agricultural University, China
Weifeng Shi,
Taishan Medical University, China
John Pasick,
Canadian Food Inspection Agency, Canada

***Correspondence:**
Shoujun Li
shoujunli@scau.edu.cn

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *13 October 2017*
Accepted: *09 February 2018*
Published: *05 March 2018*

Citation:
Fu C, Luo J, Ye S, Yuan Z and Li S (2018) Integrated Lung and Tracheal mRNA-Seq and miRNA-Seq Analysis of Dogs With an Avian-Like H5N1 Canine Influenza Virus Infection. Front. Microbiol. 9:303. doi: 10.3389/fmicb.2018.00303

INTRODUCTION

Although the natural host of influenza A virus is wild aquatic birds, the host barrier is not unbreakable, and the virus can be transmitted to other species, including dogs (Klenk, 2014). Canine influenza virus (CIV) had not been reported until 2004, and it was first discovered in racing Greyhounds in the United States. The first identified CIV was confirmed as subtype H3N8 by sequencing (Crawford et al., 2005). In addition, another subtype of CIV (H3N2) was isolated from sick farmed and pet dogs in South Korea (SK) in 2007 (Lee et al., 2009; Li et al., 2010). Moreover,

multiple subtypes of influenza A viruses are reported to infect dogs, including the deadly H5N1 avian influenza virus (Songserm et al., 2006b), influenza A (H1N1) pdm09 virus (Damiani et al., 2012; Su et al., 2014a,b; Yin et al., 2014), H10N8 avian influenza virus (Su et al., 2014b), avian-like H9N2 influenza A virus (Songserm et al., 2006a; Sun et al., 2013), avian-like CIV (Su S. et al., 2013), and H3N2 influenza virus with the PA gene of H9N2 avian influenza virus (Lee et al., 2016). Studies have reported successful experimental infection of CIV in guinea pigs and mice (Tu et al., 2009; Castleman et al., 2010). These findings indicate that dogs may represent a new bridging species for avian and human influenza viruses for interspecies transmission.

In October 2004, highly pathogenic avian influenza (HPAI) H5N1 virus was reported for the first time in Thailand in a dog with severe lung congestion and edema and bloody nasal discharge (Songserm et al., 2006b). Furthermore, previous studies have reported that H5N1 influenza virus infections in dogs have resulted in anorexia, fever, conjunctivitis, labored breathing, and coughing (Songserm et al., 2006b; Maas et al., 2007; Chen et al., 2010; Ashour et al., 2012). Being considered "man's best friend," dogs drastically influence human lives. However, influenza viruses are susceptible to mutation; hence, it is important to understand the pathogenesis of H5N1 influenza virus infection among dogs.

MiRNAs are small non-coding RNAs. MiRNA binds on target mRNA to negatively regulate biological processes such as differentiation (Shivdasani, 2006), proliferation (Song et al., 2010), growth (Suh et al., 2015), metabolism (Gomez et al., 2014; Meydan et al., 2016), and apoptosis (Cheng, 2005). Increasing evidence has shown that deregulation of miRNA contributes to antiviral activity during an influenza virus infection. MiR-342-5p participates in macrophage interferon (IFN) antiviral responses against multiple viruses including influenza A (H1N1) (Ploegh et al., 2016). MiR-146a up-regulation significantly decreases H1N1 and H3N2 viral propagation (Terrier et al., 2013). MiR-144 was reported to post-transcriptionally decrease *TRAF6* levels to attenuate the antiviral response (Lopez et al., 2017). MiR-485 binds *PB1* transcripts in H5N1 to inhibit viral replication (Ingle et al., 2015). These findings suggest that certain miRNAs may play an important role in regulating influenza virus infections in dogs.

Strikingly, recent studies have assumed that miRNAs not only develop functions but also transmit from one species to another, thereby promoting crosstalk (Zhang et al., 2016) and interfering in signal transmission and communication (Hansen et al., 2010). Numerous studies have reported that viruses adapt their own miRNAs based on host miRNA (Pfeffer et al., 2004; Cai et al., 2005; Pfeffer et al., 2005) and establish an environment conducive to viral replication (Grey et al., 2007; Samols et al., 2007; Stern-Ginossar et al., 2007; Choy et al., 2008; Murphy et al., 2008; Nachmani et al., 2009). Hence, understanding the mechanism underlying viral miRNA-mediated adaptations can further our knowledge of the cross-species communication.

When infected with a virus, animals try to defend themselves with transcriptional reprograming of the affected cells. In this process, several genes are key elements and these genes are regulated by miRNAs (Peng et al., 2011; de Cubas et al., 2013). Recently, next-generation sequencing (NGS) technology has been used to obtain comprehensive sequencing data, which was used to detect and study the miRNA and protein expression levels in dogs with influenza virus infection (Zhao et al., 2014; Su et al., 2015). However, no detailed analysis of miRNA and the mRNA transcriptome is available. In this study, we used miRNA and mRNA profiles to perform a deep analysis of critical genes, miRNAs, and pathways related to virus infections.

MATERIALS AND METHODS

Sample Collection and RNA Isolation

Canine influenza virus, A/canine/01/Guangdong/2013 (H5N1), was isolated in 2013 from a dog with severe respiratory symptoms. Eighteen 40-day-old beagles were assigned to experimental and control groups. Dogs were housed in the Laboratory Animal Center of South China Agricultural University with number SYXK (YUE) 2014-0136. All study protocols were approved by the ABSL-3 Committee of South China Agricultural University in this study. A hemagglutination inhibition (HI) assay revealed that these dogs were sero-negative for avian-origin CIV (H3N2 and H5N1) and for H1N1, H3N1, and influenza B viruses of seasonal influenza viruses. Nine beagles were randomly divided into each group, i.e., experimentally infected (I) and non-infected (NI). After beagles were anesthetized with tiletamine–zolazepam (Virbac, 10–15 mg/kg), they were inoculated intranasally with H5N1 CIV at a tissue culture infectious dose 50 ($TCID_{50}$) 10^6 and the control group similarly received 1.0 ml of phosphate-buffered saline. Nasal samples were collected from all beagles before infection and continuously for 14 d after infection. At 3 and 7 d post-infection (dpi), three beagles from each group were euthanized through a pentobarbital overdose. Lesions in the lungs and trachea were collected and frozen in liquid nitrogen. One section was immediately used for RNA isolation, and the others were stored at −80°C for further use. An RNA library was generated for each group from the total RNAs collected from three dogs. Total RNAs were isolated using TRIZOL (Takara, Otsu, Japan) in accordance with the manufacturer's protocol. The concentration and purity of RNA were determined by measuring absorbance at 260 nm and the A_{260}/A_{280} ratio using a microspectrophotometer (Nanophometer, Germany). RNA samples were stored at −80°C until further use.

Ethics Statement

All procedures in the animal experiments were approved by the South China Agricultural University Experimental Animal Welfare Ethics Committee with a reference number of 2016-07.

RNA Sequencing and Data Analysis

Eight total RNA samples were obtained to generate RNA libraries for each sample. After the samples were qualified, using mRNA Capture Beads Enriched eukaryote mRNA, mRNA was fragmented by heating. These short mRNA and random hexamers were used to generate the first cDNA and then the second cDNA was synthesized. The second cDNA was purified using VAHTSTM DNA Clean Beads (Vazyme, Nanjing, China)

and then ligated to sequencing adapters. The fragments were amplified by using polymerase chain reaction (PCR) and purified using VAHTSTM DNA Clean Beads and then sequenced using an Illumina HiSeq (Vazyme, Nanjing, China). Raw sequence data were assessed, and sequences containing adaptor tags and those of low quality were excluded. Filtered reads were used for subsequent analysis, and the unique reads were used to identify differentially expressed genes (DEGs) with $|\log2\text{Ratio}| \geq 1$ and q-value $|\text{FDR}| \leq 0.05$.

Small RNA Sequencing and Data Analysis

Eight RNA libraries were generated with total RNA from samples. Total RNA was extracted and different fragments of RNAs by polyacrylamide gel electrophoresis (PAGE) were separated. Polyacrylamide electrophoresis gels were used to purify fragments that were 18–30 nt in length and 5′- and 3′-ends adaptors were ligated. The PCR products were generated after reverse-transcription (RT)-PCR and isolated using PAGE. Then, Illumina HiSeq 2000 (Vazyme, Nanjing, China) was used to sequence the purified cDNAs. After excluding reads with 3′- and 5′-primer contaminants or a poly(A) tails shorter than 18 nt, those with low-quality or those without the insert tag, were compared to clean reads in databases to annotate all known small RNA sequences. The unannotated sequences were searched against known miRNA precursors and mature miRNAs identified as known miRNAs. Differentially expressed (DE) miRNAs between the different samples were measured by $|\log2\text{Ratio}| \geq 1$ and q-value $|\text{FDR}| \leq 0.05$.

MiRNA–mRNA Integrative Genomic Analysis

To elucidate the interaction network of miRNA–mRNA with positive and negative correlations, we constructed an miRNA–mRNA regulatory network. The DE mRNAs and miRNAs were collected from the mRNA list and miRNA list. Then, we used miRanda[1] and TargetScan[2] to confirm the relationship between miRNA and mRNA. Finally, the relationship between miRNA and mRNA was verified.

Functional Analysis

Depending on the miRNA–mRNA integrative genomic analysis, we used Gene Ontology (GO) to identify biological themes for each negatively correlated miRNA–mRNA pair. There are three ontologies in GO: molecular function, cellular component, and biological process. Negative miRNA–mRNA correlations were imported into the KEGG database[3] for pathway analysis.

Real-Time qPCR

cDNA was synthesized with oligo (dt) primer for mRNA, using PrimeScript^{TM} RT Master Mix (Takara, Otsu, Japan). qPCR was performed using the SYBR Premix Ex Taq^{TM} (Tli RNaseH Plus) (Takara, Otsu, Japan) on the LC480 Real-Time PCR System (Roche, Basel, Switzerland) in accordance with the manufacturer's instructions (the primers are listed in Supplementary Tables 1, 2). cDNA was synthesized with A tail for miRNA, miRcute Plus miRNA Firsr-Strand cDNA Synthesis Kit (Tiangen, Beijing, China), and qPCR was performed using miRcute Plus miRNA qPCR Detection Kit (Tiangen, Beijing, China) on the LC480 Real-Time PCR System (Roche, Basel, Switzerland) in accordance with the manufacturer's instructions. GAPDH was used as an endogenous control gene for mRNA and U6 was used for miRNA. We used the $2^{-\Delta\Delta C_T}$ method to analyze these data. qPCR in each reaction was performed in triplicate, and the data were expressed as the mean ± standard error ($n = 3$).

RESULTS

Experiment With Dogs Infected With H5N1 Influenza Virus

Following infection with H5N1 influenza virus, clinical symptoms were observed, including cough, running nose, and an increased temperature. After infection with the virus, nasal swabs were collected from 1 to 14 dpi, and the peak temperature was 40.6°C at the second day after infection. The highest virus titer of a nasal swab was $10^{4.56}$. Virus replication in lung and trachea tissues was detected at 3 and 7 dpi, and the mean viral titers of 3 and 7 dpi were $10^{5.6}$ and $10^{2.16}$ $TCID_{50}$/ml in lung tissues, respectively, whereas it was $10^{1.3}$ and $10^{0.5}$ $TCID_{50}$/ml, respectively, in the trachea.

Overview of the Transcriptome and miRNAome

Gene libraries and miRNA libraries

Eight gene libraries as shown in Supplementary Data Sheet 1 were collected to identify mRNA differentiation of beagles when infected with H5N1 influenza virus. The beagles in the control group (KL3, KT3, KL7, and KT7) and experiment group (LUN3, TRA3, LUN7, and TRA7) were represented in the eight libraries. The original results of sequencing data and assembling results are shown in Supplementary Tables 3, 4. Among all of these reads, >90% of the clean reads were mapped to the canine reference genome.

Total RNA from the lung and trachea were used to build small RNA libraries to assess the characteristics of miRNAs in the lung and trachea after infection. After removing the contaminant and adaptor sequences and filtering low-quality tags, 10.5–12.4 million clean reads were collected in eight samples (as shown in Supplementary Table 5). Clean reads included rRNAs, tRNAs, snRNAs, and snoRNAs. After other RNAs were removed, 291, 290, 290, 283, 280, 287, 282, and 288 mature miRNAs were found in LUN3, TR3, LUN7, TR7, KL3, KT3, KL7, and KT7, respectively.

Differentially Expressed miRNAs and Analysis

A total of 455 miRNAs were identified. The DE miRNAs were identified through $|\log2\text{Ratio}| \geq 1$ and q-value $|\text{FDR}| \leq 0.05$ (as shown in **Figure 1A**). Compared with NI tissues, H5N1-infected tissues displayed differential expression of 19 mature miRNAs

[1]http://34.236.212.39/microrna/home.do

[2]http://www.targetscan.org

[3]http://www.genome.jp/kegg/

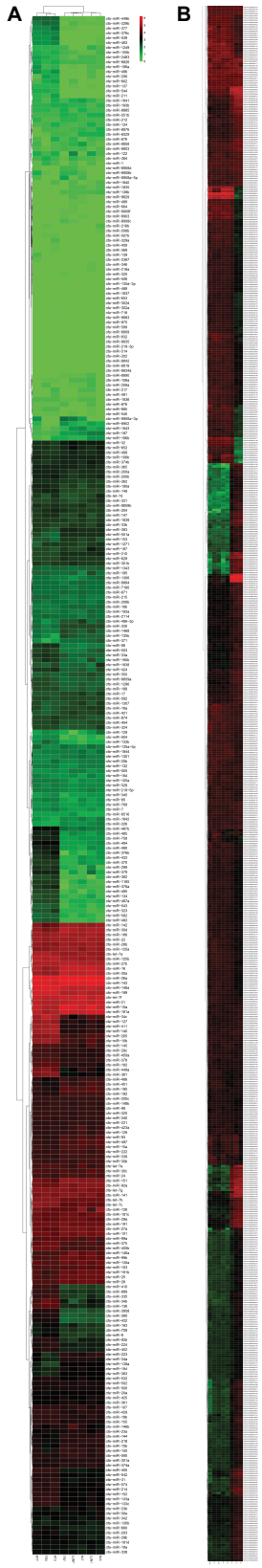


FIGURE 1 | Continued

FIGURE 1 | (A,B) Hierarchical clustering of DE miRNAs and DE mRNAs among eight RNA libraries. Heatmap of the number of DE miRNAs and DE mRNAs libraries for the DEGs between the infected groups and non-infected groups. MiRNA and mRNA heatmaps of all DE miRNAs and mRNAs are included.

(as shown in **Figure 2**). Only miR-126 was DE on day 3 in lung tissue, miR-126 has been reported to restrict the replication of H5N1 in endothelial cells to ameliorate disease condition (Sant et al., 2017), probably implying that when infected with the virus, miR-126 may play an important role in restricting viral replication. MiR-34c, miR-146b, and miR-34b were DE on day 7 in lung tissue, playing anti-inflammatory (miR-146b) (Comer et al., 2014) and resistant apoptosis (miR-34) (Catuogno et al., 2012) roles at this stage. On the seventh day, the mi-34 family is predicted to protect the lung from undergoing apoptosis to restore lung function to normal. Fifteen DE miRNAs were identified on the seventh day in the tracheal tissue (as shown in Supplementary Table 6). At this stage, miRNAs may restore the state of the organism to normal; these included miR-379, which inhibited cell proliferation invasion (Li et al., 2017), anti-inflammatory-related miR-127 (Zhang et al., 2017), and apoptosis-related miR-410 (Palumbo et al., 2016; Deng H. et al., 2017).

To determine whether differential miRNA expression was related to differences in post-infection stages, four comparisons were made. Within the tissues, 24 and 132 different miRNAs were identified from lung and tracheal tissues at different times, respectively. Furthermore, 140 and 12 miRNAs were DE at 3 and 7 dpi, respectively (Supplementary Table 7). There were more up-regulated miRNAs (15 of 24 in the lung and 90 of 132 in the trachea) than down-regulated miRNAs in the lung and trachea (**Figure 3**). Furthermore, three miRNAs, miR-379, miR-34c, and miR-34b, were DE at different times (**Figure 4**). These results suggest that when infected with H5N1 influenza virus, miR-379 and miR-34 probably play significant roles in resisting viral invasion.

Identification and Analysis of Differentially Expressed Genes

Compared with the control group, 1236 significant DEGs (**Figure 1B**) were identified. There were more down-regulated genes than up-regulated genes (**Figures 5A,B**). Among the down-regulated genes, 7 were found in the lung at different times, only 1 was found in the trachea at a different time, and 6 and 39 were found at 3 and 7 dpi, respectively; these included immune-related genes *IL2R2* (Schliemann et al., 2008), *CLEC4D* (Steichen et al., 2013), *IGFBP2* (Ambrosini-Spaltro et al., 2011), and *GPNMB* (Schwarzbich et al., 2011) and antiviral genes *FOSB* (Baumann et al., 2003) and *TNIP3* (Zhao et al., 2017).

KEGG Pathway Enrichment Analysis of Differentially Expressed (DE) mRNAs

To explore the mechanisms underlying resistance to influenza virus infection in dogs, three groups were formed. We identified 919 and 2314 significant DEGs from lung and tracheal tissues,

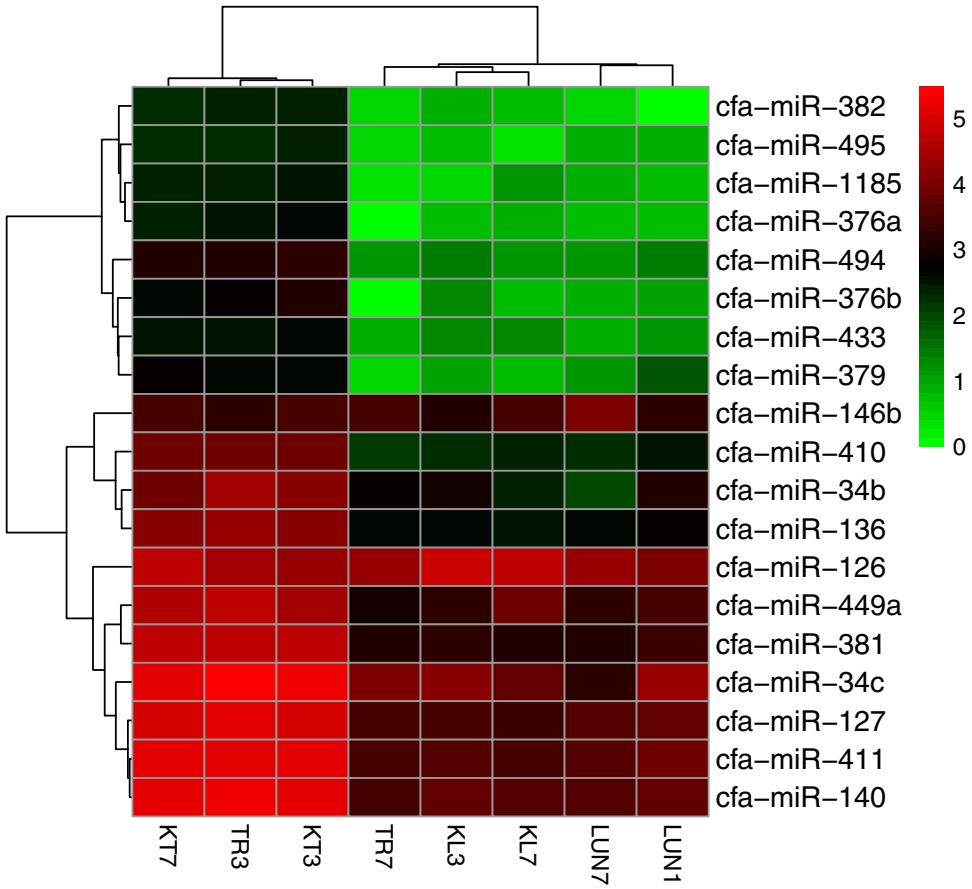


FIGURE 2 | Clustering of the 19 up-regulated and down-regulated DE miRNAs for assigning dogs to infected groups and non-infected groups. The colors in the heat map represent the normalized expression values, with lower expression values indicated in shades of green and higher expression values in shades of red.

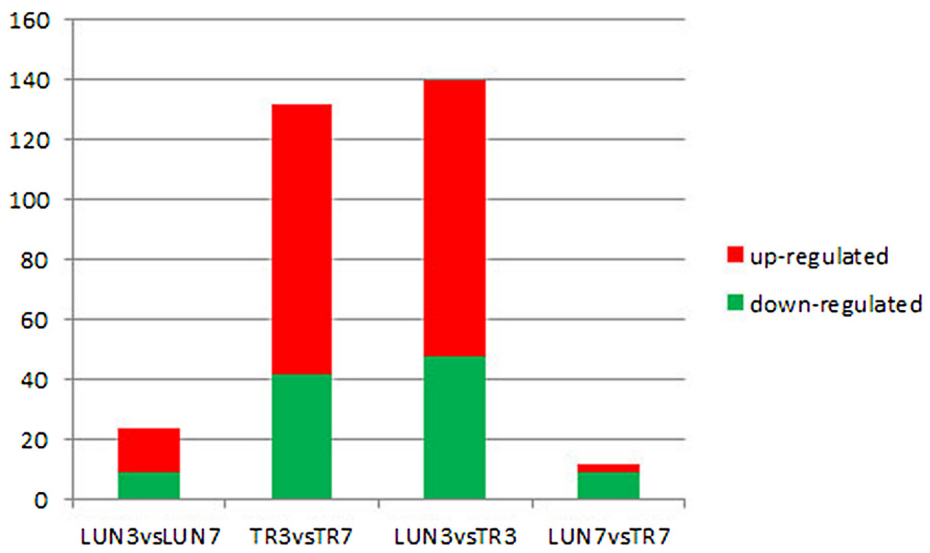


FIGURE 3 | Four comparisons of different post-infection stages, miRNAs showing a differential expression pattern, with down-regulation indicated in green and up-regulation indicated in red.

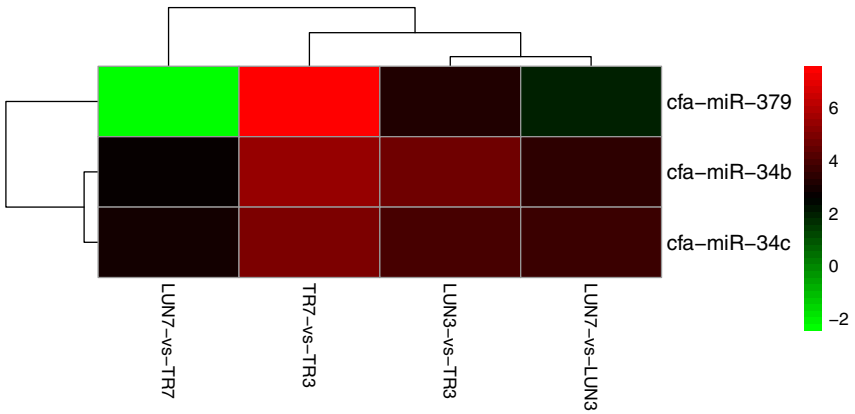


FIGURE 4 | MiR-379, miR-34c, and miR-34b were DE at all different post-infection stages in four comparisons (LUN7 vs. TR7, TR7 vs. TR3, LUN3 vs. TR3, and LUN7 vs. LUN3).

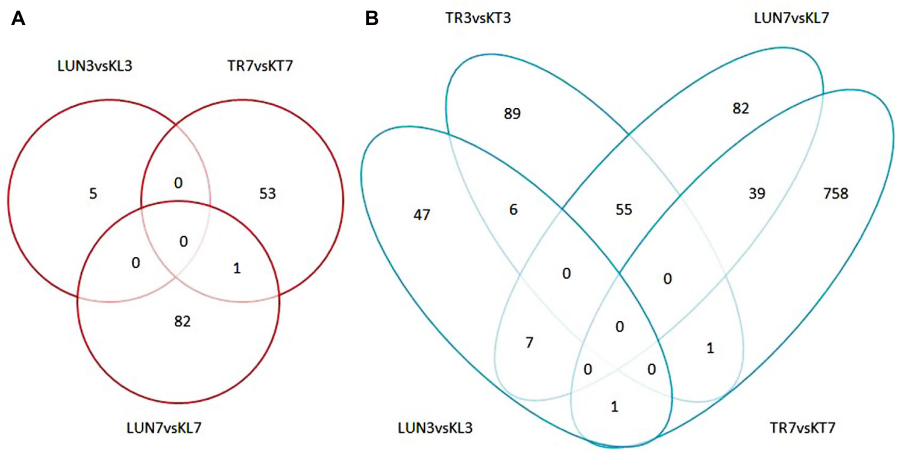


FIGURE 5 | (A) Numbers of up-regulated DEGs in infected groups and non-infected groups. **(B)** Numbers of down-regulated DEGs in infected groups and non-infected groups. An FDR of 0.05 was used to classify genes as DE. The lung and tracheal tissues from beagles in the control group, named KL3 and KT3 on the third day and KL7 and KT7 on the seventh day, and the experimental group named LUN3 and TRA3 on the third day and LUN7 and TRA7 on the seventh day.

respectively; furthermore, DE genes were also identified under the infected state in the lung and tracheal tissues; 2411 DEGs were collected under the infected state in different tissues. These data were imported into KEGG databases. Significant (p-value < 0.05) pathways were identified (**Figure 6** X2, X4, and X6). The signaling pathway and immune related were the two most represented subunits among these groups. The chemokine signaling pathway and hematopoietic cell lineage belonged to the immune system, and all participated in the regulation of these groups. All were involved in the modulation of infected CIV signal pathways including cytokine–cytokine receptor interactions, the PI3K-Akt signaling pathway, cell adhesion molecules (CAMs), and ECM–receptor interactions. Apart from these, infectious diseases were another subclass pathway. Furthermore, influenza A was in the enriched pathway in all of the groups. These results suggested that genes in these pathways may play an important role in response to influenza virus.

Integrative Genomic Analysis Associated With Negative miRNA–mRNA

To further investigate miRNA–mRNA regulatory information in dogs, transcriptome analyses were performed to identify genes that were co-expressed with miRNAs and miRNA targets were also predicted to identify genes that were bound to miRNA. A total of 268 and 251 miRNA–mRNA pairs with both positive and negative correlations, respectively, were identified in the lung and tracheal tissues, respectively; moreover, under the infected state, miRNA–mRNA pairs were also different between lung and tracheal tissues, and 255 miRNA–mRNA pairs were identified under the infected state in different tissues. However, miRNA generally negatively targeted genes. Therefore, when miRNAs are induced by the influenza virus, their target mRNAs are down-regulated and vice versa. There were 138 negative miRNA–mRNA interactions with 24 mature miRNAs and 201 validated mRNAs were DE in the lung. A total of 134 negative

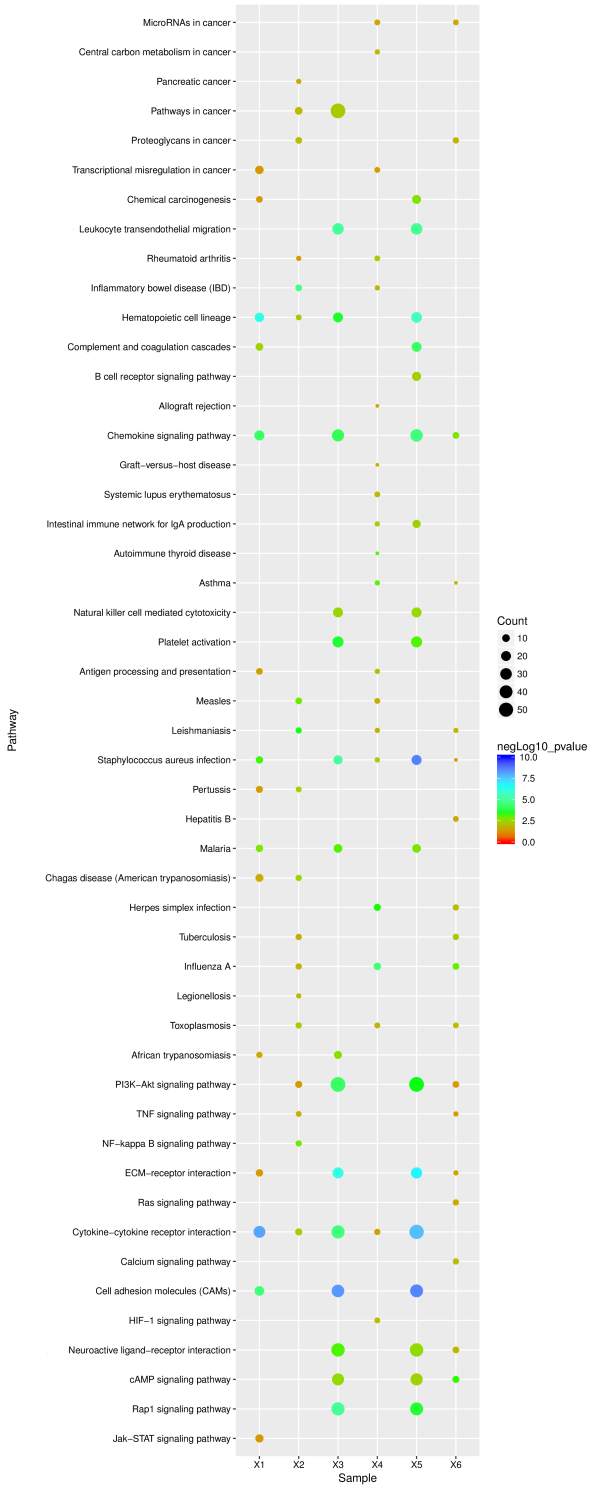


FIGURE 6 | KEGG pathway enrichment analysis of DE unigenes of negative miRNA–mRNA and DE unigenes of RNA-seq. X1, 138 negative miRNA–mRNA interactions with 24 mature miRNAs and 201 validated mRNAs DE in the lung (15 pathways); X2, 916 DE unigenes of RNA-seq in lung tissue (18 pathways); X3, 134 negative miRNA–mRNA interactions with 97 mature miRNAs and 135 mRNAs were found in the trachea (16 pathways); *(Continued)*

FIGURE 6 | Continued
X4, 2314 DE unigenes of RNA-seq in trachea tissue (20 pathways); X5, 65 mature miRNAs and 78 mRNA with 120 negative miRNA–mRNA interactions were identified under the infected state with different tissues (18 pathways); X6, 2411 DE unigenes of RNA-seq were identified under the infected state with different tissues (pathways). The ordinate represents the pathway name, the abscissa represents the sample name, the size of the dots indicates the number of DEGs in this pathway, and the point colors correspond to different negLog10 *p*-value ranges.

miRNA–mRNA interacted with 97 mature miRNAs and 135 mRNAs were found in the trachea. Moreover, 65 mature miRNAs and 78 mRNAs with 120 negative miRNA–mRNA interactions were identified under the infected state in different tissues. All target genes of the miRNAs were predicted using miRanda and TargetScan.

Functional Annotation and Pathways Affected by Relevant Negative miRNA–mRNA Correlations in Beagles

To identify enriched functional terms of these predicted target genes and to further explore the significant negatively correlated miRNA–mRNA pairs, a GO analysis was carried out on these 105 mRNAs in the lung, 84 mRNAs in the trachea, and 78 mRNAs in different infected tissues, which were up-regulated and down-regulated. GO enrichment analysis was involved in biological processes, cellular components, and molecular functions.

For the 109 mRNAs in the lung, 15 immune-related GO terms in biological process were significantly enriched ($p < 0.05$) (**Figure 7A**), and immune response was the highest fold enrichment (eightfold). Inflammatory responses, transcription, and positive regulation of fever generation were also identified. Furthermore, 17 immune-related GO terms in biological processes were significantly enriched ($p < 0.05$) in the trachea of 84 mRNAs (**Figure 7B**). Inflammatory responses and innate immune responses had the highest fold enrichment (ninefold). In addition, an infected state with different tissues was associated with 10 immune-related GO terms in biological processes (**Figure 7C**) including positive regulation of cytosolic calcium ion concentrations, positive regulation of I-kappaB kinase/NF-κB signaling, and positive regulation of the Toll-like receptor (TLR) 2 signaling pathway ($p < 0.05$). *In vivo*, different genes coordinate each other to perform biological functions. Pathway analysis helps to gain a better understanding of the biological function of genes.

Pathway enrichment analysis for 109 mRNAs of 138 negative miRNA–mRNA pairs in the lung, 84 mRNAs of 134 negative pairs in the trachea, and 78 mRNAs of 120 negative miRNA–mRNA pairs in different infected tissues. Further analysis found that among these significant ($p < 0.05$) pathways, signal- and immune-related pathways were also the most in all of these groups. In addition, infectious diseases were also another enrichment class of the pathways, and the influenza A pathway was also enriched in all groups (**Figure 6** X1, X3, and X5).

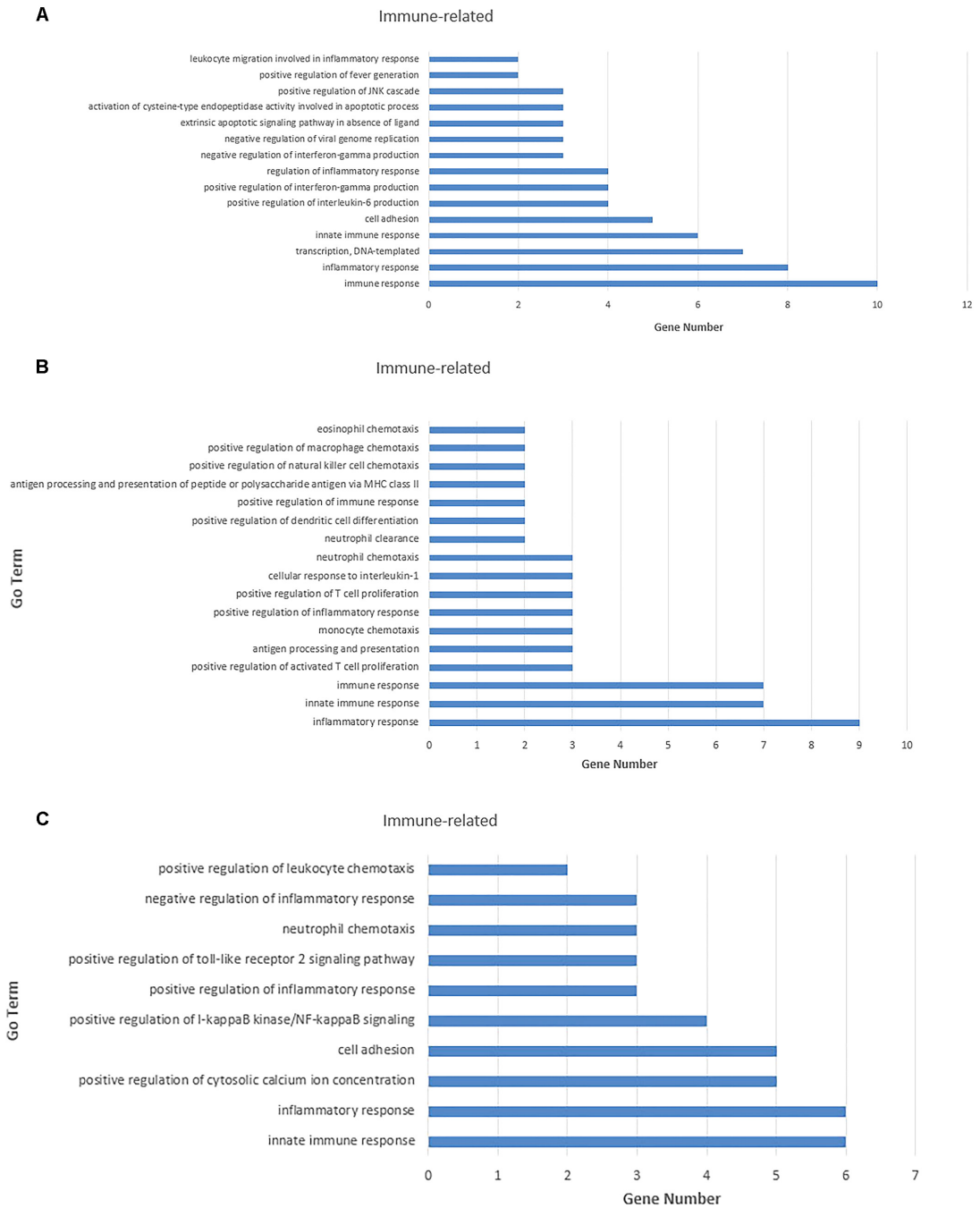


FIGURE 7 | Gene ontology (GO) enrichment analysis for negatively correlated miRNA–mRNA. **(A)** Enriched immune-related GO terms of 109 target genes of repressed DE miRNAs in the lung tissue. **(B)** Enriched immune-related GO terms of 135 target genes of repressed DE miRNAs in the trachea tissue. **(C)** Enriched immune-related GO terms of 141 target genes of repressed DE miRNAs under the infected state with different tissues. The abscissa represents the sample name, and the numbers indicate related genes.

We also found several genes played roles in multiple pathways. For example, TLR4 (miR-122) and nuclear factor kappa B subunit 1 (NFKB1) (miR-34c) were all involved in the influenza A pathway, as well as the TLR signaling pathway. While tumor necrosis factor (TNF) (miR-331) and nitric oxide synthase 3 (NOS3) (miR-335) participate in the PI3K-Akt signaling pathway and TNF signaling pathway (**Figure 8**).

qPCR Validation of DEGs

Seventeen DE mature miRNAs (cfa-miR-122, cfa-miR-129, cfa-miR-1838, cfa-miR-185, cfa-miR-23a, cfa-miR-331, cfa-miR-34c,

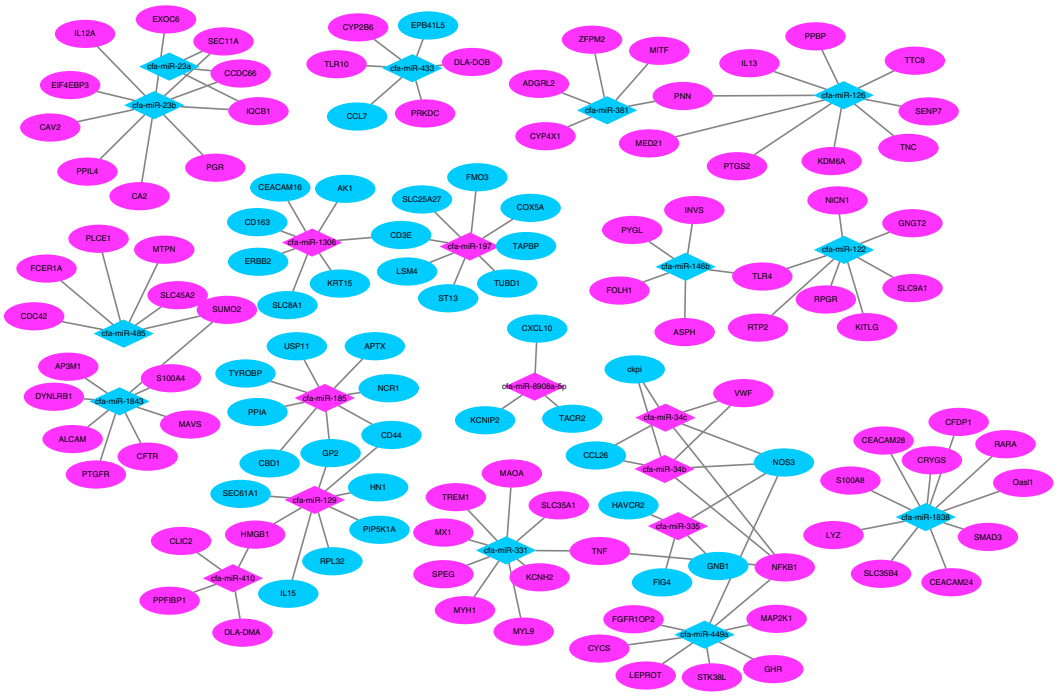


FIGURE 8 | MiRNA–mRNA negative correlation network. Red indicates up-regulation and green indicates down-regulation.

TABLE 1 | Relative miRNA expression of selected differentially expressed genes (DEGs) as revealed through miRNA-Seq and quantitative real-time PCR analyses.

miR_name	Illumina miRNA-seq (log2 fold change)	Regulation	Real-time PCR (log2 fold change)
cfa-miR-122	−1.21676769	Down	−5.63666666666667*
cfa-miR-129	3.53389948	Up	6.78333333333334*
cfa-miR-1838	−1.07414525	Down	−6.54333333333334*
cfa-miR-185	1.12327887	Up	6.42333333333334*
cfa-miR-23a	−1.2160282	Down	−6.08*
cfa-miR-331	1.06565908	Up	6.77666666666667*
cfa-miR-34c	−2.725446079	Down	−9.88333666666666*
cfa-miR-410	−5.370665561	Down	−3.22666666666667*
cfa-miR-433	−5.630137818	Down	−4.88333333333334*
cfa-miR-449a	−5.107389669	Down	−3.62666666666667*
cfa-miR-1843	−1.29413691	Down	−5.89666666666667*
cfa-miR-8908a-5p	2.78257022	Up	4.22666666666667*
cfa-miR-335	4.08439398	Up	2.80666666666668*
cfa-miR-34b	4.63988339	Up	1.66666666666667*
cfa-miR-449a	4.19807289	Up	3.915*
cfa-miR-485	4.9041897	Up	1.83333333333334*
cfa-miR-126	1.26672194	Up	3.98*
cfa-miR-370	4.9041897	Up	2.66666666666667*

Asterisk indicates statistical significance of differential gene expression with p-value < 0.05 (t-test).

cfa-miR-410, cfa-miR-433, cfa-miR-449a, cfa-miR-1843, cfa-miR-8908a-5p, cfa-miR-335, cfa-miR-34b, cfa-miR-485, cfa-miR-126, and cfa-miR-370) and 19 DEGs among the negative miRNA–mRNA interaction network were validated by qRT-PCR. The validation with qPCR was consistent with the sequencing results (as shown in **Tables 1**, **2**).

DISCUSSION

H5N1 infections are characterized with a high-fatality rate; hence, in this study, we choose beagles as the animal model of H5N1 infection. We investigated the effect of H5N1 virus infection on miRNA and mRNA, and analyzed the interaction

TABLE 2 | Relative mRNA expression of 10 selected DEGs as revealed through mRNA-Seq and Quantitative real-time PCR.

Annotation	Accession	Illumina mRNA-seq (log2 fold change)	Regulation	Real-time PCR (log2 fold change)
Toll-like receptor 4 (TLR4)	ENSCAFG00000003518	0.0576201	Up	0.480000000000004
Glycoprotein 2 (GP2)	ENSCAFG00000018023	−2.4609	Down	−2.53333333333333*
Lysozyme (LYZ)	ENSCAFG00000000426	2.36132	Up	2.09*
Interleukin 12A (IL12A)	ENSCAFG00000014200	−2.4609	Down	−0.973333333333333*
Tumor necrosis factor (TNF)	ENSCAFG00000000517	0.0758078	Up	0.663333333333327
G-protein subunit gamma transducin 2 (GNGT2)	ENSCAFG00000016918	2.38011	Up	0.583333333333329
Interleukin 13 (IL13)	ENSCAFG00000000878	1.14698	Up	0.515000000000004
Nuclear factor kappa B subunit 1 (NF-κB1)	ENSCAFG00000010730	0.0836651	Up	1.97666666666666*
Major histocompatibility complex, class II, DM alpha (DLA-DMA)	ENSCAFG00000000848	0.834172	Up	1.70500000000001*
Major histocompatibility complex, class II, DO beta (DLA-DOB)	ENSCAFG00000000819	1.58298	Up	2.85666666666667*
Nitric oxide synthase 3 (NOS3)	ENSCAFG00000004687	0.47985	Up	2.80666666666667*
C–C motif chemokine ligand 5 (CCL5)	ENSCAFG00000018171	−0.952333	Down	−0.925000000000001*
Interleukin 15 (IL15)	ENSCAFG00000003626	−1.19931	Down	−0.388333333333332
Cystic fibrosis transmembrane conductance regulator (CFTR)	ENSCAFG00000003429	0.887699	Up	0.418333333333333
Chemokine (C–X–C motif) ligand 10 (CXCL10)	ENSCAFG00000008584	−2.12935	Down	−0.9*
Nitric oxide synthase 3 (NOS3)	ENSCAFG00000004687	−1.4553	Down	−0.694999999999997
Phospholipase C epsilon 1 (PLCE1)	ENSCAFG00000007985	−0.983665	Down	−0.326666666666664
Tenascin C (TNC)	ENSCAFG00000003426	−1.1546	Down	−1.02166666666667*
Adrenoceptor alpha 1B (ADRA1B)	ENSCAFG00000017281	−1.5188	Down	−0.594999999999999

**Asterisk indicates statistical significance of differential gene expression with p-value < 0.05 (t-test).*

between miRNA and mRNA. Dysregulation of 24 miRNAs and 204 mRNAs with 138 negative miRNA–mRNA pairs was observed in the lung tissue, and dysregulation of 97 miRNAs and 135 mRNAs with 134 negative miRNA–mRNA pairs was observed in the tracheal tissue; moreover, under the infected state, between lung and trachea, miRNAs and mRNAs were also different expressed: 97 miRNAs and 141 mRNAs with 120 negative miRNA–mRNA pairs were found between different tissues. Target mRNA gene functions were analyzed with GO and the KEGG database. Several pathways were activated by relevant target genes of DE miRNAs with a negative miRNA–mRNA correlation.

Infection with H5N1 influenza virus caused differential expression of mRNA in lung and trachea tissues. Comparisons among these infected groups revealed 65 similar up-regulated genes and 7 similar down-regulated genes. These up-regulated genes included antiviral genes, such as cell surface receptor (*CD59*), calcium ion binding gene (*EFHC1* and *CAPS2*), and protein coding genes (*RIBC2*, *CFAP157*, *DRC7*, *CFAP45*, *CFAP65*, and *CCDC33*). *CD59* (Zhang et al., 2010) can inhibit local inflammatory reactions caused by the influenza virus. *EFHC1* affects cellular apoptosis (Wang et al., 2002; Katano et al., 2012). *CFAP157* (Benos et al., 2012), *DRC7* (Yang et al., 2011), and *CFAP45* (Li et al., 1999) are involved in the regulation of flagella and cilia motility for viral invasion. Of the down-regulated genes, *CD5L* combined with any modified low-density lipoprotein (LDL) or other polyanionic ligand and delivered the ligand into the cell via receptor-mediated endocytosis to regulate innate immunity (Peiser et al., 2002). SLC11A1 is a natural resistance-associated macrophage protein (NRAMP1), which is expressed only in immune-related cells (Govoni and Gros, 1998). As a divalent transition metal transporter, it was also involved in iron metabolism and assisted with host resistance to certain pathogens. In conclusion, when infected by H5N1 CIV, immune-related genes and antiviral genes are activated to resist viral infection.

Differential miRNA expression in the lung and trachea revealed that the regulatory mechanism of miRNA on host responses when infected with CIV and that between lung and trachea is different. At different post-infection stages, there were more miRNAs expressed at 3 than 7 dpi. Immunogenicity was enhanced by miR-155 (Izzard et al., 2017), miR-375 (Lin et al., 2017), miR-155, miR-146a (Zhou et al., 2017), and miR-136 (Zhao et al., 2015). Furthermore, miR-146a (Terrier et al., 2013) plays an antiviral role. Moreover, miR-29 activates cyclooxygenase and lambda-1 IFN to resist viral infection (Fang et al., 2011). As a cytolytic virus, influenza A virus induces apoptosis, which results in organ and cellular dysfunction. MiR-15b and miR-451 regulate a series of pro-inflammatory cytokine responses (Chan et al., 2011; Rosenberger et al., 2012). Moreover, miR-29c inhibits *BCL2L2* expression to regulate apoptosis induced by influenza A virus (Guan et al., 2012). These data reveal that viral infections typically induce miRNAs that regulate cytokine production and the anti-viral immune response.

The present study reveals novel findings regarding H5N1 influenza virus-infected dogs and lung and tracheal tissue profiling, since these have not been previously performed. MiRNAs and their target genes have multiple relationships (Perez et al., 2009; Fan and Wang, 2016; Nakamura et al., 2016). Dysregulated miRNAs may serve as a diagnostic and prognostic biomarker (Okkenhaug and Vanhaesebroeck, 2003; Hale et al., 2010; Haneklaus et al., 2013; Ingle et al., 2015).

Influenza, caused by influenza virus, is an acute febrile contagious respiratory infectious disease. The innate immune system recognizes invading viruses (Hale et al., 2010; Iwasaki and Pillai, 2014) through multiple mechanisms. The non-structural NS1 inhibits type I IFN production (Haye et al., 2009; Pekosz et al., 2012) by inactivating transcription factors (MacEwan, 2002; Wang et al., 2017) such as IRF3 (Wang et al., 2017), AP1, and NF-κB (miR-34) (Taganov et al., 2006). Pattern recognition receptors (PRRs) recognize viral RNA (Thompson et al., 2011; Okamoto et al., 2017) to activate a specialized immune response at mucosal surfaces to combat viral invasion. TLR4 (miR-146) has been reported to attenuate influenza virus infection with TLR4 antagonists, which may be a novel therapeutic approach to combat infection (Shirey et al., 2013). MiR-146a regulates TRAF6 when infected with H3N2 virus (Deng Y. et al., 2017). Moreover, TRAF6 and MEKK1 are critical genes that IPS-1 activates NF-κB and induces IFNs (Yoshida et al., 2008). In addition, miR-144 targets TRAF4-IRF7 through NF-κB to attenuate the host response to influenza virus (Lopez et al., 2017). Furthermore, miR-34 targets pro-apoptosis Bax through downregulation (Fan and Wang, 2016) during an influenza virus infection. Despite the paucity in the number of studies on dogs, it is important to refer to previous studies. Briefly, NF-κB and TLR4 with their target miR-34c and miR-146 through IPS-1 and TRAF family may have important roles in regulating the innate immune system during an H5N1 CIV infection.

Toll-like receptors activate IFNs and lead to antiviral responses through cytokine–cytokine receptor interactions. As chemokines (Zlotnik and Yoshie, 2000), ccl5 (miR-335) and cxcl10 (miR-8908a-5p) activate their receptors once activated, different downstream pathways are activated and may produce a series of immune responses. In addition, P13K-Akt signaling pathway regulates transcription, growth, proliferation, translation, and survival of fundamental cellular functions (Koyasu, 2003; Okkenhaug and Vanhaesebroeck, 2003; Richardson et al., 2004; Duronio, 2008; Hers et al., 2011). When cells are infected with the influenza A virus, viral NS1 binds to and activates P13K, inducing the beta-IFN and the apoptotic responses (Ehrhardt et al., 2007; Ehrhardt and Ludwig, 2009). In our findings, up-regulated GNGT2 and down-regulated miR-122 activate the P13K-Akt pathway during an influenza virus infection. Despite the paucity in the number of studies on canine microRNA and mRNA, it is important to refer to previous studies to explain the function of microRNAs in dogs and is of great significance in studies on dogs.

Although the expression levels and functions of miRNAs in the lung and trachea have been studied, the effects of beagles infected with influenza virus on mRNA and miRNA expression and influenza virus-related pathways have not been investigated. To our knowledge, this is the first study to assess the effect of beagles on miRNA–mRNA expression when infected with H5N1 influenza virus. Despite the low chance of H5N1 infection in domestic animals, direct contact of H5N1-infected dogs with humans may occur. Our results provide information for understanding the mechanisms of viral pathogenesis in dogs.

AUTHOR CONTRIBUTIONS

CF contributed to the data analysis and the writing of the manuscript. JL contributed to the drafting of the manuscript. SY and ZY contributed to the data collection and the laboratory work. CF and JL contributed to the animal experiment. SL contributed to the conception of the idea and design. All authors read and approved the manuscript.

FUNDING

This project was supported in part by The National Natural Science Foundation of China (31672563), The National Key Research and Development Program of China (2016YFD0501004), and The Guangdong Provincial Key Laboratory of Prevention and Control for Severe Clinical Animal Diseases (2017B030314142).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00303/full#supplementary-material

REFERENCES

Ambrosini-Spaltro, A., Farnedi, A., Montironi, R., and Foschini, M. P. (2011). IGFBP2 as an immunohistochemical marker for prostatic adenocarcinoma. *Appl. Immunohistochem. Mol. Morphol.* 19, 318–328. doi: 10.1097/PAI.0b013e3128052936

Ashour, M. M., Khatab, A. M., El-Folly, R. F., and Amer, W. A. (2012). Clinical features of avian influenza in Egyptian patients. *J. Egypt Soc. Parasitol.* 42, 385–396. doi: 10.12816/0006325

Baumann, S., Hess, J., Eichhorst, S. T., Krueger, A., Angel, P., Krammer, P. H., et al. (2003). An unexpected role for FosB in activation-induced cell death of T cells. *Oncogene* 22, 1333–1339. doi: 10.1038/sj.onc.1206126

Benos, P. V., Hoh, R. A., Stowe, T. R., Turk, E., and Stearns, T. (2012). Transcriptional program of ciliated epithelial cells reveals new cilium and centrosome components and links to human disease. *PLoS One* 7:e52166. doi: 10.1371/journal.pone.0052166

Cai, X., Lu, S., Zhang, Z., Gonzalez, C. M., Damania, B., and Cullen, B. R. (2005). Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5570–5575. doi: 10.1073/pnas.0408192102

Castleman, W. L., Powe, J. R., Crawford, P. C., Gibbs, E. P. J., Dubovi, E. J., Donis, R. O., et al. (2010). Canine H3N8 influenza virus infection in dogs and mice. *Vet. Pathol.* 47, 507–517. doi: 10.1177/0300985810363718

Catuogno, S., Cerchia, L., Romano, G., Pognonec, P., Condorelli, G., and de Franciscis, V. (2012). miR-34c may protect lung cancer cells from paclitaxel-induced apoptosis. *Oncogene* 32, 341–351. doi: 10.1038/onc.2012.51

Chan, L. Y., Kwok, H. H., Chan, R. W. Y., Peiris, M. J. S., Mak, N. K., Wong, R. N. S., et al. (2011). Dual functions of ginsenosides in protecting human endothelial cells against influenza H9N2-induced inflammation and apoptosis. *J. Ethnopharmacol.* 137, 1542–1546. doi: 10.1016/j.jep.2011.08.022

Chen, Y., Zhong, G., Wang, G., Deng, G., Li, Y., Shi, J., et al. (2010). Dogs are highly susceptible to H5N1 avian influenza virus. *Virology* 405, 15–19. doi: 10.1016/j.virol.2010.05.024

Cheng, A. M. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.* 33, 1290–1297. doi: 10.1093/nar/gki200

Choy, E. Y.-W., Siu, K.-L., Kok, K.-H., Lung, R. W.-M., Tsang, C. M., To, K.-F., et al. (2008). An Epstein-Barr virus–encoded microRNA targets PUMA to promote host cell survival. *J. Exp. Med.* 205, 2551–2560. doi: 10.1084/jem.20072581

Comer, B. S., Camoretti-Mercado, B., Kogut, P. C., Halayko, A. J., Solway, J., and Gerthoffer, W. T. (2014). MicroRNA-146a and microRNA-146b expression and anti-inflammatory function in human airway smooth muscle. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 307, L727–L734. doi: 10.1152/ajplung.00174.2014

Crawford, P. C., Dubovi, E. J., Castleman, W. L., Stephenson, I., Gibbs, E. P., Chen, L., et al. (2005). Transmission of Equine to dog. *Science* 310, 482–485. doi: 10.1126/science.1117950

Damiani, A. M., Kalthoff, D., Beer, M., Müller, E., and Osterrieder, N. (2012). Serological survey in dogs and cats for influenza A(H1N1)pdm09 in Germany. *Zoonoses Public Health* 59, 549–552. doi: 10.1111/j.1863-2378.2012.01541.x

de Cubas, A. A., Leandro-Garcia, L. J., Schiavi, F., Mancikova, V., Comino-Mendez, I., Inglada-Perez, L., et al. (2013). Integrative analysis of miRNA and mRNA expression profiles in pheochromocytoma and paraganglioma identifies genotype-specific markers and potentially regulated pathways. *Endocr. Relat. Cancer* 20, 477–493. doi: 10.1530/erc-12-0183

Deng, H., Chu, X., Song, Z., Deng, X., Xu, H., Ye, Y., et al. (2017). MicroRNA-1185 induces endothelial cell apoptosis by targeting UVRAG and KRIT1. *Cell Physiol. Biochem.* 41, 2171–2182. doi: 10.1159/000475571

Deng, Y., Yan, Y., Tan, K. S., Liu, J., Chow, V. T., Tao, Z.-Z., et al. (2017). MicroRNA-146a induction during influenza H3N2 virus infection targets and regulates TRAF6 levels in human nasal epithelial cells (hNECs). *Exp. Cell Res.* 352, 184–192. doi: 10.1016/j.yexcr.2017.01.011

Duronio, V. (2008). The life of a cell apoptosis regulation by the PI3/PKB pathway. *Biochem. J.* 415, 333–344. doi: 10.1042/BJ20081056

Ehrhardt, C., and Ludwig, S. (2009). A new player in a deadly game: influenza viruses and the PI3K/Akt signalling pathway. *Cell. Microbiol.* 11, 863–871. doi: 10.1111/j.1462-5822.2009.01309.x

Ehrhardt, C., Wolff, T., Pleschka, S., Planz, O., Beermann, W., Bode, J. G., et al. (2007). Influenza A virus NS1 protein activates the PI3K/Akt pathway to mediate antiapoptotic signaling responses. *J. Virol.* 81, 3058–3067. doi: 10.1128/jvi.02082-06

Fan, N., and Wang, J. (2016). MicroRNA 34a contributes to virus-mediated apoptosis through binding to its target gene Bax in influenza A virus infection. *Biomed. Pharmacother.* 83, 1464–1470. doi: 10.1016/j.biopha.2016.08.049

Fang, J., Hao, Q., Liu, L., Li, Y., Wu, J., Huo, X., et al. (2011). Epigenetic changes mediated by microRNA miR29 activate cyclooxygenase 2 and lambda-1 interferon production during viral infection. *J. Virol.* 86, 1010–1020. doi: 10.1128/jvi.06169-11

Gomez, I. G., MacKenna, D. A., Johnson, B. G., Kaimal, V., Roach, A. M., Ren, S., et al. (2014). Anti–microRNA-21 oligonucleotides prevent Alport nephropathy progression by stimulating metabolic pathways. *J. Clin. Invest.* 125, 141–156. doi: 10.1172/jci75852

Govoni, G., and Gros, P. (1998). Macrophage NRAMP1 and its role in resistance to microbial infections. *Inflamm. Res.* 47, 277–284. doi: 10.1007/s000110050330

Grey, F., Meyers, H., White, E. A., Spector, D. H., and Nelson, J. (2007). A human cytomegalovirus-encoded microRNA regulates expression of multiple viral genes involved in replication. *PLoS Pathog.* 3:e163. doi: 10.1371/journal.ppat.0030163

Guan, Z., Shi, N., Song, Y., Zhang, X., Zhang, M., and Duan, M. (2012). Induction of the cellular microRNA-29c by influenza virus contributes to virus-mediated apoptosis through repression of antiapoptotic factors BCL2L2. *Biochem. Biophys. Res. Commun.* 425, 662–667. doi: 10.1016/j.bbrc.2012.07.114

Hale, B. G., Albrecht, R. A., and García-Sastre, A. (2010). Innate immune evasion strategies of influenza virus. *Future Microbiol.* 5, 23–41. doi: 10.2217/FMB.09.108

Haneklaus, M., Gerlic, M., O'Neill, L. A. J., and Masters, S. L. (2013). miR-223: infection, inflammation and cancer. *J. Intern. Med.* 274, 215–226. doi: 10.1111/joim.12099

Hansen, A., Henderson, S., Lagos, D., Nikitenko, L., Coulter, E., Roberts, S., et al. (2010). KSHV-encoded miRNAs target MAF to induce endothelial cell reprogramming. *Genes Dev.* 24, 195–205. doi: 10.1101/gad.553410

Haye, K., Burmakina, S., Moran, T., Garcia-Sastre, A., and Fernandez-Sesma, A. (2009). The NS1 protein of a human influenza virus inhibits type I interferon production and the induction of antiviral responses in primary human dendritic and respiratory epithelial cells. *J. Virol.* 83, 6849–6862. doi: 10.1128/jvi.02323-08

Hers, I., Vincent, E. E., and Tavaré, J. M. (2011). Akt signalling in health and disease. *Cell. Signal.* 23, 1515–1527. doi: 10.1016/j.cellsig.2011.05.004

Ingle, H., Kumar, S., Raut, A. A., Mishra, A., Kulkarni, D. D., Kameyama, T., et al. (2015). The microRNA miR-485 targets host and influenza virus transcripts to regulate antiviral immunity and restrict viral replication. *Sci. Signal.* 8:ra126. doi: 10.1126/scisignal.aab3183

Iwasaki, A., and Pillai, P. S. (2014). Innate immunity to influenza virus infection. *Nat. Rev. Immunol.* 14, 315–328. doi: 10.1038/nri3665

Izzard, L., Dlugolenski, D., Xia, Y., McMahon, M., Middleton, D., Tripp, R. A., et al. (2017). Enhanced immunogenicity following miR-155 incorporation into the influenza A virus genome. *Virus Res.* 235, 115–120. doi: 10.1016/j.virusres.2017.04.002

Katano, M., Numata, T., Aguan, K., Hara, Y., Kiyonaka, S., Yamamoto, S., et al. (2012). The juvenile myoclonic epilepsy-related protein EFHC1 interacts with the redox-sensitive TRPM2 channel linked to cell death. *Cell Calcium* 51, 179–185. doi: 10.1016/j.ceca.2011.12.011

Klenk, H. D. (2014). Influenza viruses en route from birds to man. *Cell Host Microbe* 15, 653–654. doi: 10.1016/j.chom.2014.05.019

Koyasu, S. (2003). The role of PI3K in immune cells. *Nature* 4, 313–319. doi: 10.1038/ni0403-313

Lee, C., Song, D., Kang, B., Kang, D., Yoo, J., Jung, K., et al. (2009). A serological survey of avian origin canine H3N2 influenza virus in dogs in Korea. *Vet. Microbiol.* 137, 359–362. doi: 10.1016/j.vetmic.2009.01.019

Lee, I. H., Le, T. B., Kim, H. S., and Seo, S. H. (2016). Isolation of a novel H3N2 influenza virus containing a gene of H9N2 avian influenza in a dog in South Korea in 2015. *Virus Genes* 52, 142–145. doi: 10.1007/s11262-015-1272-z

Li, K., Wang, Y., Zhang, A., Liu, B., and Jia, L. (2017). miR-379 inhibits cell proliferation, invasion, and migration of vascular smooth muscle cells by targeting insulin-like factor-1. *Yonsei Med. J.* 58, 234–240. doi: 10.3349/ymj.2017.58.1.234

Li, S., Shi, Z., Jiao, P., Zhang, G., Zhong, Z., Tian, W., et al. (2010). Avian-origin H3N2 canine influenza A viruses in Southern China. *Infect. Genet. Evol.* 10, 1286–1288. doi: 10.1016/j.meegid.2010.08.010

Li, Z., Yao, K., and Cao, Y. (1999). Molecular cloning of a novel tissue-specific gene from human nasopharyngeal epithelium. *Gene* 237, 235–240. doi: 10.1016/S0378-1119(99)00234-6

Lin, J., Xia, J., Tu, C. Z., Zhang, K. Y., Zeng, Y., and Yang, Q. (2017). H9N2 avian influenza virus protein PB1 enhances the immune responses of bone marrow-derived dendritic cells by down-regulating miR375. *Front. Microbiol.* 8:287. doi: 10.3389/fmicb.2017.00287

Lopez, C. B., Rosenberger, C. M., Podyminogin, R. L., Diercks, A. H., Treuting, P. M., Peschon, J. J., et al. (2017). miR-144 attenuates the host response to influenza virus by targeting the TRAF6-IRF7 signaling axis. *PLoS Pathog.* 13:e1006305. doi: 10.1371/journal.ppat.1006305

Maas, R., Tacken, M., Ruuls, L., Koch, G., van Rooij, E., and Stockhofe-Zurwieden, N. (2007). Avian influenza (H5N1) susceptibility and receptors in dogs. *Emerg. Infect. Dis.* 13, 1219–1221. doi: 10.3201/eid1308.070393

MacEwan, D. J. (2002). TNF receptor subtype signalling: differences and cellular consequences. *Cell. Signal.* 14, 477–492. doi: 10.1016/S0898-6568(01)00262-5

Meydan, C., Shenhar-Tsarfaty, S., and Soreq, H. (2016). MicroRNA regulators of anxiety and metabolic disorders. *Trends Mol. Med.* 22, 798–812. doi: 10.1016/j.molmed.2016.07.001

Murphy, E., Vanicek, J., Robins, H., Shenk, T., and Levine, A. J. (2008). Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: implications for latency. *Proc. Natl. Acad. Sci. U.S.A.* 105, 5453–5458. doi: 10.1073/pnas.0711910105

Nachmani, D., Stern-Ginossar, N., Sarid, R., and Mandelboim, O. (2009). Diverse herpesvirus microRNAs target the stress-induced immune ligand MICB to escape recognition by natural killer cells. *Cell Host Microbe* 5, 376–385. doi: 10.1016/j.chom.2009.03.003

Nakamura, S., Horie, M., Daidoji, T., Honda, T., Yasugi, M., Kuno, A., et al. (2016). Influenza A virus-induced expression of a GalNAc transferase, GALNT3, via

microRNAs is required for enhanced viral replication. *J. Virol.* 90, 1788–1801. doi: 10.1128/jvi.02246-15

Okamoto, M., Tsukamoto, H., Kouwaki, T., Seya, T., and Oshiumi, H. (2017). Recognition of viral RNA by pattern recognition receptors in the induction of innate immunity and excessive inflammation during respiratory viral infections. *Viral Immunol.* 30, 408–420. doi: 10.1089/vim.2016.0178

Okkenhaug, K., and Vanhaesebroeck, B. (2003). PI3K in lymphocyte development, differentiation and activation. *Nat. Rev. Immunol.* 3, 317–330. doi: 10.1038/nri1056

Palumbo, T., Poultsides, G. A., Kouraklis, G., Liakakos, T., Drakaki, A., Peros, G., et al. (2016). A functional microRNA library screen reveals miR-410 as a novel anti-apoptotic regulator of cholangiocarcinoma. *BMC Cancer* 16:353. doi: 10.1186/s12885-016-2384-0

Peiser, L., Mukhopadhyay, S., and Gordon, S. (2002). Scavenger receptors in innate immunity. *Curr. Opin. Immunol* 14, 123–128. doi: 10.1016/S0952-7915(01)00307-7

Pekosz, A., Rajsbaum, R., Albrecht, R. A., Wang, M. K., Maharaj, N. P., Versteeg, G. A., et al. (2012). Species-specific inhibition of RIG-I ubiquitination and IFN induction by the influenza A virus NS1 protein. *PLoS Pathog.* 8:e1003059. doi: 10.1371/journal.ppat.1003059

Peng, X., Gralinski, L., Ferris, M. T., Frieman, M. B., Thomas, M. J., Proll, S., et al. (2011). Integrative deep sequencing of the mouse lung transcriptome reveals differential expression of diverse classes of small RNAs in response to respiratory virus infection. *mBio* 2:e198-11. doi: 10.1128/mBio.00198-11

Perez, J. T., Pham, A. M., Lorini, M. H., Chua, M. A., Steel, J., and tenOever, B. R. (2009). MicroRNA-mediated species-specific attenuation of influenza A virus. *Nat. Biotechnol.* 27, 572–576. doi: 10.1038/nbt.1542

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F. A., et al. (2005). Identification of microRNAs of the herpesvirus family. *Nat. Methods* 2, 269–276. doi: 10.1038/nmeth746

Pfeffer, S., Zavolan, M., Grässer, F. A., Chien, M., Russo, J. J., Ju, J., et al. (2004). Identification of virus-encoded microRNAs. *Science* 304, 734–736. doi: 10.1126/science.1096781

Ploegh, H. L., Robertson, K. A., Hsieh, W. Y., Forster, T., Blanc, M., Lu, H., et al. (2016). An interferon regulated microRNA provides broad cell-intrinsic antiviral immunity through multihit host-directed targeting of the sterol pathway. *PLoS Biol.* 14:e1002364. doi: 10.1371/journal.pbio.1002364

Richardson, C. J., Schalm, S. S., and Blenis, J. (2004). PI3-kinase and TOR: PIKTORing cell growth. *Semin. Cell Dev. Biol.* 15, 147–159. doi: 10.1016/j.semcdb.2003.12.023

Rosenberger, C. M., Podyminogin, R. L., Navarro, G., Zhao, G. W., Askovich, P. S., Weiss, M. J., et al. (2012). miR-451 regulates dendritic cell cytokine responses to influenza infection. *J. Immunol.* 189, 5965–5975. doi: 10.4049/jimmunol.1201437

Samols, M. A., Skalsky, R. L., Maldonado, A. M., Riva, A., Lopez, M. C., Baker, H. V., et al. (2007). Identification of cellular genes targeted by KSHV-encoded microRNAs. *PLoS Pathog.* 3:e65. doi: 10.1371/journal.ppat.0030065

Sant, A. J., Tundup, S., Kandasamy, M., Perez, J. T., Mena, N., Steel, J., et al. (2017). Endothelial cell tropism is a determinant of H5N1 pathogenesis in mammalian species. *PLoS Pathog.* 13:e1006270. doi: 10.1371/journal.ppat.1006270

Schliemann, C., Palumbo, A., Zuberbuhler, K., Villa, A., Kaspar, M., Trachsel, E., et al. (2008). Complete eradication of human B-cell lymphoma xenografts using rituximab in combination with the immunocytokine L19-IL2. *Blood* 113, 2275–2283. doi: 10.1182/blood-2008-05-160747

Schwarzbich, M.-A., Gutknecht, M., Salih, J., Salih, H. R., Brossart, P., Rittig, S. M., et al. (2011). The immune inhibitory receptor osteoactivin is upregulated in monocyte-derived dendritic cells by BCR–ABL tyrosine kinase inhibitors. *Cancer Immunol. Immunother.* 61, 193–202. doi: 10.1007/s00262-011-1096-1

Shirey, K. A., Lai, W., Scott, A. J., Lipsky, M., Mistry, P., Pletneva, L. M., et al. (2013). The TLR4 antagonist Eritoran protects mice from lethal influenza infection. *Nature* 497, 498–502. doi: 10.1038/nature12118

Shivdasani, R. A. (2006). MicroRNAs regulators of gene expression and cell differentiatio. *Blood* 108, 3646–3653. doi: 10.1182/blood-200601-030015

Song, G., Sharma, A. D., Roll, G. R., Ng, R., Lee, A. Y., Blelloch, R. H., et al. (2010). MicroRNAs control hepatocyte proliferation during liver regeneration. *Hepatology* 51, 1735–1743. doi: 10.1002/hep.23547

Songserm, T., Amonsin, A., Jam-on, R., Sae-Heng, N., Pariyothorn, N., Payungporn, S., et al. (2006a). Evidence of avian-like H9N2 influenza A virus among dogs in Guangxi, China. *Infect. Genet. Evol.* 12, 1744–1747. doi: 10.3201/eid1211.060542

Songserm, T., Amonsin, A., Jam-on, R., Sae-Heng, N., Pariyothorn, N., Payungporn, S., et al. (2006b). Fatal avian influenza A H5N1 IN a dog. *Emerg. Infect. Dis.* 12, 1744–1747. doi: 10.3201/eid1211.060542

Steichen, A. L., Binstock, B. J., Mishra, B. B., and Sharma, J. (2013). C-type lectin receptor Clec4d plays a protective role in resolution of Gram-negative pneumonia. *J. Leukoc. Biol.* 94, 393–398. doi: 10.1189/jlb.1212622

Stern-Ginossar, N., Elefant, N., Zimmermann, A., Wolf, D. G., Saleh, N., Biton, M., et al. (2007). Host immune system gene targeting by a viral miRNA. *Science* 317, 376–381. doi: 10.1126/science.1140956

Su, S., Chen, J., Jia, K., Khan, S. U., He, S., Fu, X., et al. (2014a). Evidence for subclinical influenza A(H1N1)pdm09 virus infection among dogs in Guangdong Province, China. *J. Clin. Microbiol.* 52, 1762–1765. doi: 10.1128/jcm.03522-13

Su, S., Chen, Y., Zhao, F.-R., Chen, J.-D., Xie, J.-X., Chen, Z.-M., et al. (2013). Avian-origin H3N2 canine influenza virus circulating in farmed dogs in Guangdong, China. *Infect. Genet. Evol.* 19, 251–256. doi: 10.1016/j.meegid.2013.05.022

Su, S., Qi, W., Zhou, P., Xiao, C., Yan, Z., Cui, J., et al. (2014b). First evidence of H10N8 avian influenza virus infections among feral dogs in live poultry markets in Guangdong province, China. *Clin. Infect. Dis.* 59, 748–750. doi: 10.1093/cid/ciu345

Su, S., Tian, J., Hong, M., Zhou, P., Lu, G., Zhu, H., et al. (2015). Global and quantitative proteomic analysis of dogs infected by avian-like H3N2 canine influenza virus. *Front. Microbiol.* 6:228. doi: 10.3389/fmicb.2015.00228

Suh, Y. S., Bhat, S., Hong, S.-H., Shin, M., Bahk, S., Cho, K. S., et al. (2015). Genome-wide microRNA screening reveals that the evolutionary conserved miR-9a regulates body growth by targeting sNPFR1/NPYR. *Nat. Commun.* 6:7693. doi: 10.1038/ncomms8693

Sun, X., Xu, X., Liu, Q., Liang, D., Li, C., He, Q., et al. (2013). Evidence of avian-like H9N2 influenza A virus among dogs in Guangxi, China. *Infect. Genet. Evol.* 20, 471–475. doi: 10.1016/j.meegid.2013.10.012

Taganov, K. D., Boldin, M. P., Chang, K. J., and Baltimore, D. (2006). NF-B-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12481–12486. doi: 10.1073/pnas.0605298103

Terrier, O., Textoris, J., Carron, C., Marcel, V., Bourdon, J. C., and Rosa-Calatrava, M. (2013). Host microRNA molecular signatures associated with human H1N1 and H3N2 influenza A viruses reveal an unanticipated antiviral activity for miR-146a. *J. Gen. Virol.* 94(Pt 5), 985–995. doi: 10.1099/vir.0.049528-0

Thompson, M. R., Kaminski, J. J., Kurt-Jones, E. A., and Fitzgerald, K. A. (2011). Pattern recognition receptors and the innate immune response to viral infection. *Viruses* 3, 920–940. doi: 10.3390/v3060920

Tu, J., Zhou, H., Jiang, T., Li, C., Zhang, A., Guo, X., et al. (2009). Isolation and molecular characterization of equine H3N8 influenza viruses from pigs in China. *Arch. Virol.* 154, 887–890. doi: 10.1007/s00705-009-0381-1

Wang, L., Fu, X., Zheng, Y., Zhou, P., Fang, B., Huang, S., et al. (2017). The NS1 protein of H5N6 feline influenza virus inhibits feline beta interferon response by preventing NF-κB and IRF3 activation. *Dev. Comp. Immunol.* 74, 60–68. doi: 10.1016/j.dci.2017.04.003

Wang, S., Chen, J.-Z., Zhang, Z., Huang, Q., Gu, S., Ying, K., et al. (2002). Cloning, characterization, and expression of calcyphosine 2, a novel human gene encoding an EF-Hand Ca^{2+}-binding protein. *Biochem. Biophys. Res. Commun.* 291, 414–420. doi: 10.1006/bbrc.2002.6461

Yang, Y., Cochran, D. A., Gargano, M. D., King, I., Samhat, N. K., Burger, B. P., et al. (2011). Regulation of flagellar motility by the conserved flagellar protein CG34110/Ccdc135/FAP50. *Mol. Biol. Cell* 22, 976–987. doi: 10.1091/mbc.E10-04-0331

Yin, X., Zhao, F.-R., Zhou, D.-H., Wei, P., and Chang, H.-Y. (2014). Serological report of pandemic and seasonal human influenza virus infection in dogs in southern China. *Arch. Virol.* 159, 2877–2882. doi: 10.1007/s00705-014-2119-y

Yoshida, R., Takaesu, G., Yoshida, H., Okamoto, F., Yoshioka, T., Choi, Y., et al. (2008). TRAF6 and MEKK1 play a pivotal role in the RIG-I-like helicase antiviral pathway. *J. Biol. Chem.* 283, 36211–36220. doi: 10.1074/jbc.M806576200

Zhang, C., Xu, Y., Jia, L., Yang, Y., Wang, Y., Sun, Y., et al. (2010). A new therapeutic strategy for lung tissue injury induced by influenza with CR2 targeting complement inhibitor. *Virol. J.* 7:30. doi: 10.1186/1743-422x-7-30

Zhang, H., Li, Y., Liu, Y., Liu, H., Wang, H., Jin, W., et al. (2016). Role of plant MicroRNA in cross-species regulatory networks of humans. *BMC Syst. Biol.* 10:60. doi: 10.1186/s12918-016-0292-1

Zhang, Z., Wan, F., Zhuang, Q., Zhang, Y., and Xu, Z. (2017). Suppression of miR-127 protects PC-12 cells from LPS-induced inflammatory injury by downregulation of PDCD4. *Biomed. Pharmacother.* 96, 1154–1162. doi: 10.1016/j.biopha.2017.11.107

Zhao, F.-R., Su, S., Zhou, D.-H., Zhou, P., Xu, T.-C., Zhang, L.-Q., et al. (2014). Comparative analysis of microRNAs from the lungs and trachea of dogs (*Canis familiaris*) infected with canine influenza virus. *Infect. Genet. Evol.* 21, 367–374. doi: 10.1016/j.meegid.2013.11.019

Zhao, H., Wang, L., Luo, H., Li, Q.-Z., and Zuo, X. (2017). TNFAIP3 downregulation mediated by histone modification contributes to T-cell dysfunction in systemic lupus erythematosus. *Rheumatology* 56, 835–843. doi: 10.1093/rheumatology/kew508

Zhao, L., Zhu, J., Zhou, H., Zhao, Z., Zou, Z., Liu, X., et al. (2015). Identification of cellular microRNA-136 as a dual regulator of RIG-I-mediated innate immunity that antagonizes H5N1 IAV replication in A549 cells. *Sci. Rep.* 5:14991. doi: 10.1038/srep14991

Zhou, D., Bitar, A., De, R., Melgar, S., Aung, K. M., Rahman, A., et al. (2017). Induction of immunomodulatory miR-146a and miR-155 in small intestinal epithelium of *Vibrio cholerae* infected patients at acute stage of cholera. *PLoS One* 12:e0173817. doi: 10.1371/journal.pone.0173817

Zlotnik, A., and Yoshie, O. (2000). Chemokines: a new classification system and their role in immunity. *Immunity* 12, 121–127. doi: 10.1016/S1074-7613(00)80165-X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

PERSPECTIVE
published: 25 July 2017
doi: 10.3389/fpls.2017.01167

tRNA Derived smallRNAs: smallRNAs Repertoire Has Yet to Be Decoded in Plants

*Gaurav Sablok[1,2†], Kun Yang[3†], Rui Chen[4] and Xiaopeng Wen[3]**

[1] Finnish Museum of Natural History, Helsinki, Finland, [2] Department of Biosciences, Viikki Plant Science Center, University of Helsinki, Helsinki, Finland, [3] Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region, Ministry of Education, Institute of Agro-Bioengineering and College of Life Sciences, Guizhou University, Guiyang, China, [4] Tianjin Institute of Agricultural Quality Standard and Testing Technology, Tianjin Academy of Agricultural Sciences, Tianjin, China

Among several smallRNAs classes, microRNAs play an important role in controlling the post-transcriptional events. Next generation sequencing has played a major role in extending the landscape of miRNAs and revealing their spatio-temporal roles in development and abiotic stress. Lateral evolution of these smallRNAs classes have widely been seen with the recently emerging knowledge on tRNA derived smallRNAs. In the present perspective, we discussed classification, identification and roles of tRNA derived smallRNAs across plants and their potential involvement in abiotic and biotic stresses.

Keywords: smallRNAs, tRNAs, microRNAs, functional genomics, stress

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos - Philosophische Praxis, Austria

Reviewed by:
German Martinez,
Swedish University of Agricultural Sciences, Sweden
Frantisek Baluska,
University of Bonn, Germany

****Correspondence:***
Xiaopeng Wen
xpwensc@hotmail.com

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Plant Science

Received: *28 April 2017*
Accepted: *19 June 2017*
Published: *25 July 2017*

Citation:
Sablok G, Yang K, Chen R and Wen X (2017) tRNA Derived smallRNAs: smallRNAs Repertoire Has Yet to Be Decoded in Plants. Front. Plant Sci. 8:1167. doi: 10.3389/fpls.2017.01167

smallRNAs: POST-TRANSCRIPTIONAL CHECK POINTS

Post-transcriptional regulation represents an integrated network of array of RNAs, among which regulatory RNAs play a major role in deciphering and regulating the functional omics. Rise of next generation sequencing approaches have widely elucidated several class of regulatory RNAs, often categorized into small non-coding RNAs and long non-coding RNAs (Heo et al., 2013). The functional role of these regulatory RNAs has been well described in plant genomics and has conserved or distinct functional roles in plants either through the RNA directed DNA methylation (RdDM pathway) or by epigenetically silencing the transposable elements (Chan et al., 2004). Swathing information on small non-coding RNAs have been revealed across a wide range of plant species focussing on most abundant class of smallRNAs – microRNAs (Zhang et al., 2006). Origin of microRNAs has been attributed to several key events such as the coordination of the DICER-like proteins (DCL-1) and methylated *HEN*1 for microRNA biogenesis and targeted transcriptional and translational repression through cleavage site interactions. Coordinated regulatory activities and interaction networks of these smallRNAs classes have been previously shown to play a key role in understanding the plant functional and developmental genomics (Rubio-Somoza and Weigel, 2011; Meng et al., 2011; Li and Zhang, 2016). With the rapid development in the next generation sequencing technologies and miRNAs moving to the single cell miRNAs transcriptome (Faridani et al., 2016), a new class of smallRNAs, tRNA derived smallRNAs (Hsieh et al., 2009), which have been explored widely in animals have now recently started gaining substantial importance in plant genomics with recent evidences showing their canonical interactions with AGO1 and TE like microRNAs (Alves et al., 2017; Martinez et al., 2017).

Although the biogenesis of these tRNA derived smallRNAs, cleavage efficiency and target accessibility has been addressed with limited reports in plants, signatures of association of tRNA derived smallRNAs with AGO proteins in particular AGO1, which also loads miRNAs revealed their role in the transcript repression, were also observed for tRNA derived smallRNAs (Loss-Morais et al., 2013; Martinez et al., 2017). Alongside, this class of smallRNAs has been shown to be involved in translation repression in *Cucurbita maxima* where phloem specific tRNAs fragments have been shown to interfere with ribosomal activity and represses translation (Zhang et al., 2009). Interestingly, recent comparative analysis revealed hints toward the biogenesis of these tRNA derived smallRNAs independent of DICER-like proteins (Alves et al., 2017), however, reduced abundance of tRNA derived smallRNAs has been observed in *dcl1* double mutants (Martinez et al., 2017). Using both *dcl1* and *ago1* single mutants, specific reduction of the tRNAs derived smallRNAs was observed as compared to miRNAs, which concludes the involvement of AGO1 in tRNA derived smallRNAs biogenesis pathway (Martinez et al., 2017). Taking into account this new emerging class of smallRNAs with relatively less explored associations with AGO proteins and potential roles in transcriptional and translational repression, it is an intriguing question to address in plant genomics the role and functional diversity of tRNAs derived smallRNAs and their roles in context to most widely profiled class of miRNAs.

tRNA DERIVED smallRNAs: SMALL NON-CODING FUNCTIONAL BEAST

tRNA derived smallRNAs represent a class of 19-mer smallRNAs, which have been previously widely demonstrated in humans and play an intriguing role in regulating the gene expression post-transcriptionally (Sobala and Hutvagner, 2011). In humans, their interactions with the other established class of small non-coding RNAs such miRNAs and siRNAs have been widely demonstrated (Garcia-Silva et al., 2012). In plants, however, the detection and possible association of tRNA derived smallRNAs (Nowacka et al., 2013) is still lacking with only few reports indicating the roles of tRNAs derived smallRNAs in tissues, embryogenic callus (Chen et al., 2011), phloem (Zhang et al., 2009) and recently in pollens (Martinez et al., 2017). tRNA derived smallRNAs classification and nomenclature in plants have been correlated with 5′, 3′ and CCA proximities and have been length classified as 19–25 nt (Loss-Morais et al., 2013; Alves et al., 2017). Taking into account the length variations in correlation with abundance, so far observed length variations were found to be between 19 and 25 nt, while majority of them belonging to 5′ 19-nt and has been seen conserved across dicot (*Arabidopsis thaliana*-Gly^{UCC}), *Zea mays* (C4 species) and evolutionary conserved moss (*Physcomitrella patens*) (Alves et al., 2017; Martinez et al., 2017) except monocot (*Oryza sativa*-Ala^{AGC}), where 25-nt abundance was seen as the most dominant length of these smallRNAs (Chen et al., 2011). Previously observed correlation of length and corresponding abundances of tRNA derived smallRNAs are also supported by recent observations of Martinez et al. (2017), which highlighted the abundance of the 19-nt tRNAs derived smallRNAs as a part of the 5′ tRNA processing in pollens of *Arabidopsis thaliana*. However, across polyploids such as *Triticum aestivum*, abundance of 21-nt with most of them originating from Val^{CAC} has been observed (Wang Y. et al., 2016). Despite being conserved in patterns of distribution across monocots and dicots, *Oryza sativa* AGO reveals high abundance of specific Arg^{CCT} and Arg^{TCG} tRNA derived 19-nt smallRNAs as compared to the Ala^{AGC} (Alves et al., 2017). Difference in these patterns of length abundances might be due to species ploidy or due to the cleavage asymmetry and role of anticodon loop required for tRNA processing (Wang Y. et al., 2016). Chen et al. (2011) established the first report on the identification of tRNAs derived smallRNAs from meristem associated smallRNAs sequencing and differential regulation of 5′Ala^{AGC} and Pro^{CGG} in callus and leaves, which is also supported by the recent reports in *Arabidopsis thaliana* addressing the tissue specificity of tRNAs derived smallRNAs (Alves et al., 2017). Recently pollen specific accumulation of tRNAs derived smallRNAs (Ala^{AGC}) has been found and interestingly using the *dcl1* mutants they showed the interactions of these smallRNAs with transposable elements (TE) in *Arabidopsis thaliana* (Martinez et al., 2017).

Although the biogenesis of tRNA-derived smallRNAs has not been functionally elucidated, recent reports indicate toward an independent processing machinery, which doesn't have functional interplay of DICER-like proteins (Alves et al., 2017). This is in contrast to the recent report of Martinez et al. (2017) using the *dcl1* double mutants revealing the association of DCL-1 proteins and tRNAs derived smallRNAs. Although the functional processing machinery might be different but their association with the AGO1, 2, 4 and 7 (*ARGONAUTE*) proteins have been experimentally verified using the immunoprecipitated AGO proteins (AGO-IP) (Loss-Morais et al., 2013; Alves et al., 2017; Martinez et al., 2017). Moreover, length variations also affects the association of tRNA derived smallRNAs to AGO proteins. Alves et al. (2017) demonstrated the association of the 19- and 20-nt tRNA derived smallRNAs with AGO2 and AGO5 whereas AGO4 was found to be abundant with 19-, 24-, and 25-nt respectively.

It is worth to mention that not only tRNA derived smallRNAs biogenesis pathways showed independency with respect to microRNAs, the presence of terminal 5′ nucleotide is also interestingly different, whilst tRNA derived smallRNAs preferred to have a G as 5′ terminal as compared to microRNAs, which prefer to have either a U or A as the 5′ terminal nucleotide (Loss-Morais et al., 2013). These observations are similar to the recent observation in fungal pathogen *Phytophthora sojae* (Wang Q. et al., 2016) suggesting the conservation pattern of these preference for 5′ terminal nucleotides. Conservation of cleavage site analysis provides support that the origin of these tRNA derived smallRNAs is not a random exonucleolytic digestion of the tRNAs precursors (Alves et al., 2017). Interestingly, target site analysis revealed the unusual target cleavage at multiple sites in tRNA-derived smallRNAs as compared to the single center binding sites in the miRNAs and siRNAs (Wang Y. et al., 2016). However, using the PARE-seq data, pollen specific tRNAs derived smallRNAs describe the mode of cleavage site action canonical to

miRNAs (Martinez et al., 2017). Recent investigations using the nucleotide composition analysis across the cleavage sites revealed preferable enrichment of U across the cleavage site for both 5′ and 3′ tRNAs derived smallRNAs (Wang Y. et al., 2016).

Organelle genomes play an important role in response to abiotic stress and control the photosynthetic regulatory activities under abiotic stress. Although being small in size as compared to the nuclear genome, recent reports reveal a large percentage (25%) of the tRNAs derived smallRNAs derived from organelle genomes in particular chloroplast (Cognat et al., 2017). The observed abundance of the plastid encoded tRNAs supports the previous dynamic regulation of chloroplast encoded Tyr^{GTA} during the ASGV infection in *Malus* x *domestica* (Visser et al., 2014). Interestingly, the profiled tRNAs derived smallRNAs population represented both forms tRF-5D (due to a cleavage in the D region) and tRF-3T (via a cleavage in the T region) (Cognat et al., 2017). However, specific enrichment of the tRF-5D was seen with AGO1 immunoprecipitation libraries (Cognat et al., 2017). Although high abundance of the plastid encoded tRNAs smallRNAs was observed, however, they were found to be localized outside organelle. Taking into account these reports, it is yet to be addressed the transport mechanism and the functional role of the plastid encoded tRNAs derived smallRNAs.

Although been recently discovered, genome wide implications on the role of these tRNA derived smallRNAs has been shown in abiotic and biotic stress (Thompson et al., 2008; Asha and Soniya, 2016; Wang Q. et al., 2016). Initial reports indicating the possible involvement of these tRNA derived smallRNAs such as 5′ tRF of Asp^{GTC}, and 3′ CCA tRFs of Gly^{TCC}, which were found to be over-expressing in phosphate starvation and drought (Hsieh et al., 2009; Loss-Morais et al., 2013). A compiled list of the tRNA derived smallRNAs and their potential involvement in stress has been presented as **Table 1**. Interestingly, Alves et al. (2017) highlighted the role of RNS1 (*RIBONUCLEASE 1*) using a T-DNA insertion line (*rsn1*-1, D2lk1087165C) with possible involvements in the tRNA derived smallRNAs biogenesis. RNS1 belongs to the ribonuclease T_2 family, whose another member s-RNASE has been shown to be the key member involved in self-incompatibility (Bariola et al., 1999) and has been shown to be mainly regulated under phosphate (Pi) starvation and anthocyanin regulation (Bariola et al., 1999). Potential role of tRNA derived smallRNAs has also been found ruing the heat stress in polyploids such as *Triticum aestivum*, where the observed tRNAs derived smallRNAs population was found to be predominantly coming from the mature arms of tRNAs as compared to the nascent transcripts in polyploids (Wang Y. et al., 2016). Interestingly, diversification of the functional tRNA derived from the same amino acid was seen in *Triticum aestivum*, with Met^{CAU} displayed lower expression as compared to the other processed isotype, from the same isoacceptor, which showed up-regulation during the heat stress (Wang Y. et al., 2016). This along with the increased cleavage of 3′ ends during the heat stress confers the dynamic changes in the tRNA derived smallRNAs pool during the abiotic stress.

Systematic profiling of smallRNAs during Apple stem grooving virus (ASGV) infection in *Malus* x *domestica* revealed a large proportion of tRNA derived smallRNAs. Interestingly, their association with miRNAs, tasiRNAs, phasiRNAs were not observed (Visser et al., 2014). Visser et al. (2014) reported 33-nt tRNA derived smallRNAs as the most abundant ones with the most abundant being 5′ tRNA-half originating from Asp^{GTC}. As previously revealed during the abiotic stress, differential regulation of tRNAs derived smallRNAs and also smallRNAs overlapped by tRNAs were found to be differentially regulated in ASGV infected samples, one of the abundant tRNAs (Tyr^{GTA}) was found altering the sRNAs arrangement in ASGV-infected samples. Strikingly, the infections states showed inverse correlations of fragment types (those originating from the 3′ and extending into the variable regions and those originating from the central stem region of the tRNAs). This arguable contrasting pattern might hint toward the co-existence of the separate biogenesis pathways.

In plants, pathogen associated immunity is controlled through the microbial or pathogen associated molecular patterns (MAPS or PAMPs), which upon the pathogen or the microbial infection trigger the defense response through up-regulation of defense related genes as a part of plant immunity. Asha and Soniya (2016) demonstrated the involvement of the tRNAs derived smallRNAs in regulating the expression patterns of defense related genes during *Phytophthora capsici* infection in Black Pepper (*Piper nigrum* L.). Interestingly the dominance of 5′ tRNA derived smallRNAs was found among the observed tRNAs population, which supports that in plants the major class of these smallRNAs is represented by 5′ tRNAs (Alves et al., 2017; Martinez et al., 2017). Experimental confirmation of 5′ Ala^{CGC} target sites on Non-expresser of pathogenesis related protein (NPR1) confirmed the repression of the NPR1 during the pathogen infection (Asha and Soniya, 2016). Above confirmatory results establish that tRNA derived smallRNAs indeed supresses the expression pattern of target genes as previously observed in Zhang et al. (2009), where the phloem specific tRNAs were found to interact with the ribosomal activity thus leading to translational repression.

TABLE 1 | tRNA derived sequences and their revealed roles in abiotic and biotic stress.

tRNA derived sequence	Stress/Development	Reference
Ala^{AGC}	Drought/Salt	Loss-Morais et al., 2013
Arg^{CCT}	Drought	Loss-Morais et al., 2013; Alves et al., 2017
Arg^{TCG}	Drought	Loss-Morais et al., 2013
Gly^{TCC}	Drought	Loss-Morais et al., 2013
Asp^{GTC}	Phosphate starvation	Hsieh et al., 2009
Gly^{TCC}	Phosphate starvation	Hsieh et al., 2009
Arg^{CCT}	Cold	Alves et al., 2017
Arg^{TCG}	Oxidative stress	Alves et al., 2017
Tyr^{GTA}	Oxidative stress	Alves et al., 2017
Ile^{AAT}	Pathogen	Wang Q. et al., 2016
Arg^{ACG}	Pathogen	Wang Q. et al., 2016
Ala^{CGC}	Pathogen	Asha and Soniya, 2016
Val^{CAC}, Thr^{UGU}, Tyr^{GUA}, Ser^{UGA}	Heat stress	Wang Y. et al., 2016

A key message from Loss-Morais et al. (2013) indicated the association of these tRNAs derived smallRNAs to AGO2, which is an AGO of unknown function and has been shown to have relatively higher levels during biotic infections (Zhang et al., 2011). However, recent reports of Wang Q. et al. (2016), showed contrasting association of the tRNA derived smallRNAs to AGO1 in pathogenic infection. Taking into account these observations, it can be presumed that association of the tRNA-derived smallRNAs might be species or organism specific in response to abiotic or biotic stress. However, taking together the AGO1 associations and recent reports of AGO1 and DCL1 association using the *dcl1* double mutants, it is noteworthy to highlight that these classes of smallRNAs has regulatory roles, which are yet to be discovered in plants, which can add to the understanding of the plant functional genomics. It is worthwhile to conclude that plants harbor the most abundant pool of tRNAs with clearly distinct nuclear and organelle tRNAs and therefore these classes of tRNAs, their abundance, profiling and genetic interactions with the cognate targets would expand understanding of the complex RNA'ome in plants.

AUTHOR CONTRIBUTIONS

GS conceived and drafted the manuscript, KY, RC, and XW provided revisions to the MS.

FUNDING

Financial grant (NSFC:31560549) to XW for providing the open access fees of the article.

REFERENCES

Alves, C. S., Vicentini, R., Duarte, G. T., Pinoti, V. F., Vincentz, M., and Nogueira, F. T. (2017). Genome-wide identification and characterization of tRNA-derived RNA fragments in land plants. *Plant Mol. Biol.* 93, 35–48. doi: 10.1007/s11103-016-0545-9

Asha, S., and Soniya, E. V. (2016). Transfer RNA derived small RNAs targeting defense responsive genes are induced during *Phytophthora capsici* infection in black pepper (*Piper nigrum* L.). *Front. Plant Sci.* 7:767. doi: 10.3389/fpls.2016.00767

Bariola, P. A., MacIntosh, G. C., and Green, P. J. (1999). Regulation of S-like ribonuclease levels in Arabidopsis. Antisense inhibition of RNS1 or RNS2 elevates anthocyanin accumulation. *Plant Physiol.* 119, 331–342. doi: 10.1104/pp.119.1.331

Chan, S. W., Zilberman, D., Xie, Z., Johansen, L. K., Carrington, J. C., and Jacobsen, S. E. (2004). RNA silencing genes control de novo DNA methylation. *Science* 303, 1336. doi: 10.1126/science.1095989

Chen, C. J., Liu, Q., Zhang, Y. C., Qu, L. H,, Chen, Y. Q., and Gautheret, D. (2011). Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embryogenic callus. *RNA Biol.* 8:538–547. doi: 10.4161/rna.8.3.15199

Cognat, V., Morelle, G., Megel, C., Lalande, S., Molinier, J., Vincent, T., et al. (2017). The nuclear and organellar tRNA-derived RNA fragment population in *Arabidopsis thaliana* is highly dynamic. *Nucleic Acids Res.* 45, 3460–3472. doi: 10.1093/nar/gkw1122

Faridani, O. R., Abdullayev, I., Hagemann-Jensen, M., Schell, J. P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* 34, 1264–1266. doi: 10.1038/nbt.3701

Garcia-Silva, M. R., Cabrera-Cabrera, F., Güida, M. C., and Cayota, A. (2012). Hints of tRNA-derived small RNAs role in RNA silencing mechanisms. *Genes* 3, 603–614. doi: 10.3390/genes3040603

Heo, J. B., Lee, Y. S., and Sung, S. (2013). Epigenetic regulation by long noncoding RNAs in plants. *Chromosome Res.* 21, 685–693. doi: 10.1007/s10577-013-9392-6

Hsieh, L. C., Lin, S. I., Shih, A. C. C., Chen, J. W., Lin, W. Y., Tseng, C. Y., et al. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. *Plant Physiol.* 151, 2120–2132. doi: 10.1104/pp.109.147280

Li, C., and Zhang, B. (2016). MicroRNAs in control of plant development. *J. Cell. Physiol.* 231, 303–313. doi: 10.1002/jcp.25125

Loss-Morais, G., Waterhouse, P. M., and Margis, R. (2013). Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. *Biol. Direct.* 8:6. doi: 10.1186/1745-6150-8-6

Martinez, G., Choudury, S. G., and Slotkin, R. K. (2017). tRNA-derived small RNAs target transposable element transcripts. *Nucleic Acids Res.* 45, 5142–5152. doi: 10.1093/nar/gkx103

Meng, Y., Shao, C., and Chen, M. (2011). Toward microRNA-mediated gene regulatory networks in plants. *Brief. Bioinform.* 12, 645–659. doi: 10.1093/bib/bbq091

Nowacka, M., Strozycki, P. M., Jackowiak, P., Hojka-Osinska, A., Szymanski, M., and Figlerowicz, M. (2013). Identification of stable, high copy number, medium-sized RNA degradation intermediates that accumulate in plants under non-stress conditions. *Plant Mol. Biol.* 83, 191–204 doi: 10.1007/s11103-013-0079-3

Rubio-Somoza, I., and Weigel, D. (2011). MicroRNA networks and developmental plasticity in plants. *Trends Plant Sci.* 16, 258–264. doi: 10.1016/j.tplants.2011.03.001

Sobala, A., and Hutvagner, G. (2011). Transfer RNA-derived fragments: origins, processing, and functions. *Wiley Interdiscip. Rev. RNA* 2, 853–862. doi: 10.1002/wrna.96

Thompson, D. M., Lu, C., Green, P. J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA* 14, 2095–2103. doi: 10.1261/rna.1232808

Visser, M., Maree, H. J., Rees, D. J., and Burger, J. T. (2014). High-throughput sequencing reveals small RNAs involved in ASGV infection. *BMC Genomics* 15:568. doi: 10.1186/1471-2164-15-568

Wang, Q., Li, T., Xu, K., Zhang, W., Wang, X., Quan, J., et al. (2016). The tRNA-derived small RNAs regulate gene expression through triggering sequence-specific degradation of target transcripts in the oomycete pathogen *Phytophthora sojae*. *Front. Plant Sci.* 7:1938. doi: 10.3389/fpls.2016.01938

Wang, Y., Li, H., Sun, Q., and Yao, Y. (2016). Characterization of small RNAs derived from tRNAs, rRNAs and snoRNAs and their response to heat stress in wheat seedlings. *PLoS ONE* 11:e0150933. doi: 10.1371/journal.pone.0150933

Zhang, B., Pan, X., Cobb, G. P., and Anderson, T. A. (2006). Plant microRNA: a small regulatory molecule with big impact. *Dev. Biol.* 289, 3–16. doi: 10.1016/j.ydbio.2005.10.036

Zhang, S., Sun, L., and Kragler, F. (2009). The phloem-delivered RNA pool contains small noncoding RNAs and interferes with translation. *Plant Physiol.* 150, 378–387. doi: 10.1104/pp.108.134767

Zhang, X., Zhao, H., Gao, H., Wang, H., Katiyar-Agarwal, S., Huang, H., et al. (2011). *Arabidopsis* argonaute 2 regulates innate immunity via miRNA393*-Mediated silencing of a golgi-localized SNARE gene, MEMB12. *Mol. Cell* 42, 356–366. doi: 10.1016/j.molcel.2011.04.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 20 September 2017
doi: 10.3389/fmicb.2017.01801

MicroRNA-Mediated Gene Silencing in Plant Defense and Viral Counter-Defense

Sheng-Rui Liu[1†], Jing-Jing Zhou[2†], Chun-Gen Hu[3], Chao-Ling Wei[1] and Jin-Zhi Zhang[3*]*

[1] State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China, [2] College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, China, [3] Key Laboratory of Horticultural Plant Biology (Ministry of Education), College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, China

MicroRNAs (miRNAs) are non-coding RNAs of approximately 20–24 nucleotides in length that serve as central regulators of eukaryotic gene expression by targeting mRNAs for cleavage or translational repression. In plants, miRNAs are associated with numerous regulatory pathways in growth and development processes, and defensive responses in plant–pathogen interactions. Recently, significant progress has been made in understanding miRNA-mediated gene silencing and how viruses counter this defense mechanism. Here, we summarize the current knowledge and recent advances in understanding the roles of miRNAs involved in the plant defense against viruses and viral counter-defense. We also document the application of miRNAs in plant antiviral defense. This review discusses the current understanding of the mechanisms of miRNA-mediated gene silencing and provides insights on the never-ending arms race between plants and viruses.

Keywords: defense, counter-defense, gene silencing, miRNA, virus

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos – Philosophische Praxis, Austria

Reviewed by:
John Hammond,
Agricultural Research Service (USDA), United States
Eugene I. Savenkov,
Swedish University of Agricultural Sciences, Sweden

***Correspondence:**
Chao-Ling Wei
weichl@ahau.edu.cn
Jin-Zhi Zhang
jinzhizhang@mail.hzau.edu.cn

[†] These authors have contributed equally to this work.

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *28 May 2017*
Accepted: *05 September 2017*
Published: *20 September 2017*

Citation:
Liu S-R, Zhou J-J, Hu C-G, Wei C-L and Zhang J-Z (2017) MicroRNA-Mediated Gene Silencing in Plant Defense and Viral Counter-Defense. Front. Microbiol. 8:1801. doi: 10.3389/fmicb.2017.01801

INTRODUCTION

Viruses are among the most important causal agents of infectious diseases in both animals and plants. Disease symptoms associated with viral infection in plants include stunting, yellowing, mosaic patterns, ringspot, leaf rolling, wilting, necrosis, and other developmental abnormalities (Wang et al., 2012). During the course of evolution, plants have employed versatile mechanisms against invading viruses, such as RNA silencing, hormone-mediated defense, immune receptor signaling, protein degradation and regulation of metabolism (Calil and Fontes, 2016). Evidence is accumulating that RNA silencing plays critical roles in plant immunity against viruses. RNA silencing, which is induced by small RNAs (sRNAs), is a central regulator of gene expression and an evolutionarily conserved mechanism in eukaryotic organisms (Eamens et al., 2008; Pumplin and Voinnet, 2013). Plant sRNAs are grouped into two major classes: microRNAs (miRNAs) and small interfering RNAs (siRNAs). Plants have evolved three basic RNA silencing pathways, which are represented by the miRNA pathway, the siRNA-directed RNA degradation pathway, and the siRNA-directed DNA methylation (RdDM) pathway (Baulcombe, 2004; Eamens et al., 2008; Wang and Smith, 2016).

MicroRNAs are endogenous RNAs of 20–24 nucleotides that are processed by Dicer-like (DCL) proteins from imperfectly paired hairpin precursor RNAs, and typically targeting a single site in their target mRNA (Voinnet, 2009; Axtell, 2013). siRNAs are similar sized and also require DCL proteins for biogenesis, but they are derived from perfectly paired double-stranded trigger

RNA molecules that can be endogenous or derived from introduced RNAs, transgenes, or viruses, affecting multiple sites on the target RNA (Bartel, 2004, 2005). The siRNA-mediated gene silencing serves as a general defense mechanism against plant viruses (Wang et al., 2012; Pumplin and Voinnet, 2013; Revers and Nicaise, 2014; Ghoshal and Sanfacon, 2015; Khalid et al., 2017), while miRNAs are involved in plant growth and development, signal transduction, protein degradation, and response to biotic and abiotic stresses (Voinnet, 2009; Zhang et al., 2012; Bologna and Voinnet, 2014). However, miRNAs also play critical roles in plant–virus interactions (Li et al., 2012; Ramesh et al., 2014; Tiwari et al., 2014; Ghoshal and Sanfacon, 2015; Huang et al., 2016). Nowadays, miRNA-mediated gene silencing has been applied to protect several agricultural crop species against infection by diverse viruses (Tiwari et al., 2014; Khalid et al., 2017). In this review, we (1) document the biogenesis and origin of miRNAs and the current understanding of miRNA-mediated gene silencing mechanism in plants; (2) describe the roles of miRNAs in plant–virus interactions; and (3) discuss the current applications of miRNA-mediated gene silencing and advances in the technique in plant science.

ORIGINS, BIOGENESIS AND MODES OF ACTION OF PLANT miRNAs

miRNAs are derived from single-stranded RNA transcripts (*MIR* genes) that can fold back onto themselves to produce imperfectly double-stranded stem–loop precursor structures. The mechanisms of miRNA biogenesis and modes of action are well-established in plants (**Figure 1**). The *MIR* genes are RNA polymerase II (Pol II) transcription units that produce the primary miRNA transcript (pri-miRNA), which is then cleaved by DCL1 in the nucleus, leading to production of the shorter precursor-miRNA (pre-miRNA, partially duplex molecule with a single-stranded loop, mismatches, and a single-stranded extension) with the assistance of the dsRNA-binding protein 1 (DRB1) and HYPONASTIC LEAVES1 (HYL1). Subsequently, the miRNA duplex (miRNA/miRNA* where miRNA* stands for the passenger strand) is released from the pre-miRNA stem–loop structure by the second cleavage step with the help of the combined action of DCL1 and HYL1. The mature miRNA duplex is methylated by the sRNA-specific methyltransferase HUA ENHANCER1 (HEN1) and then exported to the cytoplasm through the action of the plant Exportin-5 ortholog HASTY and other unknown factors. In the cytoplasm, the mature miRNA strand is loaded onto Argonaute 1 (AGO1) to form an RNA-induced silencing complex (RISC) with the help of Hsc70/Hsp90 chaperone and ATP, followed by the passenger strand ejection (Iwasaki et al., 2010; Nakanishi, 2016). The RISC then uses the miRNA to guide the slicer activity of AGO1 to repress the expression of complementary target mRNAs (Llave et al., 2002). Two main modes of action have been described for target repression caused by miRNAs: translational repression and cleavage of target mRNA. It is worth noting that animal miRNAs bind 3′ untranslated regions (UTRs) and function predominantly through translational repression; whereas plant miRNAs primarily target the coding regions of mRNA, and repression of gene expression is mostly by transcript cleavage. Nevertheless, recent studies have indicated that miRNA-mediated translational repression

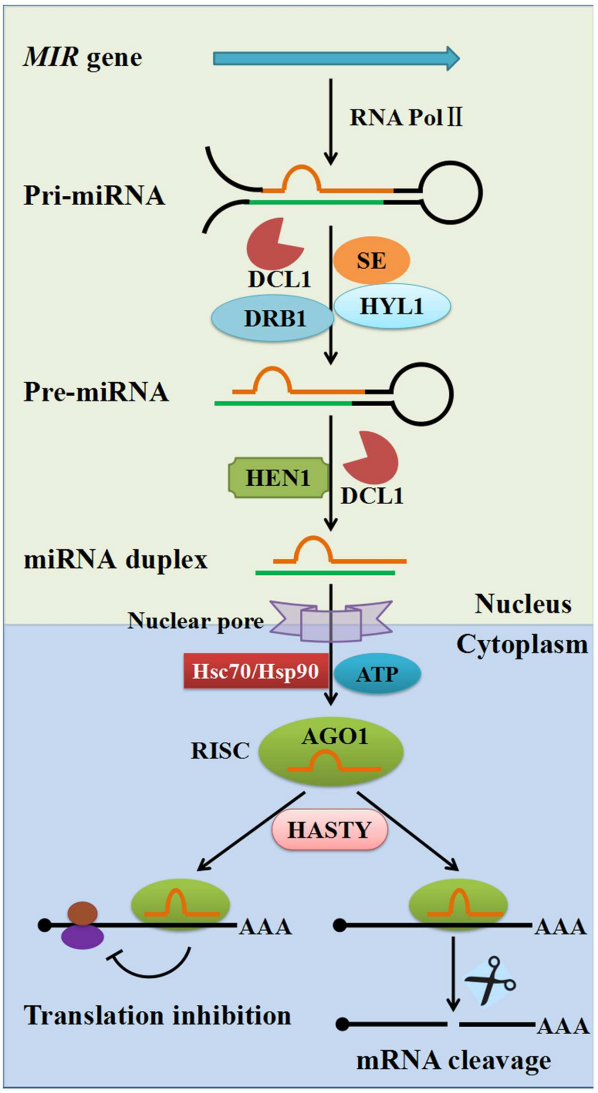


FIGURE 1 | The biogenesis and regulation mechanisms of plant miRNAs. Plant pri-miRNAs are mostly produced from *MIR* genes by RNA polymerase II (Pol II). Pri-miRNAs are cleaved into pre-miRNAs by DCL1 with the assistance of SE, dsRBP and HYL1. Pre-miRNAs are further processed into 21–24 nucleotide duplex miRNAs by the combined action of DCL1 and HEN1. Duplex miRNAs are methylated by HEN1 into mature miRNA duplexes and are exported to the cytoplasm through the action of plant exportin 5 ortholog HASTY. The guide-strand (red) is then loaded onto an AGO protein with the help of Hsc70/Hsp90 chaperone and ATP, followed by passenger strand (green) ejection, to form a RISC. There are two modes of plant miRNA action in cytoplasm: in one, the miRNA regulates its target at the protein level through translational inhibition (left); in the other, the miRNA regulates its target at the mRNA level through mRNA cleavage (right). AGO1, Argonaute 1; DCL1, Dicer-like1; SE, Serrate; HEN1, Hua enhancer1; DRB1, Double-strand RNA binding protein1; HYL1, Hyponastic leaves1; RISC, RNA-induced silencing complex.

is also commonly found in plants (Brodersen et al., 2008; Djuranovic et al., 2012; Iwakawa and Tomari, 2013; Li et al., 2013).

The first miRNA (lin-4) was discovered in *Caenorhabditis elegans* (Lee et al., 1993), and a large number of miRNAs have since been identified in animals and plants. Initially, miRNAs were considered to be a consequence of the evolution of multicellularization, but it was later discovered that the unicellular green alga (*Chlamydomonas reinhardtii*) also encodes miRNAs (Molnar et al., 2007; Zhao et al., 2007), suggesting that the miRNAs pathway evolved prior to the divergence between unicellular algae and land plants. Moreover, most miRNA families in *Arabidopsis* have homologs in other plants, and several miRNA–mRNA target pairs are consistently conserved in primitive multicellular land plants (Bartel and Bartel, 2003; Jones-Rhoades, 2012; Zhang et al., 2013), suggesting that the miRNA has an ancient origin.

Three main models for the emergence and evolution of *MIR* genes in plant genomes have been suggested (Voinnet, 2009; Zhao et al., 2015; Zhang Y. et al., 2016). First, miRNAs are generated from the inverted duplication events of their target gene sequences (Allen et al., 2004; Maher et al., 2006); second, miRNAs originate from a variety of small-to-medium sized fold-back sequences distributed throughout the genome, termed 'spontaneous evolution' (Felippes et al., 2008); and third, DNA-type non-autonomous elements, namely miniature inverted-repeat transposable elements (MITEs) can readily fold into imperfect stem–loop structures of miRNA precursors (Piriyapongsa and Jordan, 2008). Because all life forms must survive their corresponding viruses, it is conceivable that host antiviral systems are essential in all living organisms (Villarreal, 2011). Indeed, viruses are crucial in the origin and evolution of host antiviral systems (Villarreal and Witzany, 2010; Villarreal, 2011). Although plant DNA viruses such as pararetroviruses and geminiviruses generally form episomal minichromosomes, illegitimate integration of these viruses in the plant genome is well documented (Hohn et al., 2008; Ghoshal and Sanfacon, 2015). Studies have also shown that cDNA sequences of plant RNA viruses can integrate into plant genomes, although plant RNA viruses are normally replicated in the cytoplasm of the infected cells (Hohn et al., 2008; Chiba et al., 2011). In addition, somatic endogenization may occur frequently, although it remains undetected because it is not passed on to the next generation (Covey and Al-Kaff, 2000). Remarkably, 24-nt sRNAs derived from an endogenous pararetrovirus sequence were found to accumulate to high levels in *Fritillaria imperialis* L. plants (Becher et al., 2014). Therefore, plant miRNAs may originate from viruses, such as virus-encoded miRNAs or miRNAs derived from the viral genome that integrated into the host genome. Two studies suggest the existence of virus-encoded miRNAs that may have been derived from *Sugarcane streak mosaic virus* (SCSMV) and *Hibiscus chlorotic ringspot virus* (HCRSV), respectively, but their functions remain to be elucidated (Gao et al., 2012; Viswanathan et al., 2014). In contrast, virus-encoded miRNAs have been identified extensively and are critical regulators of gene expression in animal–virus interactions (Nair and Zavolan, 2006; Grundhoff and Sullivan, 2011; Wang and Smith, 2016). However, more evidence is needed for the existence of plant virus–derived miRNAs.

miRNAs AND PLANT ANTIVIRAL DEFENSE

The successful survival of plants crucially depends upon their ability to exploit numerous defense mechanisms against invading pathogens or hostile environments. siRNA-mediated gene silencing is one of the most important strategies of plants against viral infections (Wang et al., 2012; Pumplin and Voinnet, 2013; Ghoshal and Sanfacon, 2015; Moon and Park, 2016; Khalid et al., 2017). There are two main advantages of siRNA-mediated gene silencing: the defensive signal can spread, and siRNA is transitive (Lu et al., 2008; Eamens et al., 2008). However, siRNA-mediated gene silencing is triggered only after viruses have invaded the host, thus infected cells are unable send a warning message to non-infected cells until the initial attack by viruses. Therefore, siRNA-mediated gene silencing may be insufficient to resist invading viruses, and a proactive mechanism is necessary. miRNAs are endogenous RNAs, some of miRNAs which exist within a cell prior to viral invasion while some miRNAs are induced previously in response to other stimuli or pathogens, indicating that these miRNAs can serve as advance preparation to counteract or evade the invading virus (Lu et al., 2008). Plant miRNAs have evolved to optimize cleavage efficiency rather than maximize complementarily to their targets (Voinnet, 2009; Jones-Rhoades, 2012). Three or more mismatches are permitted between miRNA and its target, which thereby significantly expands the spectrum of targets and facilitates the release of the cleaved target RNAs from the RISC complex. In plants, two main modes have been suggested for the roles of miRNAs in an antiviral defense response: a direct mode through targeting viral RNAs, and an indirect mode through triggering the biogenesis of siRNA responsible for the antiviral response.

Endogenous miRNAs have been shown to play an important role in the suppression of invading viruses in mammals (Gottwein and Cullen, 2008). In plants, miR393 was the first endogenous miRNA recognized to function in antibacterial resistance by suppressing auxin signaling (Navarro et al., 2006). In the same year, Simon-Mateo and Garcia (2006) demonstrated that *Plum pox virus* (PPV) chimeras bearing plant miRNA target sequences, which have been reported to be functional in *Arabidopsis*, were affected by miRNA function in three different host plants (Simon-Mateo and Garcia, 2006). In addition, several studies have shown that miRNA-mediated post-transcriptional regulation is involved in plant defensive responses against viral infections (Amin et al., 2011; Li et al., 2012; Pacheco et al., 2012). A recent study showed that cotton plants can export miRNAs to inhibit virulence gene expression in the fungal pathogen *Verticillium dahlia* (Zhang T. et al., 2016). The authors found that two genes encoding a Ca^{2+}-dependent cysteine protease (*Clp-1*) and an isotrichodermin C-15 hydroxylase (*HiC-15*) targeted by miR166 and miR159, respectively, are both indispensable for *V. dahlia* virulence. Nevertheless, most studies provide indirect evidence for the first mode of plant miRNA function being direct

targeting of viral RNAs, and more studies are needed to clarify this mode of action.

Plant genomes contain a large number of leucine-rich repeat (LRR) and nucleotide binding (NB)-LRR immune receptors encoded by resistance (*R*) genes, which recognize specific pathogen effectors and trigger resistance responses. To a great extent, the siRNA-mediated gene silencing involved in antiviral defense occurs through regulation of these *R* genes. Studies have shown that plant miRNAs target and negatively regulate plant *R* genes by prompting the production of phased, *trans*-acting siRNAs (tasiRNAs) against these *R* genes, and this miRNA-mediated gene regulation is suppressed on bacterial or viral infection (Zhai et al., 2011; Li et al., 2012). In *Medicago truncatula*, these 'anti-*R* gene' siRNAs are produced from dsRNA with the assistance of RNA-dependent RNA polymerase 6 (RDR6), DCL4, and DRB4 following the cleavage of certain *R* gene transcripts by miR482, a scheme that is similar to that of tasiRNA production (Zhai et al., 2011) (**Figure 2**). In tomato, miR482 can target a conserved sequence from 58 coiled coil (CC)-NB-LRR proteins, resulting in cleavage of *R* gene mRNA and production of secondary siRNAs in an RDR6-dependent manner (Shivaprasad et al., 2012). In tobacco, the *R* gene *N* against TMV, the first *R* gene conferring resistance to a virus to be identified, was found to undergo regulation by miR482 (Whitham et al., 1994; Li et al., 2012). In total, the silencing of NBS-LRR genes by miR482, and their activation after miR482 down-regulation upon bacterial or viral treatments, have been widely studied in different plants (Li et al., 2012; Shivaprasad et al., 2012; Zhu et al., 2013; Yang et al., 2015). Similarly, Li et al. (2012) demonstrated that miR6019 and miR6020 in tobacco cause specific cleavage of transcripts of the *N* gene and its homologs by binding to the complementary sequence of the conserved Toll and Interleukin-1 receptors (TIR)-encoding domain of the N transcript (Li et al., 2012; Moon and Park, 2016). Moreover, synthesis of phased, secondary siRNAs (phasiRNAs) from the *N* coding sequence through overexpression of miR6019 was shown to be accompanied by reductions in *N* transcript accumulation and *N*-mediated resistance against TMV (Li et al., 2012). Taken together, these results suggest that the miRNA-mediated gene silencing response is integrated with *R* gene–mediated antiviral defense responses.

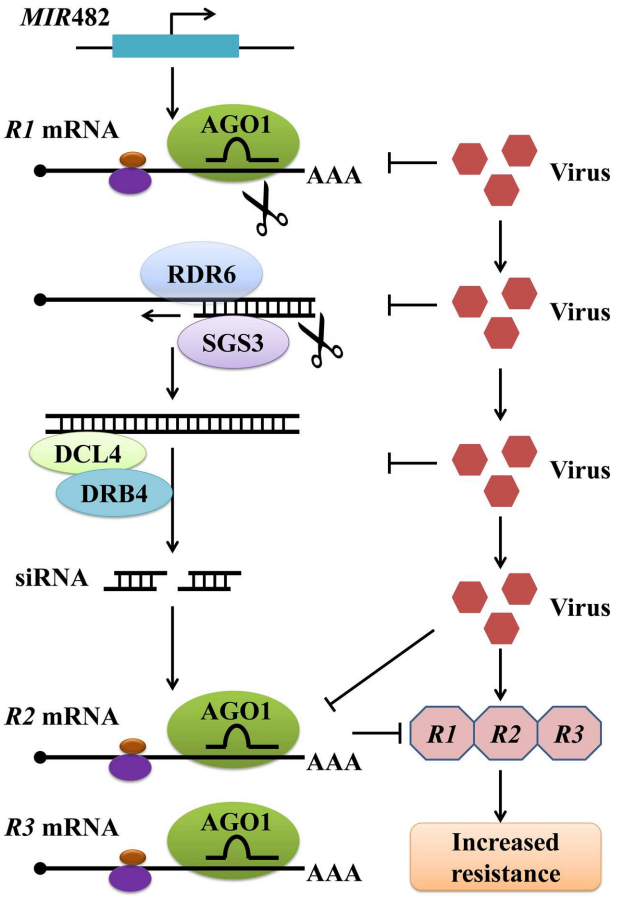


FIGURE 2 | A pathway of plant miRNA482-mediated resistance against virus by inducing the production of siRNAs responsible for regulating *R* genes. The expression of *MIR*482 triggers the simultaneous silencing of multiple *R* genes through tasiRNAs produced from dsRNA derived from a primary miR482-targeted *R* gene. Virus infection may trigger silencing suppression involved in this process at several steps, resulting in increased accumulation of multiple *R* proteins and enhanced resistance. AGO1, Argonaute 1; DCL4, Dicer-like 4; RDR6, RNA-dependent RNA polymerase 6; SGS3, Suppressor of gene silencing 3; DRB4, Double-stranded RNA binding protein 4.

miRNAs AND VIRAL COUNTER-DEFENSE

Viruses have evolved numerous strategies to counteract or evade host defenses mediated by RNA silencing, such as the deployment of decoy RNAs, specialized replication mechanisms, and sequestration of viral RNAs in large protein or membrane complexes (Ghoshal and Sanfacon, 2015; Nie and Molen, 2015). Almost all plant viruses encode viral suppressors of RNA silencing (VSRs), which in addition to their functions in viral replication, encapsidation, or movement, interfere with host RNA silencing through multiple modes of action (Burgyan and Havelda, 2011; Wang et al., 2012). VSRs contribute to viral symptoms in two main ways: facilitating virus accumulation indirectly and modifying endogenous siRNA- or miRNA-mediated regulation directly (Silhavy and Burgyan, 2004; Burgyan and Havelda, 2011). In general, most VSR-mediated inhibition of RNA silencing occurs through two modes of action: (1) some VSRs sequester small RNA duplexes by binding to short or long dsRNAs, resulting in the suppression of the assembly of AGOs into RISCs; (2) some VSRs physically interact with AGO1 to prevent siRNA or miRNA loading, impede slicing activity, or degrade the AGO1 protein (Burgyan and Havelda, 2011; Wang et al., 2012; Moon and Park, 2016).

The molecular basis of viral symptom development depends upon the ability of VSRs to interfere with plant miRNA biogenesis, eventually affecting mRNA turnover to the advantage of invaders (Chapman et al., 2004; Chen et al., 2004; Ramesh et al., 2014). The tombusvirus P19 protein is one of the best-studied VSRs that play critical roles in plant–virus interactions

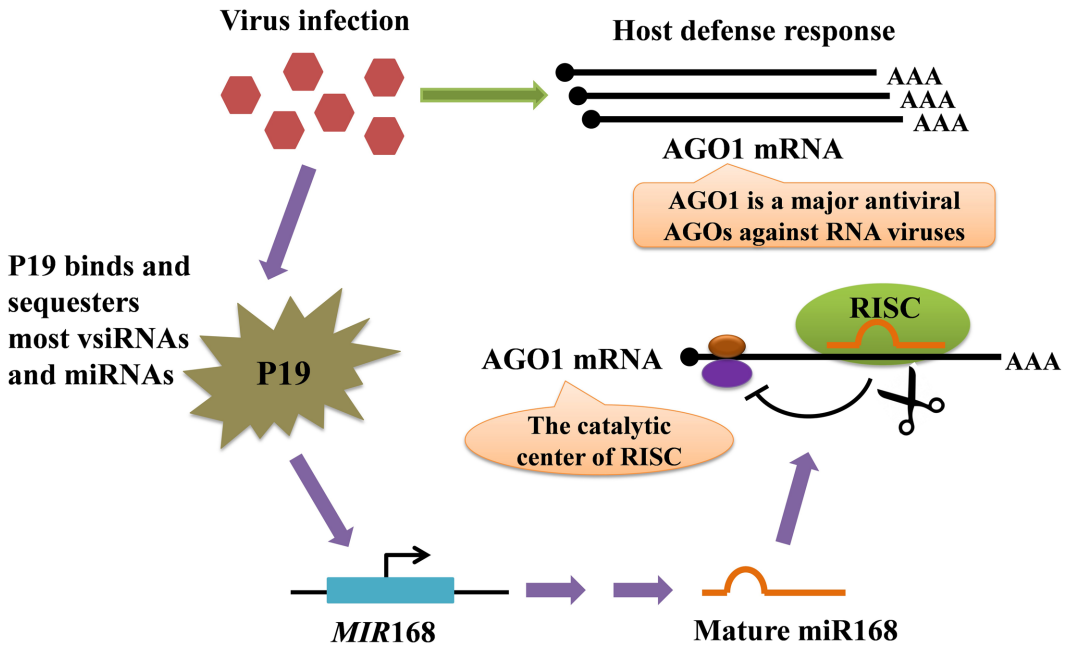


FIGURE 3 | Model for the regulation of the AGO1 mRNA level mediated by tombusviral protein P19-induced miR168. The AGO1 protein plays a central role in plant defensive response against pathogens, and viral infection induces enhanced expression of AGO1 mRNA. Meanwhile, the virus produces the P19 VSR. P19 forms head-to-tail homodimers that bind to virus-encoded siRNAs (vsiRNAs), siRNAs and endogenous miRNAs with high affinity, preventing their loading into AGO1. However, miR168 is not efficiently bound by P19, resulting in the increased loading of miR168 into AGO1. Because miR168 directly represses the AGO1 mRNA, the accumulation of antiviral AGO1 is sharply decreased. In addition, tombusvirus infection also stimulates *MIR*168 transcription, and the expression of AGO1 mRNA is consequently further repressed by the increased miR168 level. Therefore, P19 VSR can not only sequester small RNAs, but can also effectively inhibit the loading of viral siRNAs onto AGO1.

(Omarov et al., 2006; Várallyay et al., 2010) (**Figure 3**). The P19 binds and sequesters most miRNAs and virus-derived siRNAs (vsiRNAs) to suppress their activity in AGO proteins but is selectively unable to bind miR168, resulting in the increased loading of miR168 into AGO1 and the subsequent reduced accumulation of AGO1. Because miR168 directly down-regulates AGO1 mRNA stability and translation, this selective binding process not only causes the direct siRNA sequestration by P19 but also sharply reduces the cellular AGO1 levels (Várallyay et al., 2010; Pumplin and Voinnet, 2013). Tombusvirus infection also stimulates *MIR*168 transcription in a silencing inhibition–dependent manner, resulting in further increased levels of miR168 responsible for AGO1 down-regulation. Similar results have been observed during infections by other viruses, supporting that diverse VSRs convergently arrest endogenous silencing against the antiviral silencing pathway (Várallyay and Havelda, 2013). Notably, *African cassava mosaic virus* (ACMV) AC4, has been shown to bind directly to certain miRNAs, thereby making mi-RISC non-functional, and thus AC4 over-expressing transgenic plants showed reduced accumulation of miRNAs (Chellappan et al., 2005). Similarly, it is possible that *Tomato leaf curl new delhi virus* (ToLCNDV) AC4 might act to destabilize miRNAs which explains the reduction in the levels of certain miRNAs (Naqvi et al., 2010). In addition, a study demonstrated that *Rice stripe virus* (RSV) infections influenced small RNA profiles in rice, and that RSV induced the expression of novel miRNAs from conserved miRNA precursors (Du et al., 2011). These results suggest that VSRs and viral infection lead to major changes in the miRNA-mediated gene silencing pathway in plants.

Alternatively, some VSRs inhibit the activity of AGO proteins that have a central role in the antiviral RNA silencing (Wang et al., 2012; Carbonell and Carrington, 2015). For instance, *Sweet potato mild mottle virus* (SPMMV) P1 and *Turnip crinkle virus* (TCV) coat protein (CP or P38), directly interact with AGO proteins through conserved GW/WG repeat motifs, which resemble the AGO1-binding peptides on RISC (Giner et al., 2010; Moon and Park, 2016). In addition, Duan et al. (2012) demonstrated that *Cucumber mosaic virus* (CMV) 2b protein suppresses the activity of RISC by physically interacting with the PAZ domain of AGO1. These observations suggest that VSR suppression of RNA silencing may be associated with independently evolved VSRs that show functional overlap (Moon and Park, 2016).

Although some viruses can specifically disable host defense through encoding proteins, most viruses harbor limited coding capacity. Thus, the miRNAs become efficient and accessible tools to regulate their own gene expression and that of their host cells (Sullivan and Ganem, 2005; Nair and Zavolan, 2006). The first virus–encoded miRNAs were identified from a cloning experiment in human B cells latently infected with the herpesvirus Epstein-Barr virus (EBV) (Pfeffer et al., 2004). Subsequently, hundreds of animal virus–encoded miRNAs

were discovered in various viruses such as herpesviruses, polyomaviruses, and adenoviruses (Gottwein and Cullen, 2008). Some animal virus–encoded miRNAs can effectively regulate viral gene expression and modulate the host's miRNA-mediated gene silencing (Pfeffer et al., 2004; Nair and Zavolan, 2006; Roberts et al., 2011). During the counter-defense response, these animal virus–encoded miRNAs facilitate infection by regulating virus gene expression to increase virulence (Lu et al., 2008; Pumplin and Voinnet, 2013; Huang et al., 2016). The targets of viral miRNAs might be viral mRNAs or host cellular mRNAs, suggesting that viruses can employ miRNAs to regulate the cellular environment to support the viral life cycle (Roberts et al., 2011). In plants, numerous virus–derived siRNAs (vsiRNAs) or viroid–related siRNAs have been identified, and they play diverse functions in plant–virus interactions (Shimura et al., 2011; Smith et al., 2011; Avina-Padilla et al., 2015; Huang et al., 2016). In contrast, little evidence supports the existence of plant virus–encoded miRNAs, although two studies have suggested that they do exist (Gao et al., 2012; Viswanathan et al., 2014). A potential explanation for why metazoan virus–encoded miRNAs exist, while plant virus–encoded miRNAs have yet to be uncovered, may depend on the mode of action of animal infecting viruses (Ramesh et al., 2014). In fact, most of the mammalian viruses known to encode miRNAs have much larger genomes than most plant viruses, and those genomes are DNA rather than RNA, which is the most common type of genomic material for plant viruses (Wang et al., 2012). Consequently, for viruses with RNA genomes it would be at a fitness disadvantage if they encoded regions that were prone to endonucleolytic cleavage by DCL proteins or other mechanisms (Grundhoff and Sullivan, 2011; Roberts et al., 2011). The DNA viruses known to encode miRNAs replicate in the nucleus, while most plant viruses typically replicate in the cytoplasm where a miRNA precursor would be more exposed to cleavage that would likely inhibit replication of the virus carrying it as part of its genome (Grundhoff and Sullivan, 2011; Wang et al., 2012). Therefore, based on the requirements of nuclear machinery and RNA cleavage for miRNA processing, it is unsurprising that cytoplasmic replicating DNA viruses and RNA viruses have not been found to express miRNAs (Boss and Renne, 2011). Nevertheless, detection of both viral strands of *Turnip mosaic virus* (TuMV) within the nucleus showed that RNA viruses do enter the nucleus (Ramesh et al., 2014). In addition, some plant DNA viruses have been identified, such as Geminiviridae and Nanoviridae with DNA genomes which replicate through a dsDNA replicative intermediate (Hohn and Vazquez, 2011).

miRNAs INVOLVED IN THE CO-EVOLUTION OF PLANTS AND VIRUSES

During the course of evolution, plants have evolved diverse strategies to counteract viral infection. Viruses have in turn evolved multiple mechanisms to counteract silencing, most obviously through the expression of VSRs. Interestingly, plants have also evolved specific defenses against RNA-silencing suppression by pathogens (Pumplin and Voinnet, 2013; Sansregret et al., 2013). The involvement of miRNAs in the never-ending arms race between plants and viruses has been summarized in **Figure 4**. As has been shown, some plant endogenous miRNAs can inhibit the expression of the plant's own genes against invading viruses, and in addition some plant miRNAs can facilitate viral mRNA cleavage or inhibit viral mRNA translation. In the viral counter-defense mechanism, VSRs can efficiently inhibit host antiviral responses by interacting with host *R* genes, which are regulated by one or multiple miRNAs that are responsible for cellular silencing machinery. Also noteworthy here is a direct interaction between VSR and *R*-mediated defense that appears to be independent of the host RNA silencing pathways (Wang et al., 2012). For instance, the CMV 2b VSR suppressed salicylic acid–mediated defense response (Ji and Ding, 2001) while the HC-Pro VSR of *Potato virus Y* (PVY) was found to induce defense responses (Shams-Bakhsh et al., 2007), indicating that some VSRs are recognized by the host defense mechanism to induce antiviral resistance. In addition, **Figure 4** also illustrates a hypothesis that plant virus–derived miRNAs can inhibit viral mRNA, host mRNA, or both, though this remains to be verified.

In fact, miRNA-mediated gene silencing provides a selective force in shaping plant viral genomes (Ramesh et al., 2014; Ghoshal and Sanfacon, 2015). Additionally, the selective pressure of being targeted by host–encoded miRNAs and the ability of virus–encoded miRNA to target host genes may also have greatly contributed to the evolution of viral genomes (Wang et al., 2012; Incarbone and Dunoyer, 2013). Single nucleotide polymorphisms (SNPs) that inhibit viral miRNA-directed silencing of certain host genes may be positively selected in the viral genome. Likewise, sequence variations of the viral genome that prevent viruses from being targeted by host-encoded miRNAs might also be under positive selection during evolution. Viruses exist as mixtures of minor sequence variants, and their replication has a relatively high error rate. The rapid evolution of the viral genome may have contributed enormously to minimizing host miRNA-directed gene silencing in facilitating viral infection in a specific plant–virus interaction. An observation was that the viral genome can evolve rapidly against the suppression of host-derived miRNAs in PPV chimeras containing genomic miRNA target sites (Simon-Mateo and Garcia, 2006). Similarly, the evolutionary stability of amiRNA-mediated resistance against TuMV was evaluated by experiments, revealing that TuMV evade RNA silencing by rapidly accumulating mutations in the target regions (Lin et al., 2009). However, variations in a plant genome caused by viral infection can also contribute positively to its genome evolution by increasing genetic and epigenetic diversity. Notably, virus infections of endemic vegetation typically induce only mild symptoms, or the infections are latent, presumably as a result of co-evolution and selection of viruses that do not kill or seriously harm their hosts, and may even induce systemic acquired resistance against other pathogens (Lovisolo et al., 2003; Fraile and García-Arenal, 2010). In a sense, viruses are not just harmful pathogens, but also beneficial symbionts of plants (Villarreal, 2011). The co-evolution of pathogens and

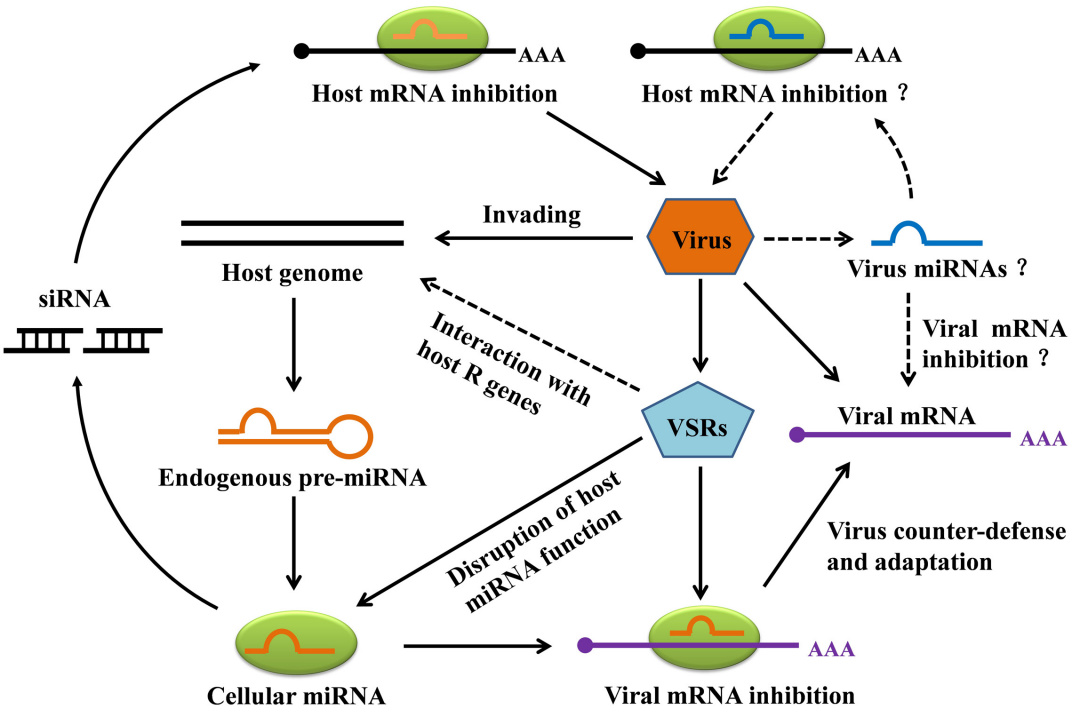


FIGURE 4 | Hypothetical model for miRNA-mediated gene silencing in plant-virus interaction. Invading viruses can trigger the production of plant endogenous miRNAs. In addition to specifically repressing the expression of host genes, the miRNAs can target viral mRNA for degradation. To counteract host miRNA-mediated gene silencing, viruses most notably express viral suppressors of RNA silencing (VSRs) to avoid host RNA silencing. VSRs can not only interfere with host miRNA function, but can also repress naturally occurring silencing of host *R* genes. Although plant virus-derived miRNAs have not yet been discovered, we illustrate hypothetical viral strategies, including miRNA-mediated repression of the host and viral transcripts.

their hosts thereby facilitates the production of diverse sRNAs. Overall, miRNAs play diverse roles in plant defensive systems, but their functions in antiviral defense are far from being completely elucidated.

THE APPLICATION OF miRNAs IN PLANT–VIRUS INTERACTIONS

Versatile plant biotechnologies, including antisense suppression, transcriptional gene silencing (TGS), virus-induced gene silencing (VIGS) and RNA interference (RNAi), are currently being used in plant antiviral biotechnology. In addition, artificial miRNA (amiRNA) is another robust biotechnology used in plants for silencing of genes, and engineering of amiRNAs has been widely applied for the targeted down-regulation of endogenous genes in various plants (**Table 1**). Given its efficacy and reliability, host-derived endogenous precursor miRNA has been commonly used as a structural backbone to replace the original ~21 nt long miRNA sequence with a region complementary to the target viral genome (Schwab et al., 2006; Ramesh et al., 2014; Khalid et al., 2017). The PPV was modified to include *Arabidopsis* miRNA target sequences, and the engineered virus had clearly impaired infectivity due to *Nicotiana clevelandii* and *Nicotiana benthamiana* miRNA, although the behaviors of PPV chimeras vary in different plants (Simon-Mateo and Garcia, 2006). Multiple-target miRNAs can also simultaneously influence several viruses. For instance, miRNA precursors containing complementary sequences with *Turnip yellow mosaic virus* (TYMV) and TuMV were designed, and the transgenic *Arabidopsis* expressing the recombinant miRNA precursors displayed specific resistance to these viruses (Niu et al., 2006; Ai et al., 2011). In wheat, Fahim et al. (2012) developed an amiRNA strategy against *Wheat streak mosaic virus* (WSMV) by incorporating five amiRNAs within one polycistronic amiRNA precursor. These designed amiRNAs replaced the natural miRNAs in each of the five arms of the polycistronic rice miR395, producing an amiRNA precursor known as *FanGuard* (FGmiR395), which was transformed into wheat, leading to the transgenic plants resistance to WSMV. Recently, Sun et al. (2016) constructed three dimeric amiRNA precursor expression vectors that target the 3-proximal part of *CP* genes of RSV and *Rice black streaked dwarf virus* (RBSDV) based on the structure of the rice osa-MIR528 precursor. The transgenic rice plants showed high resistance simultaneously against RSV and RBSDV infection at a low temperature (Sun et al., 2016). Thus far, engineering of amiRNA for antiviral resistance has been used successfully in various plant species, including *N. benthamiana* (Qu et al., 2007; Ai et al., 2011; Kung et al., 2012; Ali et al., 2013; Song et al., 2014; Mitter et al., 2016; Wagaba et al., 2016; Carbonell and Daros, 2017), *Arabidopsis* (Duan et al., 2008; Lin et al., 2009), rice (Sun et al., 2016), wheat

TABLE 1 | Engineering of plant miRNA for antiviral immunity.

Plant species	MiRNA backbone	Virus	Target viral region/gene	Reference
Arabidopsis thaliana	*Arabidopsis* pre-miR159	TYMV, TuMV	P69, HC-Pro (coat protein)	Niu et al., 2006
Nicotiana benthamiana	*Arabidopsis* miR159a, miR167b, and miR171a	PPV	P1/HC-Pro	Simon-Mateo and Garcia, 2006
Nicotiana benthamiana	*Arabidopsis* pre-miR171a	CMV	2b viral gene	Qu et al., 2007
Arabidopsis thaliana,	*Arabidopsis* pre-miR159	CMV	3′-UTR	Duan et al., 2008
Arabidopsis thaliana, Nicotiana benthamiana	*Arabidopsis* pre-miR159	TuMV	P69	Lin et al., 2009
Nicotiana tabacum	*Arabidopsis* miR159a, miR167b, and miR171a	PVY PVX	HC-Pro, TGBp1/p25 (p25)	Ai et al., 2011
Solanum lycopersicum	*Arabidopsis* pre-miR159a	CMV	2a and 2b viral genes, 3′-UTR	Zhang et al., 2011
Nicotiana benthamiana	*Arabidopsis* pre-miR159a	WSMoV	Conserved motifs of L (replicase) gene (A, B1, B2, C, D, E, AB1E, B2DC)	Kung et al., 2012
Triticum	Rice miR395	WSMV	Conserved region	Fahim et al., 2012
Vitis vinifera	*Arabidopsis* pre-miR319a	GFLV	Coat protein (CP)	Jelly et al., 2012
Nicotiana benthamiana	Cotton pre-miR169a	CLCuBuV	V2 gene	Ali et al., 2013
Solanum lycopersicum	*Arabidopsis* pre-miR319a, Tomato pre-miR319a and pre-miR168a	ToLCV	The middle region of the AV1 (coat protein), the overlapping region of the AV1 and AV2 (pre-coat protein)	Vu et al., 2013
Nicotiana benthamiana Nicotiana. tabacum	*Arabidopsis* pre-miR319a	PVY	CI, NIa, NIb, CP	Song et al., 2014
Zea mays	Maize pre-miR159a	RBSDV	Conserved region	Xuan et al., 2015
Nicotiana benthamiana	Barley pre-miR171	WDV	Conserved region	Kis et al., 2016
Oryza sativa	Rice pre-miR528	RSV, RBSDV	Middle segment, 3′ end and 3′-UTR region of the CP gene	Sun et al., 2016
Nicotiana benthamiana	*Arabidopsis* pre-miR159a	CBSV, UCBSV	P1, P3, CI, Nib and CP	Wagaba et al., 2016
Nicotiana benthamiana	*Arabidopsis* pre-miR159a	TSWV	N, NSs	Mitter et al., 2016
Nicotiana benthamiana	Six amiRNAs	PSTVd	Structural domains	Carbonell and Daros, 2017

TYMV, Turnip yellow mosaic virus (Potyviridae); TuMV, Turnip mosaic virus (Potyviridae); CMV, Cucumber mosaic virus (Bromoviridae); PPV, Plum pox virus (Potyviridae); PVY, Potato virus Y (Potyviridae); PVX, Potato virus X (Alphaflexiviridae); WSMoV, Watermelon silver mottle virus (Bunyaviridae); WSMV, Wheat streak mosaic virus (Potyviridae); GFLV, Grapevine fan leaf virus (Secoviridae); CLCuBuV, Cotton leaf curl Burewala virus (Geminiviridae); WDV, Wheat dwarf virus (Geminiviridae); RSV, Rice stripe virus (unassigned); RBSDV, Rice black streaked dwarf virus (Reoviridae); CBSV, Cassava brown streak virus (Potyviridae); UCBSV, Ugandan cassava brown streak virus (Potyviridae); TSWV, Tomato spotted wilt virus (Bunyaviridae); PSTVd, Potato spindle tuber viroid (Pospiviroidae); ToLCV, Tomato leaf curl virus (Geminiviridae).

(Fahim et al., 2012), maize (Xuan et al., 2015), tomato (Zhang et al., 2011; Vu et al., 2013), and grapevine (Jelly et al., 2012) (**Table 1**). Apart from being used in plant antiviral immune systems, engineering of amiRNA has been extensively applied in plant resistance against other pathogens such as bacteria (Navarro et al., 2006; Li et al., 2010; Boccara et al., 2014; Ma et al., 2014), and fungi (Liu et al., 2014; Ouyang et al., 2014; Xu et al., 2014). These studies indicate that plant amiRNA biotechnology could be of broad utility in increasing plant resistance against pathogens.

Previous studies revealed that the efficiency of miRNA to target viral RNAs depends not only on their nature but also on their inserted positions or the local structures of the target mRNAs (Simon-Mateo and Garcia, 2006; Duan et al., 2008). The accessibility of target sequences for amiRNA silencing is a pivotal factor for consideration. An experimental approach was used to determine the accessible cleavage hotspots on viral RNA by comparing the viral-derived siRNAs from wild-type *Arabidopsis* with sRNAs derived from those of the DCL mutants. The target viral transcript is assessed for DCL susceptibility and the vulnerable region was identified, thereby antiviral amiRNAs could be deployed (Duan et al., 2008). It is intriguing that the miRNA-mediated gene silencing mechanism or processing can also be affected by the flanking sequence in addition to the miRNA itself. The reasonable explanation is that RNA folding influences the binding sites between miRNAs and their target sequences (Lafforgue et al., 2013; Liu et al., 2016). Therefore, the insertion sites and the flanking sequence should be carefully validated when amiRNA-mediated gene silencing is established.

Engineering of amiRNAs possesses several advantages, including fewer off-target effects, high RNA promoter compatibility, high stability *in vivo*, high accuracy and the ability to degrade target genes without affecting expression of other genes, heritability of phenotypes, and environmental biosafety (Lu et al., 2008; Ramesh et al., 2014; Tiwari et al., 2014). Nevertheless, using amiRNA has several problems: (1) broad-spectrum amiRNAs are intractable to devise owing to the high sequence divergence of plant viruses; (2) the durability of

amiRNAs is a challenge if the amiRNA targets the non-conserved regions of plant viruses; (3) single amiRNA expressing transgenic plants under field conditions may be confronted with strong virus pressure, thereby the resistance of transgenic plants against viruses may not be sustained. Fortunately, considerable efforts have been made to overcome these obstacles. For example, Lafforgue and his colleagues established two alternative strategies to improve the effectiveness of amiRNA including the expression of two amiRNAs complementary to independent targets and the design of amiRNAs complementary to highly conserved RNA motifs in the viral genome (Lafforgue et al., 2013). In addition, polycistronic amiRNA-mediated resistance to WSMV was successfully and efficiently applied in wheat and barley, respectively (Fahim et al., 2012; Kis et al., 2016). Recently, the Plant Small RNA Maker Site (P-SAMS) tool[1], which serves as a high-throughput platform for the high efficiency design of amiRNA and synthetic *trans*-acting small interfering RNAs (syn-tasiRNA), has been established (Fahlgren et al., 2016). Collectively, there is still a long way to go for amiRNA engineering, although great progress has been made.

CONCLUSION

Increasing evidence has shown that miRNA-mediated gene silencing plays a critical role in plant resistance against invading viruses and other types of pathogens. Although much remains to be learned about the molecular mechanisms of miRNA-mediated gene silencing in plants, current understanding has already laid a foundation for developing molecular tools for crop improvements. Due to the multiple advantages of amiRNA-mediated gene silencing, it has emerged as a powerful technique and become one of the most important tools in genetic engineering. However, failure and inefficiency of amiRNA-mediated gene silencing have been observed in some instances, probably due to the lack of complete knowledge of miRNA processing procedures involving biochemical enzymes and miRNA recruiting machinery. Hence, understanding the overall mechanisms of miRNA biogenesis is critical, beginning with transcription initiation and extending to target gene cleavage or translational repression. In addition, elucidation of the molecular mechanisms underlying the interactions between plants and viruses with respect to miRNAs will enable us to more thoroughly obtain the benefits to be derived from the miRNA-mediated gene silencing mechanism. Future efforts should be directed not only at understanding how to explore the machinery of viruses in hijacking the host miRNA-mediated gene silencing, but also developing rapid and systemic amiRNA delivery strategies to integrate amiRNAs in the plant genome.

[1] http://p-sams.carringtonlab.org

AUTHOR CONTRIBUTIONS

S-RL wrote the paper, J-JZ, C-GH, C-LW, and J-ZZ wrote and edited the paper.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 31171608, 31360469, 31471863, 31521092, 31772252, and 31221062), the Special Innovative Province Construction in Anhui Province (15czs08032), and the Central Guiding the Science and Technology Development of the Local (2016080503b024).

REFERENCES

Ai, T., Zhang, L., Gao, Z., Zhu, C. X., and Guo, X. (2011). Highly efficient virus resistance mediated by artificial microRNAs that target the suppressor of PVX and PVY in plants. *Plant Biol. (Stuttg).* 13, 304–316. doi: 10.1111/j.1438-8677.2010.00374.x

Ali, I., Amin, I., Briddon, R. W., and Mansoor, S. (2013). Artificial microRNA-mediated resistance against the monopartite begomovirus Cotton leaf curl Burewala virus. *Virol. J.* 10:231. doi: 10.1186/1743-422X-10-231

Allen, E., Xie, Z., Gustafson, A. M., Sung, G. H., Spatafora, J. W., and Carrington, J. C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat. Genet.* 36, 1282–1290. doi: 10.1038/ng1478

Amin, I., Basavaprabhu, L. P., Briddon, R. W., Mansoo, S., and Fauquet, C. M. (2011). Common set of developmental miRNAs are upregulated in *Nicotiana benthamiana* by diverse begomoviruses. *Virol. J.* 8:143. doi: 10.1186/1743-422X-8-143

Avina-Padilla, K., Martinez de la Vega, O., Rivera-Bustamante, R., Martinez-Soriano, J. P., Owens, R. A., Hammond, R. W., et al. (2015). In silico prediction and validation of potential gene targets for pospiviroid-derived small RNAs during tomato infection. *Gene* 564, 197–205. doi: 10.1016/j.gene.2015.03.076

Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.* 64, 137–159. doi: 10.1146/annurev-arplant-050312-120043

Bartel, B. (2005). microRNA directing siRNA biogenesis. *Nat. Struct. Mol. Biol.* 12, 569–571. doi: 10.1038/nsmb0705-569

Bartel, B., and Bartel, D. P. (2003). MicroRNAs: at the root of plant development? *Plant Physiol.* 132, 709–717. doi: 10.1104/pp.103.023630

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi: 10.1016/S0092-8674(04)00045-5

Baulcombe, D. (2004). RNA silencing in plants. *Nature* 431, 356–363. doi: 10.1038/nature02874

Becher, H., Ma, L., Kelly, L. J., Kovarik, A., Leitch, I. J., and Leitch, A. R. (2014). Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome. *Plant J.* 80, 823–833. doi: 10.1111/tpj.12673

Boccara, M., Sarazin, A., Thiebeauld, O., Jay, F., Voinnet, O., Navarro, L., et al. (2014). The Arabidopsis miR472-RDR6 silencing pathway modulates PAMP- and effector-triggered immunity through the post-transcriptional control of disease resistance genes. *PLOS Pathog.* 10:e1003883. doi: 10.1371/journal.ppat.1003883

Bologna, N. G., and Voinnet, O. (2014). The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. *Annu. Rev. Plant Biol.* 65, 473–503. doi: 10.1146/annurev-arplant-050213-035728

Boss, I. W., and Renne, R. (2011). Viral miRNAs and immune evasion. *Biochim. Biophys. Acta* 1809, 708–714. doi: 10.1016/j.bbagrm.2011.06.012

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y. Y., Sieburth, L., et al. (2008). Widespread translational inhibition

by plant miRNAs and siRNAs. *Science* 320, 1185–1190. doi: 10.1126/science.1159151

Burgyan, J., and Havelda, Z. (2011). Viral suppressors of RNA silencing. *Trends Plant Sci.* 16, 265–272. doi: 10.1016/j.tplants.2011.02.010

Calil, I. P., and Fontes, E. P. (2016). Plant immunity against viruses: antiviral immune receptors in focus. *Ann. Bot.* 119, 711–723. doi: 10.1093/aob/mcw200

Carbonell, A., and Carrington, J. C. (2015). Antiviral roles of plant ARGONAUTES. *Curr. Opin. Plant Biol.* 27, 111–117. doi: 10.1016/j.pbi.2015.06.013

Carbonell, A., and Daros, J. A. (2017). Artificial microRNAs and synthetic trans-acting small interfering RNAs interfere with viroid infection. *Mol. Plant Pathol.* 18, 746–753. doi: 10.1111/mpp.12529

Chapman, E. J., Prokhnevsky, A. I., Gopinath, K., Dolja, V. V., and Carrington, J. C. (2004). Viral RNA silencing suppressors inhibit the microRNA pathway at an intermediate step. *Genes Dev.* 18, 1179–1186. doi: 10.1101/gad.1201204

Chellappan, P., Vanitharani, R., and Fauquet, C. M. (2005). MicroRNA-binding viral protein interferes with Arabidopsis development. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10381–10386. doi: 10.1073/pnas.0504439102

Chen, J., Xiang, W. L., Xie, D., Peng, J. R., and Ding, S. W. (2004). Viral virulence protein suppresses RNA silencing-mediated defense but upregulates the role of microRNA in host gene expression. *Plant Cell* 16, 1302–1313. doi: 10.1105/tpc.018986

Chiba, S., Kondo, H., Tani, A., Saisho, D., Sakamoto, W., Kanematsu, S., et al. (2011). Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLOS Pathog.* 7:e1002146. doi: 10.1371/journal.ppat.1002146

Covey, S. N., and Al-Kaff, N. S. (2000). Plant DNA viruses and gene silencing. *Plant Mol. Biol.* 43, 307–322. doi: 10.1023/A:1006408101473

Djuranovic, S., Nahvi, A., and Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science* 336, 237–240. doi: 10.1126/science.1215691

Du, P., Wu, J., Zhang, J., Zhao, S., Zheng, H., Gao, G., et al. (2011). Viral infection induces expression of novel phased microRNAs from conserved cellular microRNA precursors. *PLOS Pathog.* 7:e1002176. doi: 10.1371/journal.ppat.1002176

Duan, C. G., Fang, Y. Y., Zhou, B. J., Zhao, J. H., Hou, W. N., Zhu, H., et al. (2012). Suppression of Arabidopsis ARGONAUTE1-mediated slicing, transgene-induced RNA silencing, and DNA methylation by distinct domains of the Cucumber mosaic virus 2b protein. *Plant Cell* 24, 259–274. doi: 10.1105/tpc.111.092718

Duan, C. G., Wang, C. H., Fang, R. X., and Guo, H. S. (2008). Artificial MicroRNAs highly accessible to targets confer efficient virus resistance in plants. *J. Virol.* 82, 11084–11095. doi: 10.1128/JVI.01377-08

Eamens, A., Wang, M. B., Smith, N. A., and Waterhouse, P. M. (2008). RNA silencing in plants: yesterday, today, and tomorrow. *Plant Physiol.* 147, 456–468. doi: 10.1104/pp.108.117275

Fahim, M., Millar, A. A., Wood, C. C., and Larkin, P. J. (2012). Resistance to Wheat streak mosaic virus generated by expression of an artificial polycistronic microRNA in wheat. *Plant Biotechnol. J.* 10, 150–163. doi: 10.1111/j.1467-7652.2011.00647.x

Fahlgren, N., Hill, S. T., Carrington, J. C., and Carbonell, A. (2016). P-SAMS: a web site for plant artificial microRNA and synthetic trans-acting small interfering RNA design. *Bioinformatics* 32, 157–158. doi: 10.1093/bioinformatics/btv534

Felippes, F. F. D., Schneeberger, K., Dezulian, T., Huson, D. H., and Weigel, D. (2008). Evolution of *Arabidopsis thaliana* microRNAs from random sequences. *RNA* 14, 2455–2459. doi: 10.1261/rna.1149408

Fraile, A., and García-Arenal, F. (2010). The coevolution of plants and viruses: resistance and pathogenicity. *Adv. Virus Res.* 76, 1–32. doi: 10.1016/S0065-3527(10)76001-2

Gao, R., Liu, P., and Wong, S. M. (2012). Identification of a plant viral RNA genome in the nucleus. *PLOS ONE* 7:e48736. doi: 10.1371/journal.pone.0048736

Ghoshal, B., and Sanfacon, H. (2015). Symptom recovery in virus-infected plants: revisiting the role of RNA silencing mechanisms. *Virology* 479–480, 167–179. doi: 10.1016/j.virol.2015.01.008

Giner, A., Lakatos, L., Garcia-Chapa, M., Lopez-Moya, J. J., and Burgyan, J. (2010). Viral protein inhibits RISC activity by argonaute binding through conserved WG/GW motifs. *PLOS Pathog.* 6:e1000996. doi: 10.1371/journal.ppat.1000996

Gottwein, E., and Cullen, B. R. (2008). Viral and cellular microRNAs as determinants of viral pathogenesis and immunity. *Cell Host Microbe* 3, 375–387. doi: 10.1016/j.chom.2008.05.002

Grundhoff, A., and Sullivan, C. S. (2011). Virus-encoded microRNAs. *Virology* 411, 325–343. doi: 10.1016/j.virol.2011.01.002

Hohn, T., Richert-Poggeler, K. R., Staginnus, C., Harper, G., Schwartzacher, T., Teo, C. H., et al. (2008). "Evolution of integrated plant viruses," in *Plant Virus Evolution*, ed. M. Roossinck (Berlin: Springer), doi: 10.1007/978-3-540-75763-4-4

Hohn, T., and Vazquez, F. (2011). RNA silencing pathways of plants: silencing and its suppression by plant DNA viruses. *Biochim. Biophys. Acta* 1809, 588–600. doi: 10.1016/j.bbagrm.2011.06.002

Huang, J., Yang, M., Lu, L., and Zhang, X. (2016). Diverse functions of small RNAs in different plant-pathogen communications. *Front. Microbiol.* 7:1552. doi: 10.3389/fmicb.2016.01552

Incarbone, M., and Dunoyer, P. (2013). RNA silencing and its suppression: novel insights from in planta analyses. *Trends Plant Sci.* 18, 382–392. doi: 10.1016/j.tplants.2013.04.001

Iwakawa, H. O., and Tomari, Y. (2013). Molecular insights into microRNA-mediated translational repression in plants. *Mol. Cell* 52, 591–601. doi: 10.1016/j.molcel.2013.10.033

Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T., et al. (2010). Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Mol. Cell* 39, 292–299. doi: 10.1016/j.molcel.2010.05.015

Jelly, N. S., Schellenbaum, P., Walter, B., and Maillot, P. (2012). Transient expression of artificial microRNAs targeting Grapevine fanleaf virus and evidence for RNA silencing in grapevine somatic embryos. *Transgenic Res.* 21, 1319–1327. doi: 10.1007/s11248-012-9611-5

Ji, L. H., and Ding, S. W. (2001). The suppressor of transgene RNA silencing encoded by Cucumber mosaic virus interferes with salicylic acid-mediated virus resistance. *Mol. Plant Microbe Interact.* 14, 715–724. doi: 10.1094/MPMI.2001.14.6.715

Jones-Rhoades, M. W. (2012). Conservation and divergence in plant microRNAs. *Plant Mol. Biol.* 80, 3–16. doi: 10.1007/s11103-011-9829-2

Khalid, A., Zhang, Q., Yasir, M., and Li, F. (2017). Small RNA based genetic engineering for plant viral resistance: application in crop protection. *Front. Microbiol.* 8:43. doi: 10.3389/fmicb.2017.00043

Kis, A., Tholt, G., Ivanics, M., Varallyay, E., Jenes, B., and Havelda, Z. (2016). Polycistronic artificial miRNA-mediated resistance to Wheat dwarf virus in barley is highly efficient at low temperature. *Mol. Plant Pathol.* 17, 427–437. doi: 10.1111/mpp.12291

Kung, Y. J., Lin, S. S., Huang, Y. L., Chen, T. C., Harish, S. S., Chua, N. H., et al. (2012). Multiple artificial microRNAs targeting conserved motifs of the replicase gene confer robust transgenic resistance to negative-sense single-stranded RNA plant virus. *Mol. Plant Pathol.* 13, 303–317. doi: 10.1111/j.1364-3703.2011.00747.x

Lafforgue, G., Martinez, F., Niu, Q. W., Chua, N. H., Daros, J. A., and Elena, S. F. (2013). Improving the effectiveness of artificial microRNA (amiR)-mediated resistance against Turnip mosaic virus by combining two amiRs or by targeting highly conserved viral genomic regions. *J. Virol.* 87, 8254–8256. doi: 10.1128/JVI.00914-13

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843–854. doi: 10.1016/0092-8674(93)90529-Y

Li, F., Pignatta, D., Bendix, C., Brunkard, J. O., Cohn, M. M., Tung, J., et al. (2012). MicroRNA regulation of plant innate immune receptors. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1790–1795. doi: 10.1073/pnas.1118282109

Li, S., Liu, L., Zhuang, X., Yu, Y., Liu, X., Cui, X., et al. (2013). MicroRNAs inhibit the translation of target mRNAs on the endoplasmic reticulum in Arabidopsis. *Cell* 153, 562–574. doi: 10.1016/j.cell.2013.04.005

Li, Y., Zhang, Q., Zhang, J., Wu, L., Qi, Y., and Zhou, J. M. (2010). Identification of microRNAs involved in pathogen-associated molecular pattern-triggered plant innate immunity. *Plant Physiol.* 152, 2222–2231. doi: 10.1104/pp.109.151803

Lin, S. S., Wu, H. W., Elena, S. F., Chen, K. C., Niu, Q. W., Yeh, S. D., et al. (2009). Molecular evolution of a viral non-coding sequence under the selective pressure of amiRNA-mediated silencing. *PLOS Pathog.* 5:e1000312. doi: 10.1371/journal.ppat.1000312

Liu, J., Cheng, X., Liu, D., Xu, W., Wise, R., and Shen, Q. H. (2014). The miR9863 family regulates distinct Mla alleles in barley to attenuate NLR receptor-triggered disease resistance and cell-death signaling. *PLOS Genet.* 10:e1004755. doi: 10.1371/journal.pgen.1004755

Liu, S. R., Hu, C. G., and Zhang, J. Z. (2016). Regulatory effects of cotranscriptional RNA structure formation and transitions. *Wiley Interdiscip. Rev. RNA* 7, 562–574. doi: 10.1002/wrna.1350

Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 297, 2053–2056. doi: 10.1126/science.1076311

Lovisolo, O., Hull, R., and Rösler, O. (2003). Coevolution of viruses with hosts and vectors and possible paleontology. *Adv. Virus Res.* 62, 325–379. doi: 10.1016/S0065-3527(03)62006-3

Lu, Y. D., Gan, Q. H., Chi, X. Y., and Qin, S. (2008). Roles of microRNA in plant defense and virus offense interaction. *Plant Cell Rep.* 27, 1571–1579. doi: 10.1007/s00299-008-0584-z

Ma, C., Lu, Y., Bai, S., Zhang, W., Duan, X., Meng, D., et al. (2014). Cloning and characterization of miRNAs and their targets, including a novel miRNA-targeted NBS-LRR protein class gene in apple (Golden Delicious). *Mol. Plant* 7, 218–230. doi: 10.1093/mp/sst101

Maher, C., Stein, L., and Ware, D. (2006). Evolution of Arabidopsis microRNA families through duplication events. *Genome Res.* 16, 510–519. doi: 10.1101/gr.4680506

Mitter, N., Zhai, Y., Bai, A. X., Chua, K., Eid, S., Constantin, M., et al. (2016). Evaluation and identification of candidate genes for artificial microRNA-mediated resistance to tomato spotted wilt virus. *Virus Res.* 211, 151–158. doi: 10.1016/j.virusres.2015.10.003

Molnar, A., Schwach, F., Studholme, D. J., Thuenemann, E. C., and Baulcombe, D. C. (2007). miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* 447, 1126–1129. doi: 10.1038/nature05903

Moon, J. Y., and Park, J. M. (2016). Cross-talk in viral defense signaling in plants. *Front. Microbiol.* 7:2068. doi: 10.3389/fmicb.2016.02068

Nair, V., and Zavolan, M. (2006). Virus-encoded microRNAs: novel regulators of gene expression. *Trends Microbiol.* 14, 169–175. doi: 10.1016/j.tim.2006.02.007

Nakanishi, K. (2016). Anatomy of RISC: how do small RNAs and chaperones activate Argonaute proteins? *Wiley Interdiscip. Rev. RNA* 7, 637–660. doi: 10.1002/wrna.1356

Naqvi, A. R., Haq, Q. M., and Mukherjee, S. K. (2010). MicroRNA profiling of tomato leaf curl new delhi virus (tolcndv) infected tomato leaves indicates that deregulation of mir159/319 and mir172 might be linked with leaf curl disease. *Virol. J.* 7:281. doi: 10.1186/1743-422X-7-281

Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., et al. (2006). A Plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science* 312, 436–439. doi: 10.1126/science.aae0382

Nie, X., and Molen, T. A. (2015). Host recovery and reduced virus level in the upper leaves after Potato virus Y infection occur in tobacco and tomato but not in potato plants. *Viruses* 7, 680–698. doi: 10.3390/v7020680

Niu, Q. W., Lin, S. S., Reyes, J. L., Chen, K. C., Wu, H. W., Yeh, S. D., et al. (2006). Expression of artificial microRNAs in transgenic *Arabidopsis thaliana* confers virus resistance. *Nat. Biotechnol.* 24, 1420–1428. doi: 10.1038/nbt1255

Omarov, R., Sparks, K., Smith, L., Zindovic, J., and Scholthof, H. B. (2006). Biological relevance of a stable biochemical interaction between the Tombusvirus-encoded P19 and short interfering RNAs. *J. Virol.* 80, 3000–3008. doi: 10.1128/JVI.80.6.3000-3008.2006

Ouyang, S., Park, G., Atamian, H. S., Han, C. S., Stajich, J. E., Kaloshian, I., et al. (2014). MicroRNAs suppress NB domain genes in tomato that confer resistance to Fusarium oxysporum. *PLOS Pathog.* 10:e1004464. doi: 10.1371/journal.ppat.1004464

Pacheco, R., Garcı a-Marcos, A., Barajas, D., Martianez, J., and Tenllado, F. (2012). PVX-potyvirus synergistic infections differentially alter microRNA accumulation in *Nicotiana benthamiana*. *Virus Res.* 165, 231–235. doi: 10.1016/j.virusres.2012.02.012

Pfeffer, S., Zavolan, M., Grasser, F. A., Chien, M., Russo, J. J., Ju, J., et al. (2004). Identification of virus-encoded microRNAs. *Science* 304, 734–736. doi: 10.1126/science.1096781

Piriyapongsa, J., and Jordan, I. K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA.* 14, 814–821. doi: 10.1261/rna.916708

Pumplin, N., and Voinnet, O. (2013). RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence. *Nat. Rev. Microbiol.* 11, 745–760. doi: 10.1038/nrmicro3120

Qu, J., Ye, J., and Fang, R. (2007). Artificial microRNA-mediated virus resistance in plants. *J. Virol.* 81, 6690–6699. doi: 10.1128/JVI.02457-06

Ramesh, S. V., Ratnaparkhe, M. B., Kumawat, G., Gupta, G. K., and Husain, S. M. (2014). Plant miRNAome and antiviral resistance: a retrospective view and prospective challenges. *Virus Genes* 48, 1–14. doi: 10.1007/s11262-014-1038-z

Revers, F., and Nicaise, V. (2014). "Plant resistance to infection by viruses," in *Encyclopedia of Life Sciences*, ed. Wiley-Blackwell (Chichester: John Wiley & Sons, Ltd), doi: 10.1002/9780470015902.a0000757.pub3

Roberts, A. P., Lewis, A. P., and Jopling, C. L. (2011). The role of microRNAs in viral infection. *Prog. Mol. Biol. Transl. Sci.* 102, 101–139. doi: 10.1016/B978-0-12-415795-8.00002-7

Sansregret, R., Dufour, V., Langlois, M., Daayf, F., Dunoyer, P., Voinnet, O., et al. (2013). Extreme resistance as a host counter-counter defense against viral suppression of RNA silencing. *PLOS Pathog.* 9:e1003435. doi: 10.1371/journal.ppat.1003435

Schwab, R., Ossowski, S., Riester, M., Warthmann, N., and Weigel, D. (2006). Highly specific gene silencing by artificial microRNAs in Arabidopsis. *Plant Cell* 18, 1121–1133. doi: 10.1105/tpc.105.039834

Shams-Bakhsh, M., Canto, M., and Palukaitis, P. (2007). Enhanced resistance and neutralization of defense responses by suppressors of RNA silencing. *Virus Res.* 130, 103–109. doi: 10.1016/j.virusres.2007.05.023

Shimura, H., Pantaleo, V., Ishihara, T., Myojo, N., Inaba, J., Sueda, K., et al. (2011). A viral satellite RNA induces yellow symptoms on tobacco by targeting a gene involved in chlorophyll biosynthesis using the RNA silencing machinery. *PLOS Pathog.* 7:e1002021. doi: 10.1371/journal.ppat.1002021

Shivaprasad, P. V., Chen, H. M., Patel, K., Bond, D. M., Santos, B. A., and Baulcombe, D. C. (2012). A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell* 24, 859–874. doi: 10.1105/tpc.111.095380

Silhavy, D., and Burgyan, J. (2004). Effects and side-effects of viral RNA silencing suppressors on short RNAs. *Trends Plant Sci.* 9, 76–83. doi: 10.1016/j.tplants.2003.12.010

Simon-Mateo, C., and Garcia, J. A. (2006). MicroRNA-guided processing impairs Plum pox virus replication, but the virus readily evolves to escape this silencing mechanism. *J. Virol.* 80, 2429–2436. doi: 10.1128/JVI.80.5.2429-2436.2006

Smith, N. A., Eamens, A. L., and Wang, M. B. (2011). Viral small interfering RNAs target host genes to mediate disease symptoms in plants. *PLOS Pathog.* 7:e1002022. doi: 10.1371/journal.ppat.1002022

Song, Y. Z., Han, Q. J., Jiang, F., Sun, R. Z., Fan, Z. H., Zhu, C. X., et al. (2014). Effects of the sequence characteristics of miRNAs on multi-viral resistance mediated by single amiRNAs in transgenic tobacco. *Plant Physiol. Biochem.* 77, 90–98. doi: 10.1016/j.plaphy.2014.01.008

Sullivan, C. S., and Ganem, D. (2005). MicroRNAs and viral infection. *Mol. Cell* 20, 3–7. doi: 10.1016/j.molcel.2005.09.012

Sun, L., Lin, C., Du, J., Song, Y., Jiang, M., Liu, H., et al. (2016). Dimeric artificial microRNAs mediate high resistance to RSV and RBSDV in transgenic rice plants. *Plant Cell Tiss. Organ. Cult.* 126, 127–139. doi: 10.1007/s11240-016-0983-8

Tiwari, M., Sharma, D., and Trivedi, P. K. (2014). Artificial microRNA mediated gene silencing in plants: progress and perspectives. *Plant Mol. Biol.* 86, 1–18. doi: 10.1007/s11103-014-0224-7

Várallyay, E., and Havelda, Z. (2013). Unrelated viral suppressors of RNA silencing mediate the control of ARGONAUTE1 level. *Mol. Plant Pathol.* 14, 567–575. doi: 10.1111/mpp.12029

Várallyay, E., Válóczi, A., Agyi, A., Burgyán, J., and Havelda, Z. (2010). Plant virus-mediated induction of miR168 is associated with repression of ARGONAUTE1 accumulation. *EMBO J.* 29, 3507–3519. doi: 10.1038/emboj.2010.215

Villarreal, L. P. (2011). Viral ancestors of antiviral systems. *Viruses* 3, 1933–1958. doi: 10.3390/v3101933

Villarreal, L. P., and Witzany, G. (2010). Viruses are essential agents within the roots and stem of the tree of life. *J. Theor. Biol.* 262, 698–710. doi: 10.1016/j.jtbi.2009.10.014

Viswanathan, C., Anburaj, J., and Prabu, G. (2014). Identification and validation of sugarcane streak mosaic virus-encoded microRNAs and their targets in sugarcane. *Plant Cell Rep.* 33, 265–276. doi: 10.1007/s00299-013-1527-x

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669–687. doi: 10.1016/j.cell.2009.01.046

Vu, T. V., Choudhury, N. R., and Mukherjee, S. K. (2013). Transgenic tomato plants expressing artificial microRNAs for silencing the pre-coat and coat proteins of a begomovirus, Tomato leaf curl New Delhi virus, show tolerance to virus infection. *Virus Res.* 172, 35–45. doi: 10.1016/j.virusres.2012.12.008

Wagaba, H., Patil, B. L., Mukasa, S., Alicai, T., Fauquet, C. M., and Taylor, N. J. (2016). Artificial microRNA-derived resistance to Cassava brown streak disease. *J. Virol. Methods* 231, 38–43. doi: 10.1016/j.jviromet.2016.02.004

Wang, M. B., Masuta, C., Smith, N. A., and Shimura, H. (2012). RNA silencing and plant viral diseases. *Mol. Plant Microbe Interact.* 25, 1275–1285. doi: 10.1094/MPMI

Wang, M. B., and Smith, N. A. (2016). Satellite RNA pathogens of plants: impacts and origins-an RNA silencing perspective. *Wiley Interdiscip. Rev. RNA* 7, 5–16. doi: 10.1002/wrna.1311

Whitham, S., Dinesh-Kumar, S. P., Choi, D., Hehl, R., Corr, C., and Baker, B. (1994). The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. *Cell* 78, 1101–1115. doi: 10.1016/0092-8674(94)90283-6

Xu, W., Meng, Y., and Wise, R. P. (2014). Mla- and Rom1-mediated control of microRNA398 and chloroplast copper/zinc superoxide dismutase regulates cell death in response to the barley powdery mildew fungus. *New Phytol.* 201, 1396–1412. doi: 10.1111/nph.12598

Xuan, N., Zhao, C., Peng, Z., Chen, G., Bian, F., Lian, M., et al. (2015). Development of transgenic maize with anti-rough dwarf virus artificial miRNA vector and their disease resistance. *Chin. J. Biotechnol.* 31, 1375–1386.

Yang, L., Mu, X., Liu, C., Cai, J., Shi, K., Zhu, W., et al. (2015). Overexpression of potato miR482e enhanced plant sensitivity to *Verticillium dahliae* infection. *J. Integr. Plant Biol.* 57, 1078–1088. doi: 10.1111/jipb.12348

Zhai, J., Jeong, D. H., De Paoli, E., Park, S., Rosen, B. D., Li, Y., et al. (2011). MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev.* 25, 2540–2553. doi: 10.1101/gad.177527.111

Zhang, J. Z., Ai, X. Y., Guo, W. W., Peng, S. A., Deng, X. X., and Hu, C. G. (2012). Identification of miRNAs and their target genes using deep sequencing and degradome analysis in trifoliate orange [*Poncirus trifoliata* (L.) Raf]. *Mol. Biotechnol.* 51, 44–57. doi: 10.1007/s12033-011-9439-x

Zhang, S., Yue, Y., Sheng, L., Wu, Y., Fan, G., Li, A., et al. (2013). PASmiR: a literature-curated database for miRNA molecular regulation in plant response to abiotic stress. *BMC Plant Biol.* 13:33. doi: 10.1186/1471-2229-13-33

Zhang, T., Zhao, Y. L., Zhao, J. H., Wang, S., Jin, Y., Chen, Z. Q., et al. (2016). Cotton plants export microRNAs to inhibit virulence gene expression in a fungal pathogen. *Nat. Plants* 2:16153. doi: 10.1038/nplants.2016.153

Zhang, X., Li, H., Zhang, J., Zhang, C., Gong, P., Ziaf, K., et al. (2011). Expression of artificial microRNAs in tomato confers efficient and stable virus resistance in a cell-autonomous manner. *Transgenic Res.* 20, 569–581. doi: 10.1007/s11248-010-9440-3

Zhang, Y., Xia, R., Kuang, H., and Meyers, B. C. (2016). The diversification of plant NBS-LRR defense genes directs the evolution of MicroRNAs that target them. *Mol. Biol. Evol.* 33, 2692–2705. doi: 10.1093/molbev/msw154

Zhao, M., Meyers, B. C., Cai, C., Xu, W., and Ma, J. (2015). Evolutionary patterns and coevolutionary consequences of MIRNA genes and microRNA targets triggered by multiple mechanisms of genomic duplications in soybean. *Plant Cell* 27, 546–562. doi: 10.1105/tpc.15.00048

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G. J., Wang, X. J., et al. (2007). A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev.* 21, 1190–1203. doi: 10.1101/gad.1543507

Zhu, Q., Fan, L., Liu, Y., Xu, H., Llewellyn, D., and Wilson, I. (2013). miR482 regulation of NBS-LRR defense genes during fungal pathogen infection in cotton. *PLOS ONE* 8:e84390. doi: 10.1371/journal.pone.0084390

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 October 2017
doi: 10.3389/fmicb.2017.01998

Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches

Anja Pecman[1,2], Denis Kutnjak[1]*, Ion Gutiérrez-Aguirre[1], Ian Adams[3], Adrian Fox[3], Neil Boonham[3,4] and Maja Ravnikar[1]*

[1] Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia, [2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, [3] Fera Science Ltd., York, United Kingdom, [4] Institute for Agri-Food Research and Innovation, Newcastle University, Newcastle upon Tyne, United Kingdom

Next generation sequencing (NGS) technologies are becoming routinely employed in different fields of virus research. Different sequencing platforms and sample preparation approaches, in the laboratories worldwide, contributed to a revolution in detection and discovery of plant viruses and viroids. In this work, we are presenting the comparison of two RNA sequence inputs (small RNAs vs. ribosomal RNA depleted total RNA) for the detection of plant viruses by Illumina sequencing. This comparison includes several viruses, which differ in genome organization and viroids from both known families. The results demonstrate the ability for detection and identification of a wide array of known plant viruses/viroids in the tested samples by both approaches. In general, yield of viral sequences was dependent on viral genome organization and the amount of viral reads in the data. A putative novel *Cytorhabdovirus,* discovered in this study, was only detected by analysing the data generated from ribosomal RNA depleted total RNA and not from the small RNA dataset, due to the low number of short reads in the latter. On the other hand, for the viruses/viroids under study, the results showed higher yields of viral sequences in small RNA pool for viroids and viruses with no RNA replicative intermediates (single stranded DNA viruses).

Keywords: next generation sequencing, small RNA, ribosomal RNA depleted total RNA, detection, plant viruses, plant viroids

OPEN ACCESS

Edited by:
Guenther Witzany,
Telos - Philosophische Praxis, Austria

Reviewed by:
Claudio Ratti,
Università di Bologna, Italy
Carmen Hernandez,
Instituto de Biología Molecular y Celular de Plantas (CSIC), Spain

****Correspondence:***
Anja Pecman
anja.pecman@nib.si
Denis Kutnjak
denis.kutnjak@nib.si

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *17 July 2017*
Accepted: *28 September 2017*
Published: *13 October 2017*

Citation:
Pecman A, Kutnjak D, Gutiérrez-Aguirre I, Adams I, Fox A, Boonham N and Ravnikar M (2017) Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches. Front. Microbiol. 8:1998. doi: 10.3389/fmicb.2017.01998

INTRODUCTION

Plant viruses and viroids are important plant pathogens, causing economic losses by reducing crop quality and quantity all over the world (Loebenstein, 2008; Soliman et al., 2012). Thus, their reliable detection is of a crucial importance for plant protection. Classical methods in plant virus diagnostics can be roughly divided into specific (serological/molecular tests) and non-specific (indicator test plants, electron microscopy) approaches. Specific methods are usually targeted to one or a few viral species and require *a priori* knowledge of the pathogens being tested, whilst non-specific approaches do not require specific knowledge of the pathogens, however, frequently only classify viruses at a genus level based on the shared physical/biological characters. Discovery of new viruses/viroids and new hosts has increased rapidly after the introduction of next generation

sequencing (NGS). NGS technologies allow a generic approach (non-specific method) to virus identification that does not require any prior knowledge on the targeted pathogens but can deliver a species/strain specific result (Adams and Fox, 2016). It was first employed for plant virus detection in 2009 (Adams et al., 2009; Al Rwahnih et al., 2009; Kreuze et al., 2009). Since 2009, different sample preparation methods have been developed, relying on different nucleic acid inputs, most commonly: total RNA (totRNA); ribosomal RNA depleted total RNA (rRNA depleted totRNA); double stranded RNA (dsRNA); virus derived small interfering RNA (sRNA); RNA from purified or partially purified viral particles; polyadenylated RNA (poly(A) RNA); and RNA after subtractive hybridization with healthy plant RNA. Applications of different sample preparation methods are reviewed in Roossinck et al. (2015); Wu et al. (2015), and Adams and Fox (2016). Viruses have diverse genome organizations and use different replication strategies. Based on these two characteristics they can be classified into 7 groups (the Baltimore classification): double stranded DNA (Group I, dsDNA +/−), single stranded DNA (Group II, ssDNA +), double stranded RNA (Group III, dsRNA +/−), positive sense single stranded RNA (Group IV, ssRNA +), negative sense single stranded RNA (Group V, ssRNA −) viruses, positive sense single stranded RNA viruses that replicate through a DNA intermediate (Group VI, ssRNA-RT +), and double stranded DNA viruses that replicate through a RNA intermediate (Group VII, dsDNA-RT +/−) (Baltimore, 1971). Viroids are classified into two families: members of *Avsunviroidae* family replicate in chloroplast, whereas members of *Pospiviroidae* family replicate in nucleus (Flores et al., 2014). Considering the diversity of viruses and viroids, with different genome organizations in mind, it is conceivable that using different nucleic acid inputs for NGS could affect their overall detection.

Sample preparation methods (i.e., different nucleic acid inputs), used before NGS, can differ in their efficiency and can have specific advantages and disadvantages. For example, subtractive hybridization of the host plant nucleic acids, using tomato (*Solanum lycopersicum*) and *Pepino mosaic virus* (PepMV, RNA +, *Potexvirus, Alphaflexiviridae*) as a model system, resulted in three times more PepMV sequences in subtracted sample (Adams et al., 2009), but as it is a time consuming procedure, which requires a healthy plant of the same species as the sample to be tested (Adams and Fox, 2016), subtractive hybridization is not well suited in a high-throughput diagnostic settings. Some sample preparation methods may cause bias in the detection of a particular group of viruses. Sequencing of dsRNA was mainly used for detection of RNA + and RNA +/− viruses, since RNA—and DNA viruses could be missed (Roossinck et al., 2015) using this approach; nevertheless, a new geminivirus (DNA +) was identified using dsRNA sequencing (Al Rwahnih et al., 2013). RNA isolated from purified viral particles has been successfully used for sequencing different viruses (reviewed in Roossinck et al., 2015; Wu et al., 2015). A comparison between deep sequencing of sRNAs and RNA isolated from viral particles showed higher efficiency of the latter for the reconstruction of complete consensus *Potato virus Y* (RNA +, *Potyvirus, Potyviridae*) genomes (Kutnjak et al., 2015). However, virus purification is not applicable for un-encapsidated viruses and requires sample specific processing since it is unlikely that all viruses could be captured by a single protocol for viral particles purification (Roossinck et al., 2015; Wu et al., 2015). Poly(A) RNA based enrichment strategy has been also used for both RNA and DNA viruses but it is not applicable for the detection of viruses without a poly(A) tail (Wu et al., 2015). Data from sequencing poly(A) RNA showed a lower degree of virus genome coverage in comparison to saturated genome coverage reached with sRNA data for *Grapevine leafroll-associated virus 3* (RNA +, *Ampelovirus, Closteroviridae*), yet a comparison between poly(a)RNA and sRNA data for *Hop stunt viroid*, (*Pospiviroidae* family) showed comparable outcomes (high genome coverage) for both approaches (Visser et al., 2016).

In this study, we focused the comparison (with the detection and identification of plant viruses and viroids in mind) on the two types of RNA inputs: sequencing of sRNA and sequencing of rRNA depleted totRNA. Those two approaches seem to be the most generically applicable to viruses with different genome types and replication strategies and could be relatively easily integrated in workflows of diagnostic labs.

Sequencing and assembly of viral sRNA (Kreuze et al., 2009) has been successfully used for detection and identification of several plant viruses and viroids and their complete genome assembly (reviewed in Boonham et al., 2014; Kreuze, 2014). It has been speculated that this approach could be problematic if used to detect viruses that either do not trigger silencing responses or that express silencing suppressors (Roossinck et al., 2015). Also, de novo assembly of longer viral contigs could be complicated due to short reads lengths (Boonham et al., 2014; Roossinck et al., 2015; Adams and Fox, 2016). On the other hand, the approach is very generic, using the same protocol of sample preparation for many different plant species and doesn't require high quality of RNA input (Kutnjak et al., 2017).

Sequencing of plant viruses using total RNA as an input was first described by Adams et al. (2009) and Al Rwahnih et al. (2009), followed by several successful studies (reviewed in Boonham et al., 2014). It is also a very generic approach, however, a potential shortcoming of that method can be the low viral RNA titer within the background plant RNA. To overcome this, removal of the highly abundant plant ribosomal RNA from the total RNA pool (rRNA depleted tot RNA) has been explored, which can results in a 10-fold enrichment of viral RNA (Adams and Fox, 2016).

Recent comparison (Visser et al., 2016) of sRNA and rRNA depleted totRNA for *Citrus tristeza virus* (RNA +, *Closterovirus, Closteroviridae*) and *Citrus dwarfing viroid* (*Pospiviroidae* family) implied a preferential use of rRNA depleted totRNA for *de novo* assembly of viral genome sequences from NGS data. No wider comparison of these two approaches (including viruses with different genome characteristics) has been reported. With this in mind, our aim was to compare the two approaches, including plant viruses with different genome structures and replication strategies (belonging to different Baltimore classification groups) and viroids from different families into comparison. The aims were to compare the two approaches in terms of: (1) known virus detection and identification (2) recovery of virus/viroid reads

and (3) effectiveness of detection of new/unknown viruses by reconstruction of longer viral contigs by *de novo* assembly and read mapping analysis approaches.

MATERIALS AND METHODS

Description of Samples

Nine virus-infected plant samples were included in this study. The selection included samples of different plant species, infected with a range of plant viruses in single or mixed infections with at least one representative from each group of the Baltimore viral classification containing plant viruses, and viroids from both families (**Table 1**).

Sample Preparation and Sequencing

Total RNA was isolated from plant samples using TRIzol reagent (Life technologies, USA) following the manufacturer's instructions. Isolated total RNA was then divided in half for comparative purposes. One half was sent to Seqmatic LLC (USA) for sRNA library preparation (TailorMix miRNA Sample Preparation Kit V2, SeqMatic LLC, USA) and sequencing. The samples were multiplexed in one lane of a HiSeq 2500 (Illumina, USA) in 1 × 50 bp mode. The remaining total RNA was further purified using an RNeasy protocol including DNase treatment following the manufacturer's protocols (RNA Cleanup protocol; RNeasy Mini Kit; Qiagen, Netherlands). Ribosomal RNA was depleted from the purified total RNA and sequencing libraries were prepared using the ScriptSeq™ Complete Kit (plant leaf) (Illumina, USA). The libraries were sequenced using MiSeq (Illumina, USA) in 2 × 300 bp (V3) mode. Number and average length of sequencing reads for every sample sequenced by both approaches are in Supplementary Table 7.

Detection of Viruses in NGS Data

Reads obtained by both sequencing procedures were trimmed, filtered and further analyzed to confirm the presence of viruses and viroids. Bioinformatics pipelines used for virus detection from NGS data are detaily described in Supplementary Data 1.1. In both cases, the presence of suspected viral sequences was confirmed by mapping the reads to the complete viral genome sequences of the most similar viral isolates from the NCBI GenBank database, followed by visual inspection of individual mappings.

Confirmatory Testing

The presence of virus in each case was also confirmed by using ELISA, RT-PCR, and RT-qPCR methods (**Table 1**). ELISA was performed using polystyrene microtiter plates (nunc-Immuno™, Sigma-Aldrich Inc., USA) and kits containing virus specific reagents as follows, AMV: Cat No. 07001S (Loewe Biochemica GmbH, Germany), CaMV: Cat No. 07086 (Loewe Biochemica GmbH, Germany), PVY: Cat No. 1105 (Bioreba AG, Switzerland) and TYLCV: Cat. No. 1072 (Neogen Europe Ltd., UK). The assays were performed following the manufacturer's instructions. In each case a negative control corresponding to the same species as the test sample was used. The result was considered positive when the optical density (OD) A_{405} value after 2 h for a given sample was greater than 2× the mean OD value of the corresponding negative control. For reverse transcription quantitative PCR (RT-qPCR) and reverse transcription conventional PCR (RT-PCR), total RNA was extracted from fresh or lyophilized plant material using the RNeasy Plant Mini Kit (Qiagen), following the manufacturer instructions. RT-qPCR was performed using published methods for PepMV (Gutiérrez-Aguirre et al., 2009) and for ToMV (Boben et al., 2007). Conventional RT-PCR was performed for PNYDV (Gaafar and Ziebell, 2016), STV (Sabanadzovic et al., 2009), ToCV (Dovas et al., 2002), TMV (Kumar et al., 2011), PLMVd (Loreti et al., 1999) TASVd and CLVd (Verhoeven et al., 2004). PCR primers designed specifically to confirm the presence of novel CCyV1 were as follows: CCyV1-fw (5′-GTCTCTCTTGCGTTGAGCCA-3′) and CCyV1-rev (5′-GGTTGCGGATAGCTCTTCCT-3′). All the amplicons obtained by RT-PCR were purified and sent for Sanger sequencing (GATC Biotech AG, Germany). The Sanger sequences were aligned against the genomes of detected viral species and their identity was confirmed in all of the cases.

Construction of Consensus Viral/Viroid Genome Sequences

For every identified virus/viroid the consensus viral genomes were extracted from the sRNA read mappings (see section Detection of Viruses in NGS Data) to obtain a corrected consensus genome. Validation of each corrected consensus genome was performed by mapping the *de-novo* generated contigs obtained by both NGS approaches to corresponding corrected consensus genome. Both mapping results were visually inspected for possible differences between the *de-novo* contigs and corrected consensus genome sequence. Observed conflicts were further investigated by inspecting the read mapping results. Finally, few of the observed differences were explained as polymorphisms in viral populations. In sample III, two divergent strains (80% nucleotide identity) of PepMV were detected (PepMV-EU and PepMV-CH2). In this case, the complete genome sequences of the two most similar isolates from NCBI GenBank were used in subsequent comparisons (KF718832.1 and JX866666.1), without the corrections after reads and contigs mapping as described previously.

Comparison of sRNA and rRNA Depleted totRNA Inputs

For comparisons, all raw reads were trimmed and filtered in CLC Genomic Workbench 9 (Qiagen). For rRNA depleted totRNA datasets, reads shorter than 100 nucleotides were discarded. Then, reads were trimmed using quality scores, setting the limit to 0.05 (see CLC Genomics Workbench User Manual, Chapter 23, for explanation). For sRNA reads, first, adaptor trimming was performed, then reads shorter than 20 and longer than 24 nucleotides were discarded.

First, the viral fraction of the total nucleotides sequenced (from now on called percentage of virus/viroid nucleotides) in each of the datasets for each of the detected viruses was calculated by mapping the trimmed and filtered reads (of the corresponding dataset) to the consensus viral/viroid genomes generated in the

TABLE 1 | Samples included in the comparison with corresponding results from: NGS (viruses/viroids listed in the table were detected in corresponding samples by NGS) and other diagnostic methods (ELISA, RT-PCR and RT-qPCR).

Sample number	Virus, genus, family	Baltimore classification	Genome organization	Abbreviations	Host	Initial detection with NGS		Results of confirmatory testing	NCBI GenBank accession number	NCBI SRA accession number (sRNA/rRNA depleted totRNA)
						sRNA	rRNA depleted totRNA			
I	*Potato virus Y, Potyvirus, Potyviridae	Group IV (ssRNA +)	Linear	PVY	*Solanum tuberosum*	+	+	+[a]	KY810782	SRR5377154/SRR5377146
II	*Cauliflower mosaic virus, Caulimovirus, Caulimoviridae;	Group VII (dsDNA-RT +/–)	Circular	CaMV	*Brassica oleracea*	+	+	+[a]	KY810770	SRR5377153/SRR5377145
	Novel cabbage cytorhabdovirus 1, *Cytorhabdovirus, Rhabdoviridae*	Group V (ssRNA –)	Linear	Novel CCyV1	*Brassica oleracea*	–	+	+[b]	KY810772	
III	*Tomato Yellow Leaf Curl Virus, Begomovirus, Geminiviridae;	Group II (ssDNA +)	Circular	TYLCV	*Solanum lycopersicum*	+	+	+[a]	KY810789	SRR5377152/SRR5377144
	*Tomato chlorosis virus, Crinivirus, Closteroviridae;	Group IV (ssRNA +)	Linear, segmented	ToCV	*Solanum lycopersicum*	+	+	+[b]	KY810786 KY810787	
	*Pepino mosaic virus, Potexvirus, Alphaflexiviridae;	Group IV (ssRNA +)	Linear	PepMV	*Solanum lycopersicum*	+	+	+[c]		
	*Tomato mosaic virus, Tobamovirus, Virgaviridae;	Group IV (ssRNA +)	Linear	ToMV	*Solanum lycopersicum*	+	+	+[c]	KY810788	
	*Southern tomato virus, Amalgavirus, Amalgaviridae;	Group III (dsRNA +/–)	Linear	STV	*Solanum lycopersicum*	+	+	+[b]	KY810783	
	Columnea latent viroid, Pospiviroid, Pospiviroidae	viroid	Circular	CLVd	*Solanum lycopersicum*	+	+	+[b]	KY810771	
IV	*Alfalfa mosaic virus, Alfamovirus, Bromoviridae	Group IV (ssRNA +)	Linear, segmented	AMV	*Nicotiana tabacum*	+	+	+[a]	KY810767 KY810768 KY810769	SRR5377151/SRR5377143
V	*Pea necrotic yellow dwarf virus, Nanovirus, Nanoviridae	Group II (ssDNA +)	Circular, segmented	PNYDV	*Pisum sativum*	+	+	+[b]	KY810774 KY810775 KY810776 KY810777 KY810778 KY810779 KY810780 KY810781	SRR5377150/SRR5377142
VI	*Tobacco mosaic virus, Tobamovirus, Virgaviridae	Group IV (ssRNA +)	Linear	TMV	*Nicotiana* sp.	+	+	+[b]	KY810785	SRR5377149/SRR5377141
VII	*Peach latent mosaic viroid, Pelamoviroid, Avsunviroidae	viroid	Circular	PLMVd	*Prunus* sp.	+	+	+[b]	KY810773	SRR5377148/SRR5377140
VIII	*Tomato apical stunt viroid, Pospiviroid, Pospiviroidae	viroid	Circular	TASVd	*Solanum lycopersicum*	+	+	+[b]	KY810784	SRR5377147/SRR5377139
IX	*Chrysanthemum stem necrosis virus, Tospovirus, Tospoviridae	Group V (ssRNA –)	Linear, segmented	CSNV	*Nicotiana benthamiana*	+	+	+[c]	MF093683 MF093684 MF093685	SRR5630913/SRR5630912

*Taxonomic classification, Baltimore classification and genome organization of detected viruses are given in separate columns. Host plant information is given in the separate column. NA, not applicable; +, detected; –, not detected; *, viruses/viroids which were known to be present in the sample before NGS analysis.*

[a] *Confirmatory testing has been done using ELISA assay.*

[b] *Confirmatory testing has been done using RT-PCR assay.*

[c] *Confirmatory testing has been done using RT-qPCR assay.*

previous step. Mapping parameters are listed in Supplementary Tables 1, 4.

To further compare the effectiveness of both approaches for detection and discovery of selected viruses, we then performed a normalization by subsampling the data from each sample (for both sRNA and rRNA depleted totRNA) to the same number of nucleotides. Random subsampling was performed to different subsample sizes: 1, 10, 30, and 50 million nucleotides. This was repeated ten times for each sample/size combination, yielding in total 360 datasets (9 samples × 4 subsample sizes × 10 replicates of subsampling). For those, the following analyses were implemented: (1) reads were mapped to the corresponding consensus viral/viroid genomes and the fraction of viral/viroid genome covered by reads (from now on: genome coverage (reads)) and the average depth of sequencing (number of times a nucleotide in a reference is covered by reads averaged for the complete genome) were calculated; (2) *de novo* assembly of reads was performed using CLC Genomics Workbench 9, followed by mapping the resulting contigs to the corresponding consensus viral/viroid genomes and calculation of the fraction of viral/viroid genome covered by the *de-novo* contigs (from now on: genome coverage (contigs)). Results of these comparisons are jointly shown in **Figure 2** and visualized as dots connected with solid lane (representing rRNA depleted totRNA results) and triangles connected with dashed lines (representing sRNA results). The mapping and *de novo* assembly parameters are listed in Supplementary Tables 1–4.

RESULTS

Sample Characterization

Twelve different viruses (among those, one viral species with two divergent strains) and three viroid species were detected using NGS in the nine samples included in the analysis (**Table 1**). Nine were known to be present in the samples before the NGS analysis (marked with * in **Table 1**), whilst six virus/viroid species were detected using NGS during the study and their presence was confirmed as described in section Materials and Methods (**Table 1**). Both methods revealed the presence of 14 viral/viroid species whilst 1 virus (a putative novel viral species from the genus *Cytorhabdovirus*: CCyV1) could only be detected using the rRNA depleted totRNA approach. Seven samples (I, IV-IX) contained single viral/viroid infections, one sample (II) was infected with two viruses. Sample III was infected with five viruses and one viroid. All of the viruses and viroids detected and included in the study are listed in the **Table 1**.

Percentage of Virus/Viroid Reads Differs For Different Viruses

First, we estimated what percentage of the total sequenced nucleotides were viral/viroid nucleotides (of the complete cleaned NGS datasets) for different viral species for each of the two approaches. The percentage of viral/viroid nucleotides was in some cases higher using sRNA input and in other cases higher using rRNA depleted totRNA input (**Figure 1**). Specifically, the results showed that for 6 viruses/viroids the sRNA approach generated a higher fraction of viral/viroid sequences: TASVd, ToCV, CLVd, TYLCV, PNYDV, PLMVd, and PVY (**Figure 1**: the viruses located below the diagonal line). For the sRNA approach, the highest percentage of viral sequences was observed for PVY (50%, **Figure 2A**). The rRNA depleted totRNA approach generated more viral sequences for 6 viruses: a novel *Cytorhabdovirus*, PepMV (two isolates), CaMV, AMV, CSNV and TMV (**Figure 1**, the viruses located above the diagonal line), with the highest viral sequences fractions for TMV (83%), AMV (56%), CSNV (48%), and CaMV (48%) (**Figures 1**, **2A**). In two cases (STV and ToMV), the percentage of virus sequences were extremely low regardless of the RNA inputs (**Figures 1**, **2A**).

Comparison on Normalized Subsamples

To be able to compare the two approaches in a greater detail, we subsampled all of the datasets to the same number of nucleotides. Ten replicates of four different sizes of subsamples (1, 10, 30, and 50 million nucleotides) were generated for each dataset to enable an assessment of the impact of data rarefaction and data variability on the performance of tested parameters.

First, average depth was evaluated (**Figure 2B**). In all cases, average depth increased with the increase of subsample sizes and followed the patterns observed when comparing the fractions of viral sequences nucleotides recovered by the two approaches. Results from 10 independent replicates for each subsample size showed a low variability for PVY, ToCV, PepMV, AMV, TMV, CSNV, and CaMV. Variability between the subsamples in average depth was higher for all other viruses/viroids (Supplementary Table 5).

Secondly, we investigated how effectively the reads cover the genomes of different viruses by calculating the fraction of the genome covered by reads [genome coverage (reads)] (**Figure 2C**). Results of the analysis showed low variability between replicates of subsamples, except when mapping rRNA depleted totRNA reads to ToMV, STV, TYLCV, TASVd, and PLMVd where variation was very high (Supplementary Table 5, **Figure 2C**). In all cases, as expected, better genome coverage was achieved with the increasing subsample sizes. For the sRNA approach, complete genomes (100%) were covered for majority of the viruses/viroids at subsample size of 30 million nucleotides. The exceptions were ToMV, STV and the putative novel *Cytorhabdovirus*. For those, even at 50 million nucleotides, genome coverage was 70% or less.

For the rRNA depleted totRNA approach, for half of the viruses (PVY, PepMV, AMV, TMV, novel CCyV1, CSNV, CaMV, CLVd, and TASVd) complete genomes were covered at 10 million nucleotides. However, for some viruses/viroids (ToCV, TYLCV, PNYDV, and PLMVd) relatively low genome coverage was achieved at smaller subsample sizes (1 and 10 million nts) and even at the largest subsample size (50 million nts) the coverage did not reach 100% (**Figure 2C**). The genomes of ToMV and STV, for which very low numbers of reads were recovered (**Figures 1**, **2A**), were poorly covered even at high subsampling depths, for example, even with 50 million nucleotides, coverage remained below 50% (**Figure 2C**).

Reads from normalized datasets were *de novo* assembled into contigs, which were then mapped to the corresponding

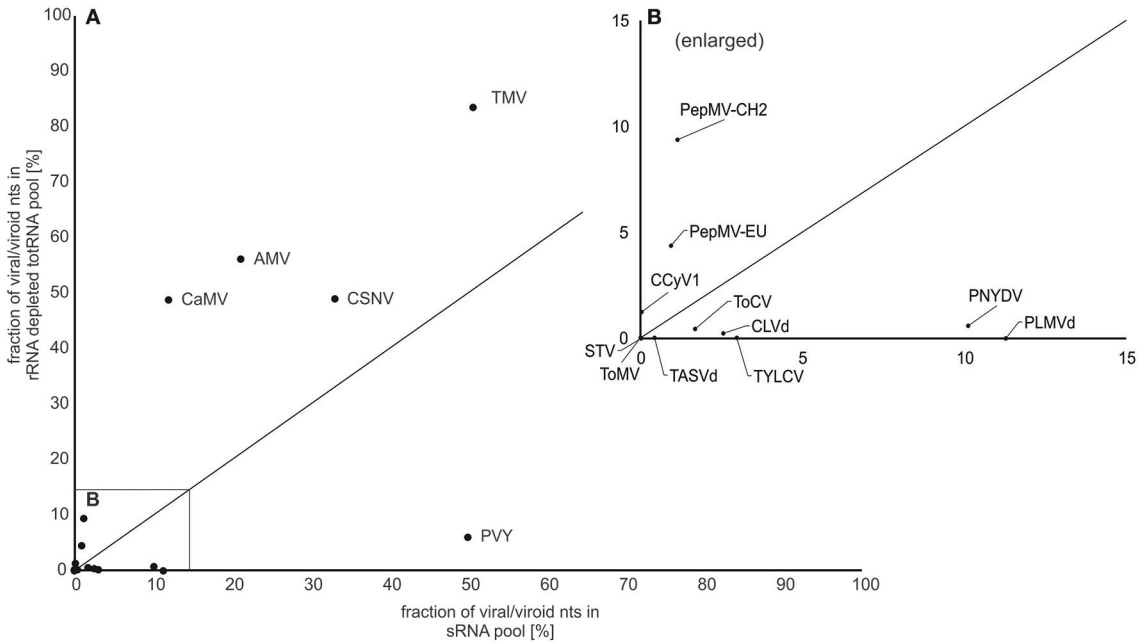


FIGURE 1 | Fraction of virus/viroid nucleotides (nt) in NGS datasets for small RNA and rRNA depleted totRNA approaches. **(A)** The plots show the percentage of nucleotides (fraction of total) that mapped to the corresponding consensus viral/viroid genome for both sRNA (x-axis) and rRNA depleted totRNA (y-axis) inputs. Dots represent the value for each virus/viroid (also listed in **Figure 2A**), with viral/viroid species abbreviations (see **Table 1**) next to the dots. The diagonal solid line represents theoretical scenario in which percentage of virus/viroid nucleotides would be equivalent for both approaches; dots above the line represent cases for which fraction of virus/viroid nucleotides was higher using rRNA depleted totRNA input, dots below the line represent cases for which fraction of viruses/viroids nucleotides was higher using sRNA input. **(B)** Enlarged part of **(A)** (0–15%).

consensus viral genomes in order to calculate the fraction of the viral genomes covered by the *de novo* assembled contigs [genome coverage (contigs)] (**Figure 2D**). The analysis of subsample replicates showed in general lower variability for sRNA datasets than rRNA depleted totRNA datasets (Supplementary Table 5). For the majority of the viruses, the coverage by contigs increased with subsample size, however, conversely, in several cases, it dropped at larger subsample sizes, i.e., TMV and PLMVd for sRNA and PepMV, CSNV, CaMV and CLVd for rRNA depleted totRNA approach (**Figure 2D**). Contigs, assembled *de novo* from rRNA depleted totRNA datasets covered higher fractions of viral genomes for almost all viruses at all subsample sizes (coverage reached 95% at 10 million nts for majority of viruses), in comparison to sRNA derived contigs (95% coverage at 10 million nts was achieved only for PVY, TMV, and CLVd). Two exceptions to this observation were TYLCV and CLVd, for which sRNA derived *de novo* contigs cover higher genome fraction than rRNA depleted totRNA contigs, for all subsample sizes.

The comparison of the *de novo* assemblies for STV and ToMV revealed that when very low numbers of viral reads are recovered, the rRNA depleted totRNA approach is more effective, since in the case of the sRNA approach, no corresponding viral contigs were generated (**Figure 2D**). A similar scenario was observed also for the putative novel *Cytrohabovirus*, where very low recovery of viral reads in the sRNA dataset resulted in no assembled contigs corresponding to this virus (**Figure 2D**).

DISCUSSION

In this study we compared the effectiveness of two NGS approaches that have been widely adopted for plant virus detection: sRNA deep sequencing and deep sequencing of rRNA depleted totRNA. When comparing the amount of virus/viroid reads recovered by one or the other approach, we observed different results for different viruses/viroids: in some cases, more viral/viroid nucleotides were recovered using sRNA and in other by rRNA depleted totRNA sequencing.

Detailed inspection of the results of the read mapping suggested higher recovery of virus reads for ssDNA viruses and viroids when using sRNA approach than when using rRNA depleted totRNA approach. For viroids, this could be the consequence of induced RNA silencing (Itaya et al., 2001; Papaefthimiou et al., 2001; Martínez de Alba et al., 2002) and, at the same time, the absence of the messenger RNA production, because, in the case of viroids, "long" RNAs are generated solely for the purpose of replication. Similarly, in the case of viruses with a circular ssDNA genome organization, a smaller fraction of viral nucleotides was recovered using rRNA depleted totRNA. In contrast with viruses with RNA genomes, for ssDNA viruses, RNA molecules are generated only during the transcription step, as messenger RNAs, which could be the reason for the lower recovery of viral nucleotides in this pool. Moreover, small RNAs could be amplified by the action of RNA-dependent RNA polymerase 6 (Borges and Martienssen, 2015)

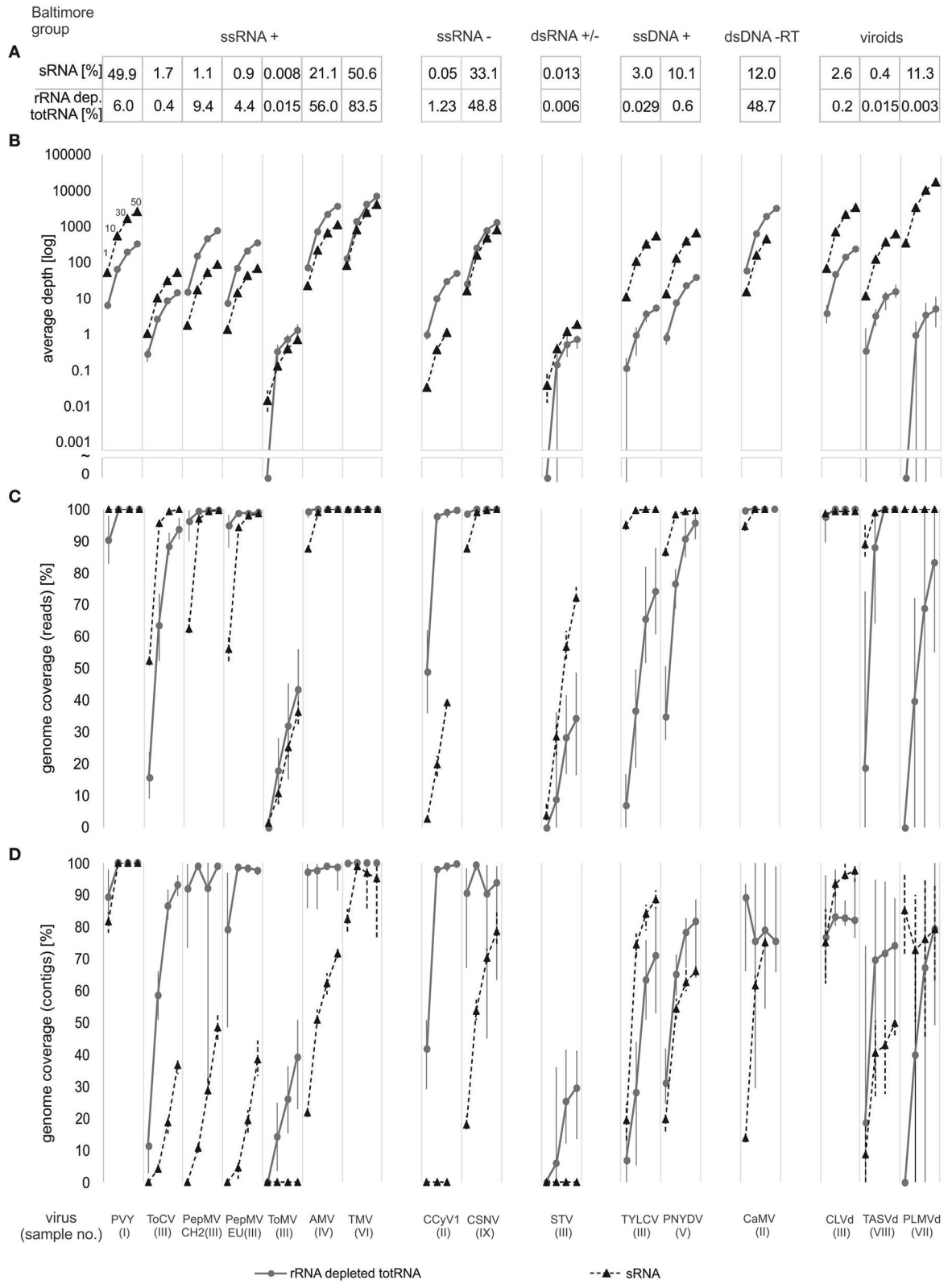


FIGURE 2 | Comparison of sRNA and rRNA depleted totRNA approaches using data size-normalized subsamples. Results for each virus included in the analysis are shown along the x-axis and are grouped according to Baltimore classification **(A)** Fraction (%) of virus nucleotides in trimmed and filtered complete NGS datasets. **(B)** Average depth (number of reads covering a position in a viral genome, averaged over the complete genome sequence) at different subsample sizes. Symbol ~

(Continued)

FIGURE 2 | Continued
indicate interruption of log scale, below, 0 values are plotted. **(C)** Fraction of viral genome (in %) covered by reads [genome coverage (reads)] at different subsample sizes. **(D)** Fraction of viral genome (in %) covered by contigs [genome coverage (contigs)] at different subsample sizes. For **(B–D)** Dots/triangles represent the mean, whereas vertical bars connect minimum and maximum results of 10 repeated analyses. Four different subsample sizes were used (1, 10, 30, and 50 million nts) and are designated in the first column, other columns follow the same logic. Triangles and dashed lines represent results for sRNA approach, dots and solid lines represent results for rRNA depleted totRNA. In some cases data points are missing, since the size of the complete dataset was smaller than the largest subsample.

during the production of secondary sRNAs. The exception among the DNA viruses in this study was CaMV (DNA-RT), for which a higher fraction of virus nucleotides was recovered by sequencing rRNA depleted totRNA. The CaMV dsDNA genome is replicated through an RNA intermediary, in addition to producing messenger RNAs through transcription (Hull, 2014), which could explain a larger proportion of viral nucleotides in this pool.

All linear viruses in our infected plant samples had a ssRNA genome organization and synthesize different types of RNA throughout their replication cycle. For most of these viruses, sequencing rRNA depleted totRNA resulted in a larger proportion of reads mapping to the viral genomes (**Figure 1**) compared with sRNA. However, a few exceptions were observed, PVY being the most notable with many more viral reads being present in the sRNA dataset. The high abundance of virus derived sRNA has already been reported for PVY (Kutnjak et al., 2015) and other potyviruses (Kreuze et al., 2009) even though they encode strong RNA silencing suppressors (Yelina et al., 2002; Ivanov et al., 2016).

In general, when read mapping was performed, 10 million nucleotides was sufficient to cover complete viral genomes using any of the two approaches (**Figure 2C**). However, in some cases (STV and ToMV in sample III) very low numbers of viral reads were recovered (by both approaches), which negatively affected all the evaluated parameters. For those two cases, the percentage of virus reads (for both approaches) was lower than 0.1%, and the average read depth remained lower than 10×, and none of the viral genomes were completely covered by the reads even at the highest subsample size (50 million) (**Figure 2C**).

When comparing *de novo* assembly of sequencing reads, the rRNA depleted totRNA approach was generally more efficient than sRNA approach; this was demonstrated in higher proportion of viral genomes covered by *de novo* generated contigs from rRNA depleted totRNA datasets. The contigs assembled from rRNA depleted totRNA data covered at least a fraction of the consensus genome even in cases where the percentage of virus/viroid reads was lower than 0.1% and average depth lower than 10 (i.e., ToMV and STV) (**Figure 2D**). In those cases, no viral contigs were assembled using sRNA datasets, probably due to a combination of low amount and small sizes of viral reads. Poorer coverage of viral genomes by sRNA derived *de novo* contigs is likely related to the more difficult assembly of very short sRNA reads into longer contigs, which has been observed previously (Kutnjak et al., 2015; Visser et al., 2016).

In some cases (PepMV, TMV, CSNV CaMV, CLVd, and PLMVd) smaller genome fractions are covered by contigs, when larger data sets are used for the assembly (corresponding to average depths > 100). This has been observed previously and is an artifact of the assembly algorithms (see CLC Analyses-related questions, 2017), which are not optimized for very high sequencing depths. After mapping reads or contigs to evaluate average depth and genome coverage (reads/contigs) we observed also the trend in generating higher or lower variability within 10 repeats. Unrepeatable random subsampling occurred when analysing smaller datasets and/or lower viral/viroid nucleotide proportion within the datasets, since all samples with this two features had greater variability.

The study has highlighted some points of difference between the compared approaches that may help to inform the choice of approach based on the purpose of the sequencing. This could be (i) screening against a list of known target organisms (e.g., at the import/export) and (ii) identification of the (possibly yet unknown) causal agent of the disease. Considering (i) screening against a list of known targets, this would be most cost effectively achieved using a method that maximizes the amount of viral sequences compared with host sequences. This study showed (**Figures 1**, **2A**) that the performance of the two compared approaches is very virus dependent. Broadly, sRNA performed better for circular ssDNA viruses and viroids, whilst rRNA depleted total RNA performed better for most of the tested linear RNA viruses with a notable exception (PVY). If considering (ii) sequencing for novel virus discovery, long contigs would provide the greatest chance of detecting very dissimilar sequences by comparing predicted amino-acid sequence from virus ORFs (e.g., with the use of BLASTx analysis or hidden Markov model based protein domain searches). The data shows that rRNA depleted total RNA generated longer contigs (which covered greater fractions of viral genomes) for most of the investigated viruses (**Figure 2D**). As the most prominent example, an important difference between the compared approaches was observed on a case of a previously un-described *Cytorhabdovirus*, which was identified from the rRNA depleted total RNA following *de novo* assembly and BLASTx analysis, whilst the virus reads could only be found in the sRNA sequence data *post-hoc* (de novo assembly of sRNA reads did not generate any matching contigs).

The results of the comparison between the two NGS approaches highlight some trends that may guide diagnostic laboratories in the selection of a method appropriate for a specific application. However, whichever method is selected it is important to be aware of the limitations, some of which are detailed in this study, and follow up putative identification using an appropriate method. The recently published framework for handling novel plant viruses detected using NGS provides guidelines for achieving this (Massart et al., 2017).

In order to examine the potential costs of each method on commonly used Illumina sequencing platforms (HiSeq/sRNA and MiSeq/rRNA depleted totRNA) staff time used and reagent

costs (in GBP) were calculated using list prices (Illumina) obtained on 1st March 2017. In general, both approaches generate more than sufficient amount of data than required to identify all of the viruses if mapping is used (50 million nts; **Figure 2**). HiSeq/sRNA sample will cost £138 and MiSeq/rRNA depleted totRNA sample will cost £159 if 24 samples (reasonable diagnostic throughput) are run per lane / flow cell, which is comparable price for output of 24 samples. Detail information about calculations is described in Supplementary data 1.2 and in Supplementary Table 6.

The outcomes presented in this study showed that all included known viruses/viroids could be identified by both NGS approaches. Both approaches successfully identified also two divergent strains of PepMV, which was, despite short fragments of sRNA already shown previously (Kutnjak et al., 2014). However, a putative novel *Cytorhabdovirus* was only detected by analysing the data generated from ribosomal RNA depleted total RNA. Additionally, the results revealed the strength of NGS technology for the simultaneous detection and identification of several different known/unknown plant viruses from a different sample material, with a different amount of viral/viroid nucleotides and in a different host plants. Similar conclusions were derived from studies using other virus enrichment approaches on single or few viral species (Adams et al., 2009; Al Rwahnih et al., 2009; Kreuze et al., 2009; Kutnjak et al., 2014; Visser et al., 2016), e.g., both, sequencing of virion-associatad nucleic acids and sRNAs enabled a discovery of a new virus, previously overlooked by other detection techniques (Candresse et al., 2014). Our study further indicates the advantages of NGS in such cases and strengthens its use as a tool in plant virus/viroid diagnostics.

AUTHOR CONTRIBUTIONS

MR, DK, and NB conceived the idea, AP, MR, DK, and NB designed the experiments. AF provided samples. AP performed laboratory part of the experiment and analyzed the data with the assistance of IA and DK. AP wrote the draft of the manuscript. All authors significantly contributed with reviewing and editing the manuscript.

FUNDING

The work was supported by COST Action FA1407 (DIVAS), thought STSM (short term scientific mission), Euphresco NGS-Detect project and Slovenian Research Agency, AP is a recipient of a Ph.D. research grant from the Slovenian Research Agency.

ACKNOWLEDGMENTS

We thank Dr. Heiko Ziebell for providing the sample material, in this paper labeled as sample V, *Pisum sativum* infected with PNYDV, Dr. Ummey Hany for help with the library preparation and sequencing and Dr. Nataša Mehle for providing the sample material in this paper labeled as sample IX, *Nicotiana benthamiana* infected with CSNV.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.01998/full#supplementary-material

REFERENCES

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., et al. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol. Plant Pathol.* 10, 537–545. doi: 10.1111/j.1364-3703.2009.00545.x

Adams, I., and Fox, A. (2016). "Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis," in *Current Research Topics in Plant Virology,* eds A. Wang and X. Zhou (Cham: Springer), 323–335. doi: 10.1007/978-3-319-32919-2_14

Al Rwahnih, M., Daubert, S., Golino, D., and Rowhani, A. (2009). Deep sequencing analysis of RNAs from a Grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a Novel virus. *Virology* 387, 395–401. doi: 10.1016/j.virol.2009.02.028

Al Rwahnih, M., Dave, A., Anderson, M. M., Rowhani, A., Uyemoto, J. K., and Sudarshana, M. R. (2013). Association of a DNA virus with grapevines affected by red blotch disease in California. *Phytopathology* 103, 1069–1076. doi: 10.1094/PHYTO-10-12-0253-R

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriol. Rev.* 35, 235–241.

Boben, J., Kramberger, P., Petrovic, N., Cankar, K., Peterka, M., Štrancar, A., and Ravnikar, M. (2007). Detection and quantification of tomato mosaic virus in irrigation waters. *Eur. J. Plant Pathol.* 118, 59–71. doi: 10.1007/s10658-007-9112-1

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., et al. (2014). Methods in virus diagnostics: from ELISA to next generation sequencing. *Virus Res.* 186, 20–31. doi: 10.1016/j.virusres.2013.12.007

Borges, F., and Martienssen, R. A. (2015). The expanding world of Small RNAs in plants. *Nat. Rev. Mol. Cell Biol.* 1–12. doi: 10.1038/nrm4085

CLC Analyses-related questions, (2017). "*Analyses-Related Questions*: *De Novo Assembly*." QIAGEN. Available online at https://secure.clcbio.com/helpspot/index.php?pg=kb.page&id=185

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., et al. (2014). Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS ONE* 9:e102945. doi: 10.1371/journal.pone.0102945

Dovas, C. I., Katis, N. I., and Avgelis, A. D. (2002). Multiplex detection of criniviruses associated with epidemics of a yellowing disease of tomato in greece. *Plant Disease* 86, 1345–1349. doi: 10.1094/PDIS.2002.86.12.1345

Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R., and Elena, S. F. (2014). *Viroids: survivors from the RNA world? Annu. Rev. Microbiol.* 68, 395–414. doi: 10.1146/annurev-micro-091313-103416

Gaafar, Y., and Ziebell, H. (2016). *Vicia Faba*, *V. Sativa* and Lens Culinaris as new hosts for pea necrotic yellow dwarf virus in Germany and Austria. *New Dis. Rep.* 34:28. doi: 10.5197/j.2044-0588.2016.034.028

Gutiérrez-Aguirre, I., Mehle, N., Delić, D., Gruden, K., Mumford, R., and Ravnikar, M. (2009). Real-time quantitative PCR based sensitive detection and genotype discrimination of Pepino Mosaic virus. *J. Virol. Methods* 162, 46–55. doi: 10.1016/j.jviromet.2009.07.008

Hull, R. (ed.). (2014). "Replication of plant viruses," in *Plant Virology, 5th Edn.* (Norwich: Academic Press), 341–421.

Itaya, A., Folimonov, A., Matsuda, Y., Nelson, R. S., and Ding, B. (2001). Potato spindle tuber viroid as inducer of RNA silencing in infected tomato. *Mol. Plant Microbe Interact.* 14, 1332–1334. doi: 10.1094/MPMI.2001.14.11.1332

Ivanov, K. I., Eskelin, K., Bašic, M., De, S., Lõhmus, A., Varjosalo, M., et al. (2016). Molecular insights into the function of the viral RNA silencing suppressor HCPro. *Plant J.* 85, 30–45. doi: 10.1111/tpj.13088

Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., et al. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology* 388, E1–E7. doi: 10.1016/j.virol.2009.03.024

Kreuze, J. (2014). "siRNA deep sequencing and assembly: piecing together viral infections,"in *Detection and Diagnostics of Plant Pathogens*, eds M. L. Gullino and P. J. M. Bonants (Dordrecht: Springer), 21–38.

Kumar, S., Udaya Shankar, A. C., Nayaka, S. C., Lund, O. S., and Prakash, H. S. (2011). Detection of tobacco mosaic virus and tomato mosaic virus in pepper and tomato by multiplex RT-PCR. *Lett. Appl. Microbiol.* 53, 359–363. doi: 10.1111/j.1472-765X.2011.03117.x

Kutnjak, D., Elena, S. F., and Ravnikar, M. (2017). Time-sampled population sequencing reveals the interplay of selection and genetic drift in experimental evolution of potato Virus Y. *J. Virol.* 91:e00690-17. doi: 10.1128/JVI.00690-17

Kutnjak, D., Rupar, M., Gutierrez-Aguirre, I., Curk, T., Kreuze, J. F., and Ravnikar, M. (2015). Deep sequencing of virus derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscape of a plant virus population. *J. Virol.* 89, 4760–4769. doi: 10.1128/JVI.03685-14

Kutnjak, D., Silvestre, R., Cuellar, W., Perez, W., Müller, G., Ravnikar, M., et al. (2014). Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach. *Virus Res.* 191, 45–50. doi: 10.1016/j.virusres.2014.07.012

Loebenstein, G., (2008). "Plant virus diseases: economic aspects" in *Desk Encyclopedia of Plant and Fungal Virology,* eds M. H. V. van Regenmortel and W. J. Mahy Brian (Oxford: Academic Press), 426–430.

Loreti, S., Faggioli, F., Cardoni, M., Mordenti, G., Babini, A. R., Poggi Pollini, C., et al. (1999). Comparison of different diagnostic methods for detection of peach latent mosaic viroid. *EPPO Bull.* 4, 433–438. doi: 10.1111/j.1365-2338.1999.tb01414.x

Martínez de Alba, A. E., Flores, R., and Hernández, C. (2002). Two chloroplastic viroids induce the accumulation of the small RNAs associated with post-transcriptional gene silencing. *J. Virol.* 76, 13094–13096. doi: 10.1128/JVI.76.24.13094-13096.2002

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., et al. (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Front. Microbiol.* 8:45. doi: 10.3389/fmicb.2017.00045

Papaefthimiou, I., Hamilton, A., Denti, M., Baulcombe, D., Tsagris, M., and Tabler, M. (2001). Replicating potato spindle tuber viroid RNA is accompanied by short RNA fragments that are characteristic of post-transcriptional gene silencing. *Nucleic Acids Res.* 29, 2395–2400. doi: 10.1093/nar/29.11.2395

Roossinck, M. J., Martin, D. P., and Roumagnac, P. (2015). Plant virus metagenomics: advances in virus discovery. *Phytopathology* 105, 716–727. doi: 10.1094/PHYTO-12-14-0356-RVW

Sabanadzovic, S., Valverde, R. A., Brown, J. K., Martin, R. R., and Tzanetakis, I. E. (2009). Southern tomato virus: the link between the families totiviridae and partitiviridae. *Virus Res.* 140, 130–137. doi: 10.1016/j.virusres.2008.11.018

Soliman, T., Mourits, M. C. M., Oude Lansink, A. G. J. M., and van der Werf, W. (2012). Quantitative economic impact assessment of an invasive plant disease under uncertainty - a case study for potato spindle tuber viroid (PSTVd) invasion into the european union. *Crop Prot.* 40, 28–35. doi: 10.1016/j.cropro.2012.04.019

Verhoeven, J., Th,. J., Jansen, C. C. C., Willemen, T. M., Kox, L. F. F., Owens, R. A., et al. (2004). Natural infections of tomato by citrus exocortis viroid, columnea latent viroid, potato spindle tuber viroid and tomato chlorotic dwarf viroid. *Eur. J. Plant Pathol.* 110, 823–831. doi: 10.1007/s10658-004-2493-5

Visser, M., Bester, R., Burger, J. T., and Maree, H. J. (2016). Next-generation sequencing for virus detection: covering all the bases. *Virol. J.* 13:85. doi: 10.1186/s12985-016-0539-x

Wu, Q., Ding, S. W., Zhang, Y., and Zhu, S. (2015). Identification of viruses and viroids by next-generation sequencing and homology- dependent and homology- independent algorithms. *Annu. Rev. Phytopathol.* 53, 425–444. doi: 10.1146/annurev-phyto-080614-120030

Yelina, N. E., Savenkov, E. I., Solovyev, A. G., Morozov, S. Y., and Valkonen, J. P. (2002). Long-distance movement, virulence, and RNA silencing suppression controlled by a single protein in Hordei- and Potyviruses: complementary functions between virus families. *J. Virol.* 76, 12981–12991. doi: 10.1128/JVI.76.24.12981-12991.2002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on