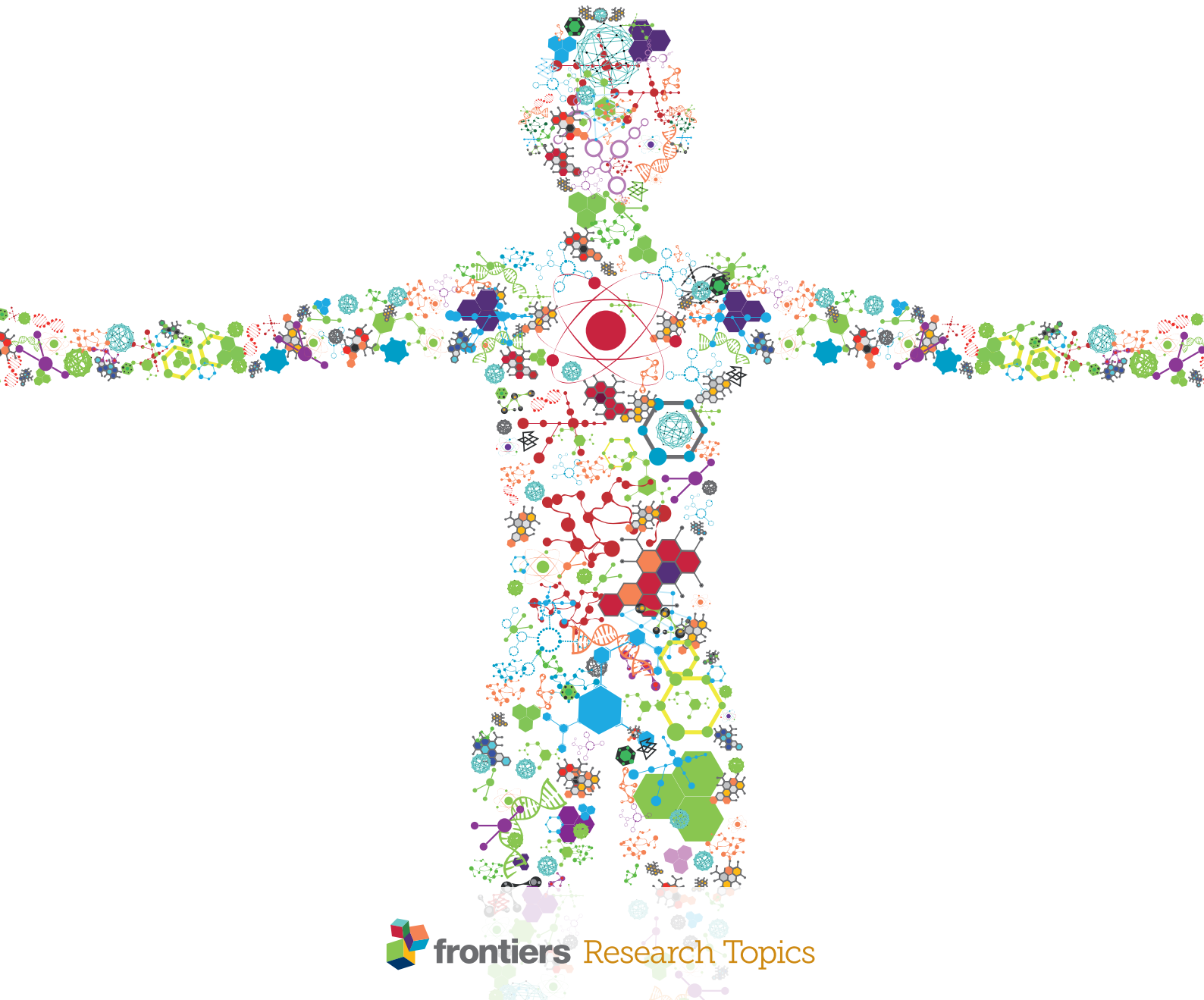
BIOINFORMATICS OF NON-CODING RNAS WITH APPLICATIONS TO BIOMEDICINE: RECENT ADVANCES AND OPEN CHALLENGES

EDITED BY: Carlo Maria Croce, Alfredo Ferro and Alessandro Laganà
PUBLISHED IN: Frontiers in Bioengineering and Biotechnology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-096-1
DOI 10.3389/978-2-88945-096-1

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**

BIOINFORMATICS OF NON-CODING RNAS WITH APPLICATIONS TO BIOMEDICINE: RECENT ADVANCES AND OPEN CHALLENGES

Topic Editors:
Carlo Maria Croce, The Ohio State University, Columbus, USA
Alfredo Ferro, University of Catania, Italy
Alessandro Laganà, Icahn School of Medicine at Mount Sinai, USA

The recent discovery of small and long non-coding RNAs (ncRNAs) has represented a major breakthrough in the life sciences. These molecules add a new layer of complexity to biological processes and pathways by revealing a sophisticated and dynamic interconnected system whose structure is just beginning to be uncovered. Genetic and epigenetic aberrations affecting ncRNA gene sequences and their expression have been linked to a variety of pathological conditions, including cancer, cardiovascular and neurological diseases. Latest advances in the development of high throughput analysis techniques may help to shed light on the complex regulatory mechanisms in which ncRNA molecules are involved. Bioinformatics tools constitute a unique and essential resource for non-coding RNA studies, providing a powerful technology to organize, integrate and analyze the huge amount of data produced daily by wet biology experiments in order to discover patterns, identify relationships among heterogeneous biological elements and formulate functional hypotheses.

This Research Topic reviews current knowledge, introduces novel methods, and discusses open challenges of this exciting and innovative field in connection with the most important biomedical applications. It consists of four reviews and six original research and methods articles, spanning the full scope of the Research Topic.

Citation: Croce, C. M., Ferro, A., Laganà, A., eds. (2017). Bioinformatics of Non-Coding RNAs with Applications to Biomedicine: Recent Advances and Open Challenges. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-096-1

Table of Contents

frontiers in Bioengineering and Biotechnology

EDITORIAL
published: 08 October 2015
doi: 10.3389/fbioe.2015.00156

Editorial: Bioinformatics of non-coding RNAs with applications to biomedicine: recent advances and open challenges

Alessandro Laganà[1]*, *Alfredo Ferro*[2] *and Carlo Maria Croce*[3]

[1] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA [2] Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy [3] Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA

Keywords: ncRNA, miRNA, lncRNA, RNA editing, NGS, biomedicine

The recent advances in the functional characterization of non-protein-coding RNAs (ncRNAs) have represented a major breakthrough in the life sciences. Large-scale projects, such as ENCODE, have shown that the human genome is pervasively transcribed and that a large proportion of the mammalian transcriptome consists of ncRNA transcripts. ncRNA genes can be roughly classified into short ncRNAs (<200 nt) and long ncRNAs (>200 nt).

The first class includes well characterized, infrastructural molecules, such as rRNA, tRNA, and snoRNA, which have a housekeeping role in essential processes like splicing and translation, and regulatory ncRNA such as miRNA and piRNA, which are involved in post-transcriptional regulation of gene expression and in the silencing of transposable elements during germ line development, respectively.

The second class of ncRNA consists of longer transcripts that are still poorly characterized mostly due to their heterogeneity in size, structure, and biogenesis. LncRNA genes outnumber short ncRNAs and are probably more abundant than protein coding genes. Such RNAs exhibit various degrees of conservation and are often polyadenylated and tissue-specific. Accumulating evidence indicates that they likely have a broad range of functions, including chromatin remodeling, gene regulation, and protein transport and trafficking.

Genetic and epigenetic aberrations affecting ncRNA gene sequences and their expression have been linked to a variety of pathological conditions, including cancer, cardiovascular, and neurological diseases.

Recently, high-throughput sequencing techniques have enabled the study of entire transcriptomes at single nucleotide resolution, providing unprecedented details of their organization, expression, modifications, and structure. Bioinformatics tools constitute an essential resource for ncRNA research, providing a powerful means to organize, integrate, and analyze the huge amount of data generated by such technologies.

The aim of this Research Topic is to review current knowledge, introduce novel methods, and discuss open challenges of this exciting and innovative field in connection with the most important biomedical applications. We have collected five original research and methods articles and four reviews, spanning the full scope of the Research Topic.

Two excellent reviews focus on the discovery of ncRNA from NGS data. Kang and Friedländer (2015) surveyed computational tools to predict animal miRNAs from short RNA sequencing data (RNAseq). The authors covered the basics of miRNA prediction, reviewed several methods, described the algorithms, and discussed their strengths. They also described algorithms for specific cases, such as prediction from massively pooled data or in species without reference genomes, and discussed challenges and future directions of the field. Veneziano et al. (2015), instead, provided

Edited and reviewed by:
Richard D. Emes, University of Nottingham, UK

***Correspondence:**
Alessandro Laganà
alessandro.lagana@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *17 September 2015*
Accepted: *25 September 2015*
Published: *08 October 2015*

Citation:
Laganà A, Ferro A and Croce CM (2015) Editorial: Bioinformatics of non-coding RNAs with applications to biomedicine: recent advances and open challenges. Front. Bioeng. Biotechnol. 3:156. doi: 10.3389/fbioe.2015.00156

a more general state-of-the-art coverage of the computational approaches for the discovery and analysis of small and long ncRNA through NGS techniques.

The detection of miRNAs from NGS data becomes an even more challenging task when sequence variants, termed isomiRs, are taken into account. IsomiRs were initially considered sequencing artifacts, but evidence showed that they are functional variants with a specific biological role. Muller et al. (2014) introduced IsomiRage, a streamlined pipeline to identify and analyze isomiRs from next generation sequencing data. The tool is able to distinguish canonical miRNAs from templated and non-templated isomiRs, including 5′- and 3′-extended and trimmed variants.

Two articles of this collection concern RNA editing, a dynamic, widespread process that alters the sequence of RNA transcripts. In particular, A-to-I editing is the most common RNA post-transcriptional modification in human and involves the deamination of adenosine (A) to inosine (I), which is recognized as guanosine (G) by all cellular machineries. Editing may alter both coding and non-coding sequences, with important functional consequences. Nigita et al. (2015) presented a comprehensive state-of-the-art review of databases and computational approaches for the discovery and the analysis of RNA editing, with particular emphasis on ncRNA. They summarized current knowledge and discussed potential consequences of RNA editing on ncRNA, pointing out the lack of tools specifically designed for the detection of editing alterations in lncRNA sequences. This gap was actually addressed by a methods article in our collection by Picardi et al. (2014). They described a novel computational approach to reliably detect A-to-I editing events in human lncRNAs through NGS, based on their previously published package called REDItools. In the presented article, the authors showed the potential of their tools in recovering A-to-I candidates from RNAseq data and provided guidelines to improve RNA editing detection in non-coding RNAs, with specific focus on lncRNAs.

This collection also includes three articles concerning data integration and functional analysis. In the first one, Bonnici et al. (2014) introduced ncRNA–DB, a novel database of ncRNA interaction in human. The database integrates associations among ncRNA, protein coding genes and diseases. It can be searched by a web-based or a command line interface and is also accessible through a Cytoscape app called ncINetView. The second paper, by Alaimo et al. (2014), described ncPred, a novel tool that predicts ncRNA-disease association through tripartite network-based inference. The results of the experimental analysis show that the tool is able to predict more biologically significant associations than its competitors. In the third paper, Li et al. (2015) performed a large-scale integration of publicly available RNA binding protein (RBP) binding sites generated by high-throughput CLIP-Seq technology and identified thousands of RBP–lncRNA interactions. The authors reported combinatorial effects among RBPs and discovered hundreds of disease-related SNPs in RBP binding sites in lncRNA.

The collection also includes a review of design principles and computational tools for the design of synthetic RNAs for gene regulation (Lagana et al., 2014). The article provided guidelines for the design of siRNA, artificial miRNA, antagomiRs, miRNA sponges, and small guide RNA for CRISPRi, and presented strengths and limitations of the different technologies.

Bioinformatics of ncRNA is a vast and rich field and the papers that we selected for this Research Topic address some of its most exciting and pressing challenges. We believe that this volume represents a valuable and useful resource and hope it will be of interest to the many researchers involved in ncRNA research.

REFERENCES

Alaimo, S., Giugno, R., and Pulvirenti, A. (2014). ncPred: ncRNA-disease association prediction through tripartite network-based inference. *Front. Bioeng. Biotechnol.* 2:71. doi:10.3389/fbioe.2014.00071

Bonnici, V., Russo, F., Bombieri, N., Pulvirenti, A., and Giugno, R. (2014). Comprehensive reconstruction and visualization of non-coding regulatory networks in human. *Front. Bioeng. Biotechnol.* 2:69. doi:10.3389/fbioe.2014.00069

Kang, W., and Friedländer, M. R. (2015). Computational prediction of miRNA genes from small RNA sequencing data. *Front. Bioeng. Biotechnol.* 3:7. doi:10.3389/fbioe.2015.00007

Lagana, A., Shasha, D., and Croce, C. M. (2014). Synthetic RNAs for gene regulation: design principles and computational tools. *Front. Bioeng. Biotechnol.* 2:65. doi:10.3389/fbioe.2014.00065

Li, J.-H., Liu, S., Zheng, L.-L., Wu, J., Sun, W.-J., Wang, Z.-L., et al. (2015). Discovery of protein-lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. *Front. Bioeng. Biotechnol.* 2:88. doi:10.3389/fbioe.2014.00088

Muller, H., Marzi, M. J., and Nicassio, F. (2014). IsomiRage: from functional classification to differential expression of miRNA isoforms. *Front. Bioeng. Biotechnol.* 2:38. doi:10.3389/fbioe.2014.00038

Nigita, G., Veneziano, D., and Ferro, A. (2015). A-to-I RNA editing: current knowledge sources and computational approaches with special emphasis on non-coding RNA molecules. *Front. Bioeng. Biotechnol.* 3:37. doi:10.3389/fbioe.2015.00037

Picardi, E., D'Erchia, A. M., Gallo, A., Montalvo, A., and Pesole, G. (2014). Uncovering RNA editing sites in long non-coding RNAs. *Front. Bioeng. Biotechnol.* 2:64. doi:10.3389/fbioe.2014.00064

Veneziano, D., Nigita, G., and Ferro, A. (2015). Computational approaches for the analysis of ncRNA through deep sequencing techniques. *Front. Bioeng. Biotechnol.* 3:77. doi:10.3389/fbioe.2015.00077

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 Bioengineering and Biotechnology

MINI REVIEW
published: 03 June 2015
doi: 10.3389/fbioe.2015.00077

Computational approaches for the analysis of ncRNA through deep sequencing techniques

Dario Veneziano[1]*, *Giovanni Nigita*[1] *and Alfredo Ferro*[2]

[1] Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University, Columbus, OH, USA, [2] Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

The majority of the human transcriptome is defined as non-coding RNA (ncRNA), since only a small fraction of human DNA encodes for proteins, as reported by the ENCODE project. Several distinct classes of ncRNAs, such as transfer RNA, microRNA, and long non-coding RNA, have been classified, each with its own three-dimensional folding and specific function. As ncRNAs are highly abundant in living organisms and have been discovered to play important roles in many biological processes, there has been an ever increasing need to investigate the entire ncRNAome in further unbiased detail. Recently, the advent of next-generation sequencing (NGS) technologies has substantially increased the throughput of transcriptome studies, allowing an unprecedented investigation of ncRNAs, as regulatory pathways and novel functions involving ncRNAs are now also emerging. The huge amount of transcript data produced by NGS has progressively required the development and implementation of suitable bioinformatics workflows, complemented by knowledge-based approaches, to identify, classify, and evaluate the expression of hundreds of ncRNAs in normal and pathological conditions, such as cancer. In this mini-review, we present and discuss current bioinformatics advances in the development of such computational approaches to analyze and classify the ncRNA component of human transcriptome sequence data obtained from NGS technologies.

Keywords: RNA-seq, miRNA, lncRNA, circRNA, bioinformatics

OPEN ACCESS

Edited by:
Fengfeng Zhou, Shenzhen Institutes of Advanced Technology, China

Reviewed by:
Yi Zhao, Chinese Academy of Sciences, China
Raffaele A. Calogero, University of Torino, Italy

***Correspondence:**
Dario Veneziano, Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University, 460 W 12th Avenue, Columbus, OH 43210, USA
dario.veneziano@osumc.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *31 October 2014*
Accepted: *14 May 2015*
Published: *03 June 2015*

Citation:
Veneziano D, Nigita G and Ferro A (2015) Computational approaches for the analysis of ncRNA through deep sequencing techniques. Front. Bioeng. Biotechnol. 3:77. doi: 10.3389/fbioe.2015.00077

Introduction

For over five decades, the central dogma of molecular biology has represented the basis of genetics (Crick, 1970), essentially describing the genetic information flow of life in which DNA and protein, as respectively repository and functional incarnation of that information, have been viewed as the two main actors in the life of the cell, confining RNA simply to the role of template for protein synthesis. Nevertheless, this view of the biological role of RNA, initially apparently exhaustive, has been over time subjected to challenges, as firstly suggested by Gilbert in 1986 (Gilbert, 1986).

As interest on the hypothesized "RNA world" grew, subsequent studies allowed to explore the potential of such new vision (Lee et al., 1993; Fire et al., 1998), eventually leading to one of the most significant biological discoveries of the past decade: the existence of several types of RNAs, each with their specific functions in eukaryotic cells (Eddy, 2001; Todd and Karbstein, 2007). As the ENCODE project has confirmed, most of the human genome is in fact transcribed, but only a very small fraction of it encodes for proteins (Birney et al., 2007; Elgar and Vavouri, 2008). Indeed, the larger remaining portion of the transcribed genomic

output is represented by a diverse family of untranslated transcripts that play crucial roles in many biochemical cellular processes (Mattick, 2001).

These non-coding RNAs (ncRNAs) are divided into two major categories most commonly according to their nucleotide sequence length: small (<200 bp) and long (200 bp or more). Within each category, there are several distinct classes, each one with its own three-dimensional folding and specific function.

From the more popular classes of small structural ncRNAs, such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), focus has shifted in the last 10 years to a set of small RNA classes involved in post-transcriptional regulation: microRNAs (miRNAs) whose precursors (pre-miRNAs) form a peculiar hairpin structure; small interfering RNAs (siRNAs); piwi-interacting RNAs (piRNAs).

Growing interest has more recently emerged also toward long ncRNAs (lncRNAs) which constitute the majority of the non-protein-coding transcripts (Ponting et al., 2009). Having length >200 nt, lncRNAs, already thought of potentially regulating transcription via chromatin modulation, may be also involved in post-transcriptional regulation, organization of protein complexes, and cell–cell signaling (Meldrum et al., 2011).

Finally, an additional class of ncRNAs is represented by circular RNAs (circRNAs) which have been proven to be untranslated, very stable, abundant, and conserved RNA molecules in animals (Jeck et al., 2013).

Yet, despite having been more than a decade since the human genome was sequenced, most transcribed regions are still of unknown molecular function and biological significance.

A potential approach to solving this problem is provided by the ever increasing application of high throughput sequencing technology (HTS), also known as next-generation sequencing (NGS). In fact, numerous transcriptomic sequencing projects are accumulating with increasing rapidity, generating data which are enabling the identification of different types of ncRNAs, and the quantification of their expression levels in different tissues, conditions, and developmental stages.

Why NGS?

Deep sequencing provides a very promising tool. NGS can produce millions of sequences at lower cost in shorter time than before (Meldrum et al., 2011) delivering greater sensitivity and accuracy than previous technologies. Its sensitivity and specificity are above microarray techniques ('t Hoen et al., 2008; Wang et al., 2009); it does not rely on target probe hybridization, permitting the sequencing of the exact transcript on a single nucleotide resolution (Zhou et al., 2011), thus allowing the identification of variations in length or composition (Jung et al., 2010); it requires no previous transcript information (Isakov et al., 2012), utilizing any relevant database to compare and characterize the sequence population (Ronen et al., 2010); it provides high depth of coverage for any library of nucleic acids and it can be modified to study specific properties, e.g., small RNA-seq (sRNA-seq) (Landgraf et al., 2007); it can be used on species for which a full-genome sequence is not yet available; RNA editing events can be detected, and knowledge of polymorphisms can provide direct measurement of allele-specific expression (Malone and Oliver, 2011).

Several HTS platforms are commercially available, each characterized by specific data throughput, read length, error rate, and price (Zhou et al., 2011), providing a wide choice of options.

Current Computational Approaches for ncRNA Analysis from NGS Output

Earlier attempts at whole genome identification of ncRNAs generally had already focused on distinct expression patterns and novel RNA structural families to better characterize the properties of ncRNAs. An example of this is the *incRNA* pipeline employed by Lu et al. (2011) who have developed a comprehensive machine-learned model integrating sequence, structure, and large-scale expression data, both deep sequencing and array. This proves how the complementary nature of combined features can clearly separate ncRNAs from other genomic elements and potentially differentiate between distinct ncRNA types, representing an important advantage of integrative approaches.

Such characterization studies have provided methods that can be adapted to different organisms to identify novel ncRNAs from unannotated genomic regions, paving the way for the development of integrated tools.

Moreover, the large amount of data generated by HTS experiments has made it absolutely necessary to dispose of bioinformatics methods in order to properly store, analyze, and visualize such data.

Generally, a ncRNA bioinformatics analysis system can be comprised of three essential components: a post-sequencing data analysis pipeline for ncRNA detection, classification and expression analysis representing the core of the system; a data module to provide annotation information and storage for the analysis results; a visualization/query system for viewing and functionally analyzing raw data and elaborated results.

As proven by Cordero et al. (2012), statistical detection of differential expression of NGS data gives efficient results when computational strategies employ statistical models based on NB distribution [i.e., baySeq (Hardcastle and Kelly, 2010)] or on variance [i.e., DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014)], as opposed to non-parametric methods which are frequently used for microarray-generated data but are very sensitive to background composition when applied to NGS data.

In order to satisfy the urgent demand for intuitive and efficient data exploration and relieve the growing pressure on handling massive quantities of short-read sequences, several NGS-based RNA transcriptome bioinformatics analysis tools/pipelines have been developed (**Tables 1** and **2**), and below we give an overview of the current most popular ones.

Small ncRNA Transcription Investigation Approaches

Throughout the last decade, the study of the small RNA transcriptome has been gradually recognized to be essential to fully comprehend the complex scenario of transcriptional regulation. For this reason, most currently available tools/pipelines for transcriptome investigation through NGS concentrate on detection/prediction/expression quantification of small RNAs, especially miRNAs.

TABLE 1 | Small non-coding RNA Tool comparison.

		miRDeep	miRDeep*	miRSpring	DARIO	CPSS	ncPRO-seq	CoRAL	RNA-CODE
Package	Online server				✓	✓	✓		
	Stand-alone	✓	✓	✓			✓	✓	✓
Applicable to	Raw data	✓	✓			✓	✓		✓
	Mapped data		✓	✓	✓		✓	✓	
Input format	FASTQ/FASTA	✓	✓			✓	✓		✓
	BAM/SAM		✓	✓	✓		✓	✓	
	BED				✓				
	GFF/GTF							✓	
Assembly	*De novo*								✓
	Reference genome sequence	✓	✓	✓	✓	✓	✓	✓	
Known miRNA detection		✓	✓	✓	✓	✓	✓	✓	✓
Known other ncRNA detection					✓	✓	✓	✓	✓
Novel ncRNA prediction		✓	✓		✓		✓	✓	
Expression analysis		✓	✓	✓	✓	✓	✓	✓	✓
miRNA target prediction			✓			✓			
miRNA target functional enrichment						✓			

TABLE 2 | Long non-coding RNA Tool comparison.

		CoRAL	RNA-CODE	lncRScan	iSeeRNA	CIRI	Annocript	LncRNA2Function
Package	Online server				✓			✓
	Stand-alone	✓	✓	✓	✓	✓	✓	
Applicable to	Raw data		✓				✓	✓
	Mapped data	✓		✓	✓	✓		
Input format	FASTQ/FASTA		✓			✓	✓	✓
	BAM/SAM	✓				✓		
	BED				✓			
	GFF/GTF	✓		✓	✓	✓		
Assembly	*De novo*		✓				✓	
	Reference genome sequence	✓		✓	✓	✓		✓
Known miRNA detection		✓	✓					
Known other ncRNA detection		✓	✓					
Novel ncRNA prediction		✓		✓	✓	✓	✓	✓
Expression analysis		✓	✓	✓	✓	✓		

miRDeep (Friedländer et al., 2008) is believed to be the first stand-alone tool used to analyze large-scale sRNA-seq data in order to detect both known and novel miRNAs. miRDeep employs Bayesian probability controls along the steps of miRNA biogenesis to estimate the false-positive rate and the sensitivity of predictions. The algorithm assumes that if a read is truly related to a pre-miRNA, then it must be a portion either of the loop sequence or of one of the potential two mature sequences in the hairpin. Thus, given the higher abundance of the dominant mature sequence in the cell compared to any other sequence of a pre-miRNA, the higher number of reads in the data will likely correspond to mature sequences, while less frequent reads may map to other parts of the hairpins. Algorithms for mapping and evaluation of free energy, previously under user control, are carried out by Bowtie and Randfold in miRDeep2 (Bonnet et al., 2004; Langmead et al., 2009; Friedländer et al., 2012) in which species conservation has been a key addition as well (Mackowiak, 2011).

Modeled off miRDeep, *mirDeep** (An et al., 2013) employs a miRNA precursor prediction strategy which the authors have proved to outperform both versions of miRDeep as it adopts a different strategy to excise the potential precursor locus range, resulting in a lower number of false negatives. Users can also apply the original miRDeep prediction algorithm, as well as the TargetScan (Lewis et al., 2005) algorithm in order to predict targets for identified known and novel miRNAs.

Great innovation in terms of portability and the elaboration of miRNA processing information is provided by the *miRspring* software (Humphreys and Suter, 2013). The tool generates a small portable interactive miRNA Sequence Profiling document capable of completely reproducing all the information from a significantly larger mapped sequencing data file in bam format (i.e., from a miRNA-Seq experiment), along with providing miRNA processing statistics. In fact, it is the first software that allows to visualize the processing features, seed distribution and relative expression levels of genomic clustered miRNAs from a whole miRNA data set.

Aside miRNA-specific approaches, other software focuses on small RNAs in general.

The first integrated tool ever developed for the analysis and prediction of several classes of small ncRNAs on RNA-seq data originating from arbitrary sequencing platforms is the web service

DARIO (Fasold et al., 2011). The software provides a straighforward interface which allows users to quantify ncRNAs in a completely platform independent way. DARIO annotates reads with information provided by several ncRNA public databases, and excludes mapping loci overlapping with exonic regions, while setting apart those that overlap with introns and intergenic regions for non-annotated ncRNA prediction. An extension of this system to plants has recently been published (Patra et al., 2014).

The web server *CPSS* (Zhang et al., 2012) takes things a step further. The tool can analyze small RNA deep sequencing data coming from single or two paired samples, with special emphasis on miRNAs. Data are classified into several categories of small ncRNAs according to several referred annotations. Matched mapped reads are then quantified for expression analysis (differential in case of two samples), while unmatched ones are employed to predict novel miRNAs also through miRDeep (Friedländer et al., 2008). CPSS also provides users with the possibility to predict target genes for differentially expressed novel/known miRNAs but, like no precedent approach, it also performs functional enrichment analysis of those targets for further experimental or computational studies.

Differently, *ncPRO-seq* (Chen et al., 2012), a stand-alone, comprehensive and flexible ncRNA analysis pipeline, systematically investigates all small ncRNA species in a given annotation family in an unbiased way, providing the user with detailed descriptions of read distribution. Furthermore, the tool defines novel small ncRNA families by identifying regions significantly enriched with short reads not classified under any known ncRNA species, allowing the discovery of previously unknown ncRNA- or siRNA-coding regions.

To address the limitations of RNA function prediction methods in classifying ncRNA classes, the machine learning package *CoRAL* (Leung et al., 2013) classifies RNA transcripts from sRNA-seq data into functional categories by relying on biologically interpretable features more informative than sequence or alignment information, like certain aspects of small RNA biogenesis. Leveraging on the assumption that such biological properties should be consistent within classes of ncRNAs sharing the same molecular function (i.e., across different tissues or organisms), CoRAL was trained in order to identify the most informative features in regard to the molecular mechanisms and metabolic processes of each functional ncRNA class. Based on fragment length, cleavage specificity, and antisense transcription, CoRAL can effectively classify six distinct ncRNA classes among miRNAs and transposon-derived RNAs. Outperforming previous tools such as DARIO and miRDeep2, CoRAL provides the opportunity to annotate ncRNAs in other less well-characterized organisms.

Another tool for ncRNA annotation in NGS data lacking reference genomes is the software *RNA-CODE* (Yuan and Sun, 2013). As ncRNA homology search takes advantage of both sequence and secondary structure similarity, optimization for NGS data is still widely absent, especially when a reference genome is missing. To compensate for this, RNA-CODE combines secondary structure based homology search with *de novo* assembly, adjusting the assembly parameters in a family specific fashion. The sofware assumes that true ncRNA reads sequenced from the same gene can be assembled into contigs with significantly high alignment scores against their native families, while reads aligned by chance tend to share poor overlaps and thus are not likely to be assembled. Sensitivity and accuracy of short reads classification is thus greatly improved. Biogenesis-based properties and homology search results are instead employed for ncRNAs, such as miRNAs, which could not as easily be assembled into contigs. The classification results can then be used to quantify the expression levels of different types of ncRNAs, both small and long, in RNA-seq data of non-model organisms.

Circular RNA Detection Algorithms

The works done by the Brown (Salzman et al., 2012) and Sharpless (Jeck et al., 2013) groups are forerunners of a series of algorithmic approaches to effectively identify circRNA, attempting to compensate the non-uniformity of RNA-seq data sets.

Most algorithms have focused on junction read detection whether leveraging on annotated exon boundaries (Salzman et al., 2012), adopting a two-segment alignment for split reads (Memczak et al., 2013) or relying on RNAase-treated sequencing (Jeck et al., 2013). Nevertheless, all these methods are annotation-dependent and unable to detect certain types of circRNAs having complex alignments and/or subject to experimental bias.

A very recent computational tool proven to outperform any precedent approach in the detection of circRNAs from NGS is *CIRI* (Gao et al., 2015). CIRI is an unbiased, annotation-independent approach employing a *de novo* algorithm able to accurately detect novel circRNAs based on paired chiastic clipping (PCC) signals combined with a filtering system able to remove false positives. CIRI has been able to specifically identify for the first time the prevalence of intronic/intergenic circRNAs as well as fragments specific to them in the human transcriptome, providing novel targets for further functional studies.

Long ncRNA Transcription Investigation Approaches

Long ncRNA investigation is a challenging task, as many more NGS reads are required to achieve adequate coverage compared to mRNAs or other types of ncRNAs. Here below, we describe a few recent computational tools which very well represent the general approach employed by several studies so far (Guttman et al., 2010; Cabili et al., 2011; Pauli et al., 2012).

The pipeline employed by Sun et al. (Sun et al., 2012) makes use of a software they have specifically developed to detect novel lncRNA, called *lncRScan*. The pipeline aims at tackling three of the major technical problems encountered in studying lncRNAs through RNA-seq: eliminating partial transcripts and artifacts in the assembled transcriptome due to RNA-seq-specific issues; identifying lncRNA from the complexity of assemblies; distinguishing lncRNAs from protein-coding mRNAs. After mapping and assembly, the data obtained are compared to a set of combined gene annotations in order to maximize detection and facilitate category labeling of novel transcripts, retaining only multi-exon ones not possessing any annotation for downstream processing. After quality control, the remaining assemblies are given as input to lncRScan for novel lncRNA detection. The tool identifies the candidate lncRNAs through a five-step filtering process: first it organizes input transcripts into five broad categories

according to their genomic location in relation to annotated gene transcripts; transcripts longer than 200 nt are selected, filtering out those with open reading frame (ORF) >300 nt; in the last two steps, phylogenetic analysis and potential aminoacidic sequences of the remaining transcripts are performed in order to exclude any protein-coding potential. Performance evaluation of the pipeline has shown its great ability to filter out mRNAs from the candidate set, while revealing a stringent prediction of true lncRNAs from a test set.

iSeeRNA (Sun et al., 2013) is an SVM-based classifier which can accurately and quickly identify lincRNAs from large datasets, employing conservation, ORF- and nucleotide sequences-based features in order to appropriately distinguish lincRNAs from protein-coding transcripts (PCTs). The best classification results on test sets were produced leveraging on 10 features from the three categories mentioned above: sequence conservation score, being lincRNA less conserved than PCTs in general; ORF length and ORF proportion compared to transcript total length; frequencies of seven di- or tri-nucleotide sequences. Homolog search-based features were instead not included due to lack of annotation for novel PCTs which could foster misclassification. Trained in a species-dependent manner, iSeeRNA allows the user, however, to build additional customized SVMs for other species of interest. Supporting file formats widely used by the RNA-Seq assemblers, iSeeRNA can be easily integrated into transcriptome data analysis pipelines.

More recently, Soreq et al. (Soreq et al., 2014) have integrated full profile characterization of lncRNAs into their comprehensive RNA-seq analysis workflow. The pipeline, based on sample specific database construction, is able to analyze count information from RNA-seq data originating from several platforms and mapping analysis methods. After sequence reads have been mapped, their genome coordinates are intersected with those of the largest available database of reconstructed transcript models for lncRNAs, GENECODE (Derrien et al., 2012). Following appropriate filtering, differential expression of the detected lncRNA candidates is performed using the Bioconductor edgeR package (Robinson et al., 2010) which accounts for biological and technical variability as well as moderating the degree of overdispersion across transcripts, thus improving the reliability of the results.

Musacchia et al. (2015) provide instead a pipeline combining the identification of both coding and long non-coding RNAs in *de novo* generated transcriptomes, without the support of comparative data. *Annocript* identifies putative lncRNAs by leveraging on public annotation databases and sequence analysis software to verify lack of protein/domain similarity, lack of long ORFs, and high non-coding potential.

Finally, an innovative approach in the functional annotation of lncRNAs is provided by Jiang et al. (2015). *LncRNA2Function* provides the first ontology-driven user-friendly web system based on the idea that similar expression patterns across multiple conditions may share similar functions and biological pathways. The tool functionally annotates a single or a set of lncRNAs with the functional terms significantly associated to the set of protein-coding genes significantly co-expressed with the lncRNAs. Standard mapping and assembly are thus followed by the computation of Pearson Correlation Coefficients for all lncRNA–mRNA gene pairs, assigning to each lncRNA a set of significantly co-expressed protein-coding genes which provides the lncRNA with functional and pathway annotations significantly enriched in such set. The tool thus allows to browse the results obtained from an RNA-seq dataset of 19 human normal tissues in order to retrieve the set of lncRNAs associated to a specific functional term, the set of functional terms associated to a lncRNA or assign functional terms to a set of lncRNAs, thus providing a precious resource for lncRNA function investigation.

Acknowledgments

GN was supported by Italian Foundation for Cancer Research (FIRC – 15046). DV was supported by Italian Foundation for Cancer Research (FIRC – 16572). The authors would like to thank the reviewers for their useful suggestions and Fabio Ferri for his help in editing the manuscript.

References

An, J., Lai, J., Lehman, M. L., and Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res.* 41, 727–737. doi:10.1093/nar/gks1187

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi:10.1038/nature05874

Bonnet, E., Wuyts, J., Rouzé, P., and Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* 20, 2911–2917. doi:10.1093/bioinformatics/bth374

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi:10.1101/gad.17446611

Chen, C.-J., Servant, N., Toedling, J., Sarazin, A., Marchais, A., Duvernois-Berthet, E., et al. (2012). ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics* 28, 3147–3149. doi:10.1093/bioinformatics/bts587

Cordero, F., Beccuti, M., Arigoni, M., Donatelli, S., and Calogero, R. A. (2012). Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis. *PLoS ONE* 7:e31630. doi:10.1371/journal.pone.0031630

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563. doi:10.1038/227561a0

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi:10.1101/gr.132159.111

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2, 919–929. doi:10.1038/35103511

Elgar, G., and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* 24, 344–352. doi:10.1016/j.tig.2008.04.005

Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., and Hoffmann, S. (2011). DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* 39, W112–W117. doi:10.1093/nar/gkr357

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

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26, 407–415. doi:10.1038/nbt1394

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40, 37–52. doi:10.1093/nar/gkr688

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16, 4. doi:10.1186/s13059-014-0571-3

Gilbert, W. (1986). Origin of life: the RNA world. *Nature* 319, 618–618. doi:10.1038/319618a0

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510. doi:10.1038/nbt.1633

Hardcastle, T. J., and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11:422. doi:10.1186/1471-2105-11-422

Humphreys, D. T., and Suter, C. M. (2013). miRspring: a compact standalone research tool for analyzing miRNA-seq data. *Nucleic Acids Res.* 41, e147–e147. doi:10.1093/nar/gkt485

Isakov, O., Ronen, R., Kovarsky, J., Gabay, A., Gan, I., Modai, S., et al. (2012). Novel insight into the non-coding repertoire through deep sequencing analysis. *Nucleic Acids Res.* 40, e86. doi:10.1093/nar/gks228

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi:10.1261/rna.035667.112

Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., et al. (2015). LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC Genomics* 16(Suppl. 3):S2. doi:10.1186/1471-2164-16-S3-S2

Jung, C.-H., Hansen, M. A., Makunin, I. V., Korbie, D. J., and Mattick, J. S. (2010). Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. *BMC Genomics* 11:77. doi:10.1186/1471-2164-11-77

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., et al. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129, 1401–1414. doi:10.1016/j.cell.2007.04.040

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi:10.1186/gb-2009-10-3-r25

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

Leung, Y. Y., Ryvkin, P., Ungar, L. H., Gregory, B. D., and Wang, L.-S. (2013). CoRAL: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Res.* 41, e137–e137. doi:10.1093/nar/gkt426

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi:10.1016/j.cell.2004.12.035

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Lu, Z. J., Yip, K. Y., Wang, G., Shou, C., Hillier, L. W., Khurana, E., et al. (2011). Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.* 21, 276–285. doi:10.1101/gr.110189.110

Mackowiak, S. D. (2011). Identification of novel and known miRNAs in deep-sequencing data with miRDeep2. *Curr. Protoc. Bioinformatics* Chapter 12, Unit12.10. doi:10.1002/0471250953.bi1210s36

Malone, J. H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* 9:34. doi:10.1186/1741-7007-9-34

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

Meldrum, C., Doyle, M. A., and Tothill, R. W. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* 32, 177–195.

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

Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., and Sanges, R. (2015). Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. *Bioinformatics*. doi:10.1093/bioinformatics/btv106

Patra, D., Fasold, M., Langenberger, D., Steger, G., Grosse, I., and Stadler, P. F. (2014). plantDARIO: web based quantitative and qualitative analysis of small RNA-seq data in plants. *Front Plant Sci* 5:708. doi:10.3389/fpls.2014.00708

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* 22, 577–591. doi:10.1101/gr.133009.111

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

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Ronen, R., Gan, I., Modai, S., Sukacheov, A., Dror, G., Halperin, E., et al. (2010). miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics* 26, 2615–2616. doi:10.1093/bioinformatics/btq493

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., and Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* 7:e30733. doi:10.1371/journal.pone.0030733

Soreq, L., Guffanti, A., Salomonis, N., Simchovitz, A., Israel, Z., Bergman, H., et al. (2014). Long non-coding RNA and alternative splicing modulations in Parkinson's leukocytes identified by RNA sequencing. *PLoS Comput. Biol.* 10:e1003517. doi:10.1371/journal.pcbi.1003517

Sun, K., Chen, X., Jiang, P., Song, X., Wang, H., and Sun, H. (2013). iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* 14(Suppl. 2):S7. doi:10.1186/1471-2164-14-S2-S7

Sun, L., Zhang, Z., Bailey, T. L., and Perkins, A. C. (2012). Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. *Bioinformatics* 13, 331. doi:10.1186/1471-2105-13-331

't Hoen, P. A. C., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H. A. M., de Menezes, R. X., et al. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 36, e141.

Todd, G., and Karbstein, K. (2007). RNA takes center stage. *Biopolymers* 87, 275–278. doi:10.1002/bip.20824

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484

Yuan, C., and Sun, Y. (2013). RNA-CODE: a noncoding RNA classification tool for short reads in NGS data lacking reference genomes. *PLoS ONE* 8:e77596. doi:10.1371/journal.pone.0077596

Zhang, Y., Xu, B., Yang, Y., Ban, R., Zhang, H., Jiang, X., et al. (2012). CPSS: a computational platform for the analysis of small RNA deep sequencing data. *Bioinformatics* 28, 1925–1927. doi:10.1093/bioinformatics/bts282

Zhou, L., Li, X., Liu, Q., Zhao, F., and Wu, J. (2011). Small RNA transcriptome investigation based on next-generation sequencing technology. *J. Genet. Genomics* 38, 505–513. doi:10.1016/j.jgg.2011.08.006

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
BIOENGINEERING AND BIOTECHNOLOGY

REVIEW ARTICLE
published: 26 January 2015
doi: 10.3389/fbioe.2015.00007

Computational prediction of miRNA genes from small RNA sequencing data

Wenjing Kang and Marc R. Friedländer*

Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Noam Shomron, Tel Aviv University, Israel
Patrick Xuechun Zhao, Samuel Roberts Noble Foundation, USA

***Correspondence:**
Marc R. Friedländer, Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Box 1031, Solna 17121, Sweden
e-mail: marc.friedlander@scilifelab.se

Next-generation sequencing now for the first time allows researchers to gage the depth and variation of entire transcriptomes. However, now as rare transcripts can be detected that are present in cells at single copies, more advanced computational tools are needed to accurately annotate and profile them. microRNAs (miRNAs) are 22 nucleotide small RNAs (sRNAs) that post-transcriptionally reduce the output of protein coding genes. They have established roles in numerous biological processes, including cancers and other diseases. During miRNA biogenesis, the sRNAs are sequentially cleaved from precursor molecules that have a characteristic hairpin RNA structure. The vast majority of new miRNA genes that are discovered are mined from small RNA sequencing (sRNA-seq), which can detect more than a billion RNAs in a single run. However, given that many of the detected RNAs are degradation products from all types of transcripts, the accurate identification of miRNAs remain a non-trivial computational problem. Here, we review the tools available to predict animal miRNAs from sRNA sequencing data. We present tools for generalist and specialist use cases, including prediction from massively pooled data or in species without reference genome. We also present wet-lab methods used to validate predicted miRNAs, and approaches to computationally benchmark prediction accuracy. For each tool, we reference validation experiments and benchmarking efforts. Last, we discuss the future of the field.

Keywords: miRNA, microRNA, gene prediction, next-generation sequencing data

miRNA BIOLOGY

microRNAs (miRNAs) are a class of small RNAs (sRNAs) around 22 nucleotides in length. They are never translated, but post-transcriptionally reduce the output of protein coding genes (Kloosterman and Plasterk, 2006; Bushati and Cohen, 2007; Farazi et al., 2008; Ghildiyal and Zamore, 2009). They have been found in all animals studied, in numbers that appear to correlate with organismal complexity, for instance, nematodes have around 200 miRNA genes while humans have more than 3000 (Kozomara and Griffiths-Jones, 2011; Friedländer et al., 2014). Mutant animals that are void of miRNAs either die at early embryonic stages or have severe developmental defects, showing the importance of the regulation they infer (Bernstein et al., 2003; Giraldez et al., 2005; Morita et al., 2007; Wang et al., 2007). More than half of all protein coding transcripts are estimated to be under regulation of miRNAs in one or more cellular contexts (Friedman et al., 2009). Thus, it is not surprising that miRNAs are involved in numerous biological contexts, ranging from formation of cell identify to development (Stefani and Slack, 2008).

miRNA BIOGENESIS

The majority of miRNAs are transcribed by Polymerase II and have features similar to protein coding transcripts: a 5′ cap, exons, and a poly(A)-tail (**Figure 1**). Each of the primary transcripts harbors one or more characteristic RNA hairpin structures around 60 nucleotides in length. While in the nucleus, these structures can be recognized by the Microprocessor complex, consisting of Drosha and DGCR8 proteins, which cleave the hairpin out of the primary transcript (Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). The hairpin is then exported to the cytosol, where it undergoes a second cleavage by Dicer, a canonical component of the RNA interference pathway (Bernstein et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Knight and Bass, 2001). The cleavage releases three products: the mature miRNA guide strand, the miRNA passenger strand, and the loop. These three products fall in determined positions: the guide and the passenger form an RNA duplex with two nucleotides 3′ overhangs, while the loop consists of the terminal end of the hairpin, positioned between the guide and the passenger strands (Ha and Kim, 2014). While the loop and the passenger strands are generally degraded as bi-products of the biogenesis, the guide miRNA remains bound to an Argonaute protein, which is part of the miRNP complex. It is not always the same strand that is fated to be bound to the Argonaute protein, in the case of many miRNA hairpins either strand can be incorporated and repress targets (Okamura et al., 2008; Guo and Lu, 2010; Yang et al., 2011). The mature miRNA can guide the effector complex to target sites, typically located in 3′ UTRs of mRNAs, through partial base complementarity (Lai, 2002; Bartel, 2009). Once bound, the complex reduces protein output of the transcript, either by destabilizing it through shortening of the poly-A tail, inhibiting its translation or by re-localizing it to subcellular ribo-protein particles, where it is inaccessible to the translation machinery (Filipowicz et al., 2008; Huntzinger and Izaurralde, 2011). Some miRNAs follow non-canonical biogenesis

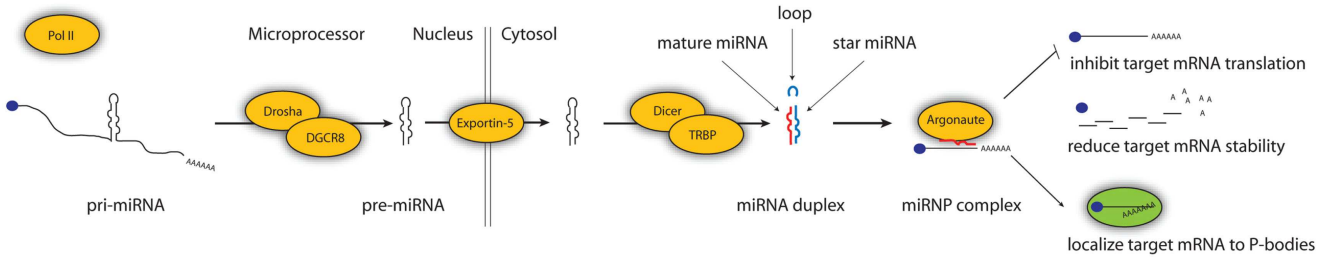


FIGURE 1 | miRNA biogenesis and function. miRNAs are transcribed as primary transcripts or are sometimes derived from exons or introns of hosts transcripts. Characteristic hairpin RNA structures are recognized by Drosha and DGCR8 and cleaved out. The hairpin is exported to the cytosol and cleaved by Dicer, which is a part of the canonical RNA interference pathway, releasing three products: the two miRNA strands (the "mature" or "guide" strand and the "star" or "passenger" strand) and the terminal loop. The guide strand is then bound by an Argonaute protein, which is part of the miRNP effector complex. Once thus bound, the miRNA can bind to target sites, often located in the 3′ UTR of protein coding transcripts, and guide the effector complex to inhibit translation of the target, cause its degradation, or relocate it to subcellular foci, where they are no longer accessible to the translation machinery.

pathways, but are believed to function like the canonical sequences (Ha and Kim, 2014). Altogether, it is estimated that around 60% of all human protein coding transcripts are regulated by miRNAs in one or more cellular conditions (Friedman et al., 2009).

miRNAs IN HUMAN DISEASE

Given the prevalence of miRNA regulation, it is not surprising that miRNAs have been involved in numerous human diseases. These regulators appear to play particularly critical roles in cancers, where they can function as onco-genes or tumor suppressors. For instance, the miR-17–92 cluster is found to be up-regulated in several cancers (He et al., 2005), and miR-15 and miR-16 are often deleted in leukemias (Cimmino et al., 2005). Although some miRNAs can function as onco-genes, they are in most cases down-regulated individually or collectively in cancers (Medina and Slack, 2008). miRNAs are important in cell differentiation and formation of cell identity, and often cancer cells revert to more undifferentiated states. In addition to cancers, miRNAs have been involved in many types of diseases including: cardiovascular, immunological, neurodegenerative, and psychiatric (Taft et al., 2010; Esteller, 2011). In disease, miRNA function can be perturbed in several ways: by down-regulation of the biogenesis factors (Hill et al., 2009), by mutation in the miRNA locus (Mencia et al., 2009), by loss or gains of the miRNA genes (Zhang et al., 2006b), or by epigenetic changes such as hypermethylation (Davalos et al., 2012). There are also cases where disease is caused by mutations that destroy (Christensen et al., 2009) or create (Abelson et al., 2005) target sites in the 3′ UTR of protein coding transcripts.

Before the role of a miRNA in a given disease can be investigated, it must be discovered and annotated. Many miRNAs have specific expression patterns and may not be highly expressed outside the particular tissue that is studied, and may not yet have been discovered. Therefore, miRNA prediction is an important first analysis step of sRNA-seq analysis in clinical context. miRNA prediction can also be used for basic research, when annotating the complement of regulatory RNAs in emerging model systems. The purpose of this review is to present the methods used to discover new animal or human miRNA genes from sRNA-seq data. We will focus on published methods that can be downloaded and run, without the user needing to implement algorithms as software by him/herself. We will discuss the strengths of the distinct methods, and will reference the studies in which the methods have been benchmarked computationally. Thus, this review can serve as a platform for the reader to decide which method is ideally suited for his miRNA prediction use case. Finally, we will present low and high-throughput methods to validate the discovered miRNA candidates.

miRNA PREDICTION

PREDICTION FROM GENOME SEQUENCE

The biogenesis of miRNAs is key to their discovery. When the field was still young and little data were available, researchers would search the genome sequences for loci that would give rise to RNA hairpin structure if transcribed. These methods have combined structure prediction with either scoring (Lai et al., 2003; Lim et al., 2003; Ohler et al., 2004; Wang et al., 2005) or rules-based (Dezulian et al., 2006; Zhang et al., 2006a) or machine-learning classification (Nam et al., 2005; Jiang et al., 2007; Sheng et al., 2007) of the hairpin features. Some of the methods have incorporated conservation information into the prediction; in fact, one approach has used phylogenetic shadowing to detect the characteristic conservation profile of miRNAs, where the miRNA strands are more conserved in sequence than the terminal loop (Berezikov et al., 2005). However, it is impossible to know from the genome DNA sequence if a locus is really transcribed and gives rise to mature miRNAs. Thus, considering the size of most animal genomes, these methods yield many false positive hairpins that are either not transcribed or do not interact with the biogenesis factors. For instance, in the human genome, around 11 million loci would give rise to hairpin structures if transcribed (Bentwich, 2005), but only a few thousands of them are actually cleaved to mature miRNAs (Kozomara and Griffiths-Jones, 2011; Friedländer et al., 2014).

SANGER SEQUENCING

For an unbiased detection of miRNAs, methods were developed to directly sequence sRNAs. This was done by separating them from other transcripts on high-resolution gels, and sequencing by Sanger sequencing (Lagos-Quintana et al., 2001; Lau et al.,

2001; Lee and Ambros, 2001). Because of the limited throughput of this technology, typically just a few hundreds of sRNAs were detected, and many of these would be degradation products of longer transcripts such as mRNAs, rRNAs, and tRNAs, or even from un-annotated transcripts. To ensure that the predicted miRNAs were genuine, researchers would filter out sequences mapping to known non-miRNA transcript annotations, and would require that the predicted miRNA was located in a loci that could give rise to a hairpin transcript (Ambros et al., 2003). More specifically and in accordance with miRNA biogenesis, the predicted sequence should be located on a hairpin arm. Further, if two sequences should locate to the same hairpin, it was required that they should form a duplex with two nucleotide 3′ overhangs, as expected from Dicer processing.

NEXT-GENERATION SEQUENCING

In 2006, the first next-generation sequencing instruments became commercially available, allowing orders of magnitude increase in data generation. For instance, the current Illumina HiSeq 2500 instruments can sequence around one billion sRNAs in <2 days. This sequencing power can be distributed between several experiments, but still sRNA-seq studies detect millions of transcripts per sample. Since a mammalian cell typically contains on the order of 100,000 miRNA transcripts (Calabrese et al., 2007), this means that sequences that are present in less than one molecule per cell can still be detected. This also holds for other clades, for instance, the *lsy-6* miRNA, which is expressed in only a single neuron in the entire nematode body (Johnston and Hobert, 2003), is now routinely detected in sRNA-seq experiments (unpublished results).

The sensitivity of these sequencing methods means that very lowly expressed sRNAs other than miRNAs are also detected. These can include short interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs) but can also be rare degradation products of longer transcripts like rRNAs, tRNAs, and mRNAs or un-annotated transcripts. In addition to this, there is now emerging evidence that transcripts like tRNAs can undergo endonucleolytic cleavage at specific positions to produce functional sRNAs (Chen and Heard, 2013). Altogether, this means that sRNAs sequenced in a single experiment can originate from millions of distinct loci in the human genome (Friedländer et al., 2008). The methods that were developed to predict miRNAs from Sanger sequencing should only handle a few thousand loci. Therefore, they are not specific enough to be applied to next-generation sequencing data, and produce numerous false positives. These false positives are transcribed and form hairpins, but the sRNAs generated from them are degradation products resulting from normal RNA turnover. Thus, accurately identifying the miRNAs in this complex landscape of sRNAs is a daunting task.

To reduce false positives, methods to predict miRNAs from sRNA-seq employ post-filtering steps beyond what is used for Sanger sequencing. The next-generation discovery methods almost all require the presence of a hairpin structure, and the formation of a duplex if both miRNA strands are detected. In addition, many methods require that the candidate precursors do not overlap known non-miRNA annotations (Berninger et al., 2008). Hairpins that pass these requirements are then exposed to a further filter step. These steps can be rule-based or can involve probabilistic scoring or machine learning (see below). The features that are evaluated can be divided into *structure* features and *signature* features (Friedländer et al., 2008). The first reflect how well the hairpin structure conforms to known miRNA precursors. For instance, most of the nucleotides in the putative duplex should be base paired, and the hairpin should not contain large bulges besides the terminal loop. Some methods also require that the structure should be energetically stable, as this is a hallmark of genuine miRNA hairpins. The *signature* is a measure of how well the distribution of sequenced RNAs fit in the hairpin structure. For instance, every sequenced RNA should correspond to either guide or passenger strand, or to the terminal loop. The guide and passenger RNAs should form a duplex with two nucleotide 3′ overhangs, as is typical of Dicer processing. Further, it is expected that the candidate miRNA guide strand is detected several times, given the sensitivity of next-generation sequencing. Last, since it is known that processing of Drosha and Dicer produces clearly defined 5′ ends, the sequenced RNAs should align neatly in this end (Ruby et al., 2006).

Besides the core prediction methods, source for predicting miRNAs differ in other respects. This includes the mapping tool, whether read pre-processing is provided, whether the tool has a graphic user interface or must be operated on the command line and whether additional analyses like expression analyses and target predictions are supported. Also, some methods are not just applicable for animal miRNAs, but also for plant sequences. Finally, some methods have been tested by computational benchmarking in several studies and their predictions validated in the wet-lab. In the following section, we describe the tools of the field in alphabetical order (**Table 1**).

SPECIFIC ALGORITHMS

deepBlockAlign

deepBlockAlign is innovative in that it provides advanced scoring of the read signature, but does not evaluate the RNA structure (Langenberger et al., 2012; Pundhir and Gorodkin, 2013). deepBlockAlign uses a variant of Needleman–Wunsch to identify blocks of mapped reads that have similar features, including read begin positions and block height. In a second step, similar groups of blocks are identified using a variant of the Sankoff algorithm. These groups of blocks correspond to gene loci. To predict novel miRNAs, the method finds loci that have block features similar to known miRNAs. While the profiles might be different for plants and animals, or specific to particular tissues or pathological conditions, the method can compare to all known profiles from the entire miRBase database of miRNAs, giving it good coverage. Since this method does not evaluate the RNA structure, it can predict miRNAs that do not have canonical structure, or whose conformation is not easily predicted by computational methods. Alternatively, it can be combined with down-stream structure analysis, to further improve specificity[1].

miRanalyzer

miRanalyzer first removes reads that map to known miRNAs or other transcripts (Hackenberg et al., 2009). The remaining reads

[1] http://rth.dk/resources/dba/

Table 1 | Tools for predicting animal miRNAs from sRNA-seq data.

Tool	Algorithm	Mapping tool	Tested in plants	Performance comparison	Validated in wet-lab	Pre-process data	Quantifies expression	Target prediction	User interface
GENERAL TOOLS									
deepBlockAlign	Read block alignment	Not included	Yes	Langenberger et al. (2012), and Pundhir and Gorodkin (2013)	No	No	No	No	Graphics, webserver
miRanalyzer	Random forest	Prefix tree	No	Hackenberg et al. (2009)	See below	Partial	Differential expression	MiRanda and TargetScan	Graphics, webserver
miRanalyzer (update)	Random forest	Bowtie	Yes	An et al. (2013), Friedländer et al. (2012), Hackenberg et al. (2011) Hansen et al. (2014), Pundhir and Gorodkin (2013), and Williamson et al. (2013)	RT-PCR (Smith et al., 2013), Northern blot (Mayoral et al., 2014)	Yes	Differential expression	TargetSpy	Graphics, webserver, and standalone
miRCat	Rules-based	PatMaN	Yes	Moxon et al. (2008)	RT-PCR (Kohli et al., 2014, and Pandey et al., 2014), Northern blot (Donaszi-Ivanov et al., 2013)	Yes	Yes (mirprof), differential expression (colide)	PAREsnip	Graphics, webserver, and standalone
miRDeep	Bayesian	Megablast	No	An et al. (2013), Friedländer et al. (2008, 2012), Hendrix et al. (2010), and Williamson et al. (2013)	Northern blot (Friedländer et al., 2008, 2009), RT-PCR (Friedländer et al., 2012)	No	Yes	No	No graphics, standalone
miRDeep2	Bayesian	Bowtie	No	An et al. (2013), Friedländer et al. (2012), Hansen et al. (2014), and Williamson et al. (2013)	Knock-down (Friedländer et al., 2012), RT-PCR (Metpally et al., 2013)	Yes	Yes	No	Graphics, standalone
miRDeep*	Bayesian	Bowtie (java version)	No	An et al. (2013), and Hansen et al. (2014)	RT-PCR, knock-down (An et al., 2013)	Yes	Yes	TargetScan	Graphics, standalone (java software)
MIReNA	Rules-based	Megablast	Yes	An et al. (2013), Friedländer et al. (2012), and Mathelier and Carbone (2010)	Knock-down (Friedländer et al., 2012)	No	No	No	No graphics
miREvo	Bayesian	Bowtie	No	No	No	Yes	Yes	No	Graphics, standalone
miRExpress	Sequence homology	Custom mapping tools	No	No	No	Yes	Yes	No	No graphics, standalone
miRTRAP	Rules-based	Not included	No	An et al. (2013), Friedländer et al. (2012), and Hendrix et al. (2010)	Knock-down (Friedländer et al., 2012), Northern blot (Hendrix et al., 2010)	No	No	No	No graphics

(Continued)

Table 1 | Continued

Tool	Algorithm	Mapping tool	Tested in plants	Performance comparison	Validated in wet-lab	Pre-process data	Quantifies expression	Target prediction	User interface
MASSIVELY POOLED DATA									
miRdentify	Feature scoring	Bowtie	No	Hansen et al., 2014	RT-PCR (Hansen et al., 2014)	Yes	No	No	No graphics
PREDICTION WITHOUT REFERENCE GENOME									
MirPlex	Support vector machine	Not included	Yes	Mapleson et al. (2013)	Knock-out (Mapleson et al., 2013)	No	No	No	No graphics
MIRPIPE	Sequence homology	BLASTN	No	Kuenne et al. (2014)	No	Yes	Yes	No	Graphics, webserver, and standalone

Algorithm: the core algorithm for identifying miRNAs. Mapping tool: software used to trace sequenced RNAs to the reference sequences. Tested in plants: if the method has been benchmarked with plant data. Performance comparison: studies that have benchmarked the performance of the tool. Validated in wet-lab: studies that have validated predicted miRNA candidates with experimental methods. Given the overall number of miRNA studies, this list may not be exhaustive. Pre-process data: tools that prepare the FASTQ sequence data for the mapping and prediction steps. Quantifies expression: tools that report estimated miRNA abundances. In addition, some tools report miRNAs that are differentially expressed between samples. Target prediction: tools that predict targets of candidate miRNAs. User interface: tools that have a graphic user interface (as opposed to being operated from the command line). Tools that are run on a webserver (as opposed to being installed and run on a local machine).

are considered as potential new miRNAs. They are evaluated as miRNAs using a random forest machine learning approach. The classifier is initially trained on a set of known miRNAs from human, rat, or nematode and dozens of features are considered, including energetics, structure, bulges, and the number of reads mapping. The tool has fitted parameters for each species analyzed and on publication provided packages for seven commonly used species. miRanalyzer is available through a webserver, making it easily accessible for biologists with little computational experience[2].

miRanalyzer (UPDATE)

miRanalyzer (update) is an improved version with several new features. It uses bowtie (Langmead et al., 2009) for much faster and less memory-intensive mapping, and it includes parameter packages for 31 species, including 6 plants (Hackenberg et al., 2011). In addition, it can perform differential expression analysis of the profiled miRNAs and can predict targets using the TargetSpy tool. In addition to the web server version, it has a stand-alone version that can be downloaded and run on local machines. miRanalyzer predictions have been validated with several wet-lab methods (Smith et al., 2013; Mayoral et al., 2014). Since miRanalyzer often predicts more new miRNAs than do other tools, it is well suited for studies where the predictions will be filtered by additional computational tools or by high-throughput wet-lab validations[2].

miRCat

miRCat has been used successfully to predict miRNAs in several plants (Szittya et al., 2008; Pantaleo et al., 2010; Mohorianu et al., 2011) and has recently been adapted to animal sequences, including butterflies (Surridge et al., 2011). miRCat uses a rules-based approach that eliminates candidates with features that are not consistent with miRNA biogenesis (Moxon et al., 2008; Stocks et al., 2012). Numerous features are investigated, including the number of read stacks in the locus, the number of reads mapping anti-sense to the locus, the size of bulges in the candidate miRNA duplex, the number/fraction of paired nucleotides in the duplex and in the hairpin, and the energetic stability of the hairpin. miRCat is part of a suite, the UEA workbench, which includes numerous computational tools, some which can be applied to the analysis of non-miRNA small RNA sequences. miRCat predictions have been validated in several systems (Donaszi-Ivanov et al., 2013; Kohli et al., 2014; Pandey et al., 2014). Since it was developed for plant miRNAs that are more variable in structure, it could be well suited for detecting animal miRNA hairpins that are not typical for this clade[3].

miRDeep

miRDeep first filters all candidates whose structure and read signature are inconsistent with Drosha/Dicer processing (Friedländer et al., 2008). In the next step, the fit of the structure and signature to an explicit model of miRNA biogenesis is scored using Bayesian statistics. Specifically, miRDeep scores the number of reads supporting biogenesis, the presence of a miRNA passenger strand, the

[2]http://bioinfo2.ugr.es/miRanalyzer/standalone.html

[3]http://srna-workbench.cmp.uea.ac.uk/tools/analysis-tools/mircat/)

presence of a conserved miRNA seed and the absolute and relative energetic stability of the hairpin. While miRDeep can be run on data filtered for known non-miRNA annotations, it can perform robust prediction without this filtering. This means that miRNAs derived from non-canonical host transcripts, such as snoRNAs, can be identified (Ender et al., 2008). Further, it does not require parameters fitted to specific species, meaning that it is not at a disadvantage when mining emerging model systems. The tool has been extensively benchmarked and validated by experimental methods (Friedländer et al., 2008, 2009; Metpally et al., 2013), and has been adapted by several other research groups (Yang and Li, 2011; Yang and Qu, 2012; Wu et al., 2013)[4].

miRDeep2

miRDeep2 improves the previous version, primarily by making more robust predictions when faced with very deep sequencing data (Mackowiak, 2011; Friedländer et al., 2012). This includes improved excision of candidate hairpins from the genome, allowing for anti-sense miRNAs and moRs (see miRTRAP below). In addition, the tool has been improved in terms of computational efficiency, implementing better tools like bowtie (Langmead et al., 2009), and it features graphics output. Last, it has been tested in seven species, using the exact same parameters, and introduces knock-down of key proteins necessary for miRNA maturation to validate that novel candidates depend on the miRNA biogenesis pathways for their expression[4].

miRDeep*

miRDeep* is an extension of the first miRDeep algorithm, and incorporates many improvements similar to miRDeep2, although it was developed by a separate research group (An et al., 2013). It features pre-processing, bowtie mapping, improved precursor excision, and target prediction for known and novel miRNAs. The tool has an extensive graphical user interface and is implemented entirely in java without requiring any pre-dependent computational tools, making it portable and easy to install. The computational efficiency makes it run on a home computer[5].

MIReNA

MIReNA is a flexible tool to predict novel miRNAs from known miRNA sequences, next-generation sequencing data, long transcripts, or hairpin precursors (Mathelier and Carbone, 2010). It uses a rules-based scheme with sharp cut-offs to classify miRNAs based on five criteria: the lack of base pairing in the mature miRNA, the difference in length between the two candidate miRNA strands, the fraction of base-paired nucleotides in the hairpin, and two measures of energetic stability. As a second filtering step, it considers only hairpins where the sequenced RNAs map in consistence with Drosha/Dicer processing. MIReNA can consider several potential miRNA duplexes within one precursor structure, e.g., within multiple stem precursors, giving it the potential to predict non-canonical miRNAs[6].

miREvo

miREvo build on the miRDeep2 predictor (above) but extends it for evolutionary analyses (Wen et al., 2012). Specifically, it uses whole-genome alignments to identify miRNA homologs in related species. It also includes tools to compare expression of miRNA homologs across species, if sRNA-seq data are available for both species. It uses modified prediction parameters for plant analyses[7].

miRExpress

miRExpress is a tool for profiling miRNA expression from sRNA-seq data (Wang et al., 2009). However, it includes a function to predict miRNAs based on sequence homology. It maps each read that does not correspond to a known reference miRNA against miRBase sequences, keeping only perfect matches. These reads are then mapped against the reference genome, and the structure evaluated with the mfold structure prediction software (Zuker, 2003)[8].

miRTRAP

miRTRAP uses a rules-based approach with two filtering steps (Hendrix et al., 2010). In the first one, all candidate miRNAs whose structure and read signatures do not conform to Drosha/Dicer processing are eliminated. In the second step, all candidates that are not located in sRNA deserts are removed. This second step builds on the observation that miRNAs typically generates blocks of sRNAs with few or no sequenced RNA mapping to the anti-sense strand or in the general vicinity. In addition to this innovative filtering step, miRTRAP has high accuracy when predicting miRNAs with moRs, which are sRNAs generated from the flanks of the precursor hairpin. This development was necessary, as the tool was initially developed for identifying miRNAs in sea squirt, a species unusually rich in moRs (Shi et al., 2009)[9].

SPECIAL APPLICATIONS

MASSIVELY POOLED DATA

Many researchers who apply miRNA prediction tools to sequencing data want to mine their own in-house data. These could be sequences from an emerging model organism, or from a human tissue of interest. The tools described above are all optimized for analyzing a limited number of data sets, ranging from maybe 1 to 20 sets. However, some studies compile all the available sRNA-seq data for a given species to give the best possible miRNA annotation. There are numerous advantages to pooling tens or hundreds of datasets (Friedländer et al., 2014). First, if the guide and passenger strands are detected in two distinct data sets, combining the information can allow analysis of the duplex features. Second, lowly expressed miRNAs might not be well profiled in single datasets, where it is difficult to evaluate the read signature. Third, since sRNA-seq library preparation involves a PCR amplification step, there is no guarantee that 10 sequencing reads in 1 dataset do not correspond to a single over-amplified sRNA. In contrast, if the same sequence is detected in data from 10 distinct tissues, this provides independent evidence that the biogenesis is common.

[4]https://www.mdc-berlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep
[5]http://www.australianprostatecentre.org/research/software/mirdeep-star
[6]http://www.lgm.upmc.fr/mirena/index.html
[7]http://omictools.com/mirevo-s962.html
[8]http://mirexpress.mbc.nctu.edu.tw
[9]http://flybuzz.berkeley.edu/miRTRAP.html

Massively pooled sRNA-seq data have previously been used to predict miRNAs in general (Friedländer et al., 2014), or of the specific mirtron class (Ladewig et al., 2012). These are hairpins, which are released by intronic splicing rather than Drosha cleavage. Some mirtrons are short and their hairpin ends are defined by the splice signals, while others are longer, and one end is trimmed to define the hairpin end (Berezikov et al., 2007; Okamura et al., 2007; Ruby et al., 2007). In addition, the miRBase database employs massively pooled data to refine the miRNA annotations and define a high-confidence set of sequences (Kozomara and Griffiths-Jones, 2014). The software used in these studies has, however, not been published, so the methods are not described in detail here.

miRdentify

miRdentify has recently been released to the public to analyze massive pooled data (Hansen et al., 2014). It requires that both guide and passenger miRNA strands are detected and evaluates 10 features of the structure and signature, including precision of 5′ end processing, two nucleotide 3′ overhangs, and several aspect of stability. For each feature, the cut-off is set so that 1% of known miRNAs is excluded. Together, the requirement for detection of both strands and the 10 features constitute stringent criteria that produce miRNA candidates with features similar to known hairpins (Hansen et al., 2014). The method thus, to some extent, trades off sensitivity to report high-quality candidates[10].

PREDICTION WITHOUT A REFERENCE GENOME

The majority of miRNA prediction tools require a reference genome as input to enable the excision of miRNA hairpin sequences, whose RNA structures and signatures are considered as key features for miRNA prediction. However, even though the price of next-generation sequencing technologies decreases, only a handful of model species have fully assembled high-quality reference genomes. Thus, many researchers rely on emerging model species without reference genomes, and novel methods are needed to discover new miRNAs in order to further study their function. One way to address this problem is to use a closely related species genome as proxy reference sequence to identify conserved miRNA. Such a study has been undertaken to discover mosquito miRNAs by mapping the sRNA-seq against the genomes of three related insect species (Etebari and Asgari, 2014). For this purpose, the miRanalyzer tool was used, and it was found that the prediction accuracy is affected by the evolutionary distance between the species of interest and the proxy species. Overall, the most abundant and conserved miRNAs were identified in this study, but the approach might be less successful for species that do not have closely related species with genome sequences.

MirPlex

MirPlex is a tool that requires only sRNA datasets as input with no genome sequences needed (Mapleson et al., 2013). It uses a multi-stage process to identify genuine miRNA duplexes. First, all overlapping sequences are assembled into contigs, and contigs that are too long to be miRNAs are discarded (> 30 nucleotides). Second, the remaining sequences are copied into two duplicate datasets followed with separate filter pathways to obtain candidate miRNA guide and miRNA passenger sequences. Last, the candidate miRNA guide and miRNA passenger sequences are then paired into duplexes for the classification. The core algorithm of MirPlex uses a support vector machine to classify genuine miRNA duplexes based on 20 features that divided into three categories: the size of sequences in the duplex, the stability of the duplex, and the nucleotide composition of the duplex. However, MirPlex depends on the presence of both strands in a miRNA duplex for prediction, and so cannot discover miRNAs unless the less abundant passenger strand is also detected by the sequencing[11].

MIRPIPE

MIRPIPE identifies miRNAs through sequence homology (Kuenne et al., 2014). It collapses duplicate reads and removes those that have only been sequenced few times. It then further collapses sequences that only differ in the 3′ end and last maps the remaining sequences against known miRBase mature sequences, using the flexible BLAST mapping (Altschul et al., 1990). Since the method relies completely on the presence of known homologs, the prediction accuracy will improve as more miRNAs are deposited to miRBase. However, it cannot identify species-specific miRNAs[12].

miRNA VALIDATION

NORTHERN BLOT ANALYSIS

To resolve if a predicted miRNA is genuine, it is often necessary to validate it with methods other than next-generation sequencing. In this respect, Northern blot analysis can be considered as the gold standard (Lee et al., 1993; Ambros et al., 2003). First, the RNA from the cells or tissues of interest is extracted and run on a high-resolution gel. Then, the gel is treated with probes that are complementary in sequence to the predicted miRNA strand. If the strand is expressed in the cells of interest, a band corresponding to 22 nucleotides will show, and in some cases the precursor, which is around 60 nucleotides, will also show. Although this double-band constitutes compelling evidence of miRNA biogenesis, Northern blot analysis has low sensitivity, so many miRNAs that can be reliably profiled by sequencing is below Northern blot detection limit (**Table 2**).

PCR-BASED METHODS

In contrast, real-time polymerase chain reaction (RT-PCR) methods can profile and thus validate miRNAs of very low abundance. These methods use sequence-specific primers to bind to the miRNAs and amplify them through reverse transcription and polymerase reaction (Lu et al., 2005). The abundances of amplified sequences are measured by fluorescence, and can be used to estimate the expression of the profiled miRNA. Some systems use stem-loop primers that fold around the 3′ end of the miRNA and can only amplify sequences with that particular end, increasing the specificity of the measurements (Chen et al., 2005). Although RT-PCR methods are considered reliable, the custom primers and probes for newly predicted miRNAs can be costly and the methods are rarely used to validate large sets of sequences.

[10] http://www.ncrnalab.dk/#mirdentify/mirdentify.php

[11] http://www.uea.ac.uk/computing/mirplex

[12] https://bioinformatics.mpi-bn.mpg.de

Table 2 | Methods for miRNA validation.

Method	Throughput	Pros	Cons
Northern blot analysis	Low	Length of transcripts observed, possibility of "double-band"	Work-intensive, lack of sensitivity
PCR-based methods	Low	Specific to transcript 3′ end, sensitive	Costly for large-scale validation
Ectopic RNA hairpin expression	Low	miRNA biogenesis is directly tested	Work-intensive, impractical for large-scale validation
Association with Argonaute proteins	Low/high	Directly shows interaction with effector proteins	Method is not always specific for miRNAs
Inhibition of miRNA biogenesis pathways	Low/high	Directly shows dependence on biogenesis proteins	Knock-downs are transient and sometimes weak, generating knock-outs is time-consuming
Experimentally identified target sites	Low/high	Directly demonstrates target interaction or repression	Reporter assays are work-intensive
Conservation and population selection pressure	Sequence analysis	No wet-lab experiments required	Non-conserved miRNAs can be functional

ECTOPIC RNA HAIRPIN EXPRESSION

In some cases, an miRNA is very lowly expressed, but researchers want to know if the miRNA biogenesis machinery would process it, were it highly expressed. It is possible to synthesize the DNA sequence of the candidate hairpin and clone it into a bacterial or viral vector (Chiang et al., 2010). The vector is then transfected into a cell culture, and the hairpin sequence is expressed. If the hairpin is recognized and cleaved by the miRNA biogenesis machinery, the predicted miRNA strand will accumulate in cells, and can then be detected by less sensitive methods, such as Northern blot analysis. A disadvantage of this method is that it is time-consuming, in that just a few miRNAs can be tested in parallel in one experiment.

ASSOCIATION WITH ARGONAUTE PROTEINS

Since miRNAs associate with Argonaute proteins, showing that a predicted miRNA interacts with these proteins constitutes strong evidence of its function. There are now anti-bodies for Argonaute proteins in mammals (Ender et al., 2008), meaning that these proteins can be isolated in immuno-precipitation and their associated sRNAs studied. This profiling was previously done by Northern blot analysis or RT-PCR, but is now often done by next-generation sequencing, allowing transcriptome-wide validation. In some cases, the interaction between protein and RNA is stabilized by crosslinking (Licatalosi et al., 2008; Hafner et al., 2010), and some studies also investigate interaction with other proteins known to interact with miRNAs, such as DGCR8 (Macias et al., 2012). However, immuno-precipitation studies also have caveats as they are often performed in cell lines, which may not have the same complements of miRNAs as the tissues from which the sequences are sometimes predicted. Further, sRNAs other than miRNAs are sometimes immune-precipitated with Argonaute proteins (Ender et al., 2008), and it is not understood if these reflect genuine biological realities, or rare artifacts introduced during the experiment. Thus, the presence of an miRNA candidate in such an experiment does not constitute final evidence that it is genuine.

INHIBITION OF miRNA BIOGENESIS PATHWAYS

It is a hallmark of canonical miRNAs that they depend on the presence of Drosha, Dicer, and DGCR8 for their expression. Thus, if an miRNA candidate is depleted in cells that are void of one or more of these proteins, it constitutes strong evidence that the candidate is genuine. The expression of the proteins can be knocked down through RNA interference, where artificial sRNAs complementary in sequence to the Drosha, Dicer, or DGCR8 mRNAs are introduced into cells (Friedländer et al., 2012, 2014). The sRNAs can bind to the mRNAs and reduce protein output transiently. The genes can also be conditionally knocked out using genetic methods (Babiarz et al., 2008). In this case, Drosha, Dicer, or DGCR8 genes are deleted, leading to a collapse of the miRNA populations. Both with RNA interference and genetic methods, it is possible to use next-generation sequencing to profile miRNA expression transcriptome-wide before and after the loss of the biogenesis pathways. A limitation of the knock-down approach is that effects on the sRNA expression level are often subtle and transient (Friedländer et al., 2012). The genetic knock-outs give much clearer results, but require generation of mutant animals or cells, which is not trivial, even with the advances made with the CRISPR/Cas9 system (Cong et al., 2013; Mali et al., 2013).

EXPERIMENTALLY IDENTIFIED TARGET SITES

Arguably, demonstrating the function of a miRNA constitutes stronger evidence than demonstrating its biogenesis or association with proteins. For this purpose, reporter constructs can be designed that are fusions of a target 3′ UTR and a reporter gene that express a marker such as luciferase (Zeng and Cullen, 2003). If the fluorescence is specifically reduced in the presence of the guide miRNA, this indicates an miRNA–target interaction. These reporter assays can be designed to simulate natural cell conditions, with endogenous miRNA and target levels and a natural number of target sites. While this method is time-consuming and only tests a single miRNA in one experiment, new genomics data can profile miRNA–target interaction transcriptome-wide (Helwak et al.,

2013; Grosswendt et al., 2014). These methods use exogenous or endogenous ligases to crosslink miRNAs and their targets, and subsequently sequence these chimeric sequences, yielding information on miRNA–target pairs. These data have been found to contain novel miRNA candidates linked to mRNA sites that have typical target features (Friedländer et al., 2014).

CONSERVATION AND POPULATION SELECTION PRESSURE

Some miRNAs, like *let-7*, are deeply conserved and retain almost the exact same sequence in all animals with bilateral body types, ranging from nematode to fruit fly to human (Pasquinelli et al., 2000). Thus miRNA validation is transitive: if a validated miRNA is conserved in a new species, it is likely to be genuine. There are numerous criteria for defining if an miRNA is conserved, but some parts are more likely to be under negative selection. Often homologous genome sequences from numerous species are aligned and the conservation studied to see which parts are most conserved. The nucleotides 2–8 in the 5′ end of the miRNA (the "seed") are important for target specificity and are often conserved in evolution (Lai, 2002). In fact, miRNAs are grouping into functional gene families based on their seed sequence. The remaining part of the miRNA guide strand also confers binding specificity (Bartel, 2009) and the passenger strand is important for forming duplex with the guide. Last, the sequences flanking the two miRNA strands often exhibit some conservation, as these regions are important for the hairpin structure, and for recruiting proteins during biogenesis (Han et al., 2006). There are examples of miRNAs that are species-specific, yet have well-defined and important functions (Hu et al., 2012). In these cases, cross-species conservation patterns cannot be used, but intra-species population studies can reveal selection pressures (Friedländer et al., 2014). However, since these selection pressures can be very subtle, large numbers of novel miRNA genes are needed to detect trends, so the population approaches are not applicable to most studies. Further, sequences can to some extent be conserved by chance, so it often does not constitute definite evidence of function.

COMPUTATIONAL BENCHMARKING

Wet-lab experiments include gold standards for demonstrating that a given miRNA candidate is genuine. But computational benchmarking can give some estimates to the performance of methods to predict miRNAs, and can compare strengths and weaknesses of distinct algorithms. An advantage of benchmarking is further that it is easily undertaken by computational research groups, while performing Northern blot analyses, for instance, may require substantial investment of time and funds.

Some of the most widely used measures of prediction performance are sensitivity, specificity, and accuracy (**Table 3**). Sensitivity is the fraction of known distinct miRNAs in the data that are recovered by the method. Specificity is the fraction of (assumed) non-miRNA sequences that are correctly discarded by the algorithm. The false positive rate is the fraction of non-miRNA sequences that are incorrectly reported as miRNAs, or 1 – sensitivity. Accuracy is the fraction of distinct sequences that are correctly classified by the method, summing over all miRNAs and non-miRNAs. Another common measure of prediction performance is the area under curve (AUC) of receiver operating characteristic

Table 3 | Sensitivity, specificity, and accuracy.

		miRNA state	
		Genuine miRNA	**Not genuine miRNA**
miRNA prediction	Positive	True positives (TP)	False positives (FP)
	Negative	False negatives (FN)	True negatives (TN)
Formulas	Sensitivity or true positive rate		TP/(TP + FN)
	Specificity or true negative rate		TN/(FP + TN)
	Accuracy		(TP + TN)/(TP + FP + FN + TN)

(ROC) Curve. The sensitivity is plotted as a function of the false positive rate, showing the trade-off between sensitivity and specificity. The area under the curve indicates performance, with the full area (100%) corresponding to perfect prediction, while half area (50%) corresponding to prediction that is no better than random.

However, the problem of predicting miRNAs from sRNA-seq data is often a skewed one. That is, if tens of thousands of candidate hairpins are being investigated, the number of genuine miRNA precursors is typically in the hundreds. In other words, the number of negatives often vastly outnumbers the positives. Therefore, a modest reduction in sensitivity can often be tolerated, while a modest reduction in specificity can result in an unmanageable number of false positives. For instance, a reduction in sensitivity from 99 to 90% will mean a 9% loss of genuine miRNAs, while a corresponding reduction in specificity will cause a 10-fold increase in false positives, potentially rendering the resulting predictions useless. To address this, true positives and false positives are often reported as absolute numbers, to give a concrete idea of the number of sequences a user of the methods will encounter. Some methods, like miRDeep and miRDeep2, include computational controls to give the user an idea of the number of false positives generated by each run.

Most studies presenting tools to predict miRNA genes include benchmarking of their own method, often comparing it to competitor methods. A summary of these comparisons would be too comprehensive for this review; however, we have listed all the benchmarking in **Table 1**. However, two independent studies have been undertaken to compare the prediction performance of miRNA discovery tools. One study found miRExpress to be the most sensitive method and the mirTools suite (which uses miRDeep for prediction) to be the most accurate method (Li et al., 2012). However, we caution against relying too much on the findings of this study, as the inferred performance of the distinct tools differs widely from other performance comparisons (as referenced in **Table 1**). Another independent study has been undertaken to compare the prediction performance of miRDeep, miRDeep2, and miRanalyzer (updated version), which are some of the most widely used methods in the field (Williamson et al., 2013). One tool, DSAP, which quantifies miRNAs in sRNA-seq was also included in the study, but is not described here as it does not predict new miRNAs. The tools were tested against six biological datasets from cell lines and one simulated negative control data set. miRDeep2 was overall found to have the highest sensitivity, while miRanalyzer reported the most novel miRNA candidates. However, it

also reported miRNAs from the simulated data, suggesting that some of the ones reported from the biological data are false positives. miRDeep had the best overall trade-off between sensitivity and specificity, as measured by AUC, followed by miRDeep2. It should be mentioned that this benchmarking just represents performance in a few use cases, and more independent studies should be undertaken to evaluate the strengths and weaknesses of the existing methods.

VISUAL INSPECTION OF STRUCTURE AND READ SIGNATURE

Many tools for miRNA prediction generate graphics of the novel candidates, showing the RNA structure and the positions of the sequenced RNAs relative to the hairpin. With experience, it is possible to make estimates which of the novel candidate miRNAs can be validated in wet-lab experiments, and which will turn out to be false positive predictions. The human eye is a sensitive tool that can discriminate subtle features that are difficult to score computationally without loss of sensitivity. For instance, the miRNA hairpin structure will rarely contain large bulges, but will also rarely form a tight stem. Also, the processing of miRNA 5′ ends tends to be more precise than processing of the 3′ end (Ruby et al., 2006). Spending some time looking at gold standard known miRNAs can teach a researcher to identify these and more features. Of course, visual inspection of structure and read signature is no substitute for validation, but it can give the trained miRNA researcher an estimate of the quality of his predictions.

FUTURE DIRECTIONS OF THE FIELD

RESOLVING AMBIGUOUS SEQUENCES

Any miRNA prediction depends on read mappings that trace the sequenced RNAs to the genome loci from which they were transcribed. sRNA-seq presents difficulties that are rarely encountered in mRNA sequencing. We know from biology that each deep sequenced RNA has been transcribed from exactly one genome locus. However, when sequenced sRNAs are mapped to the reference genome, many map to more than one locus. This is in some cases because the RNA is transcribed from a gene with many copies in the genome, like a transposable element. In some cases, it will be "spurious" mappings, meaning that a short sequence can have chance matches to biologically unrelated positions in the genome, especially when the reference genome is large. A solution to the problem could be to assume that most deep sequencing reads have originated from a relatively small number of genome loci, and attempt to map the reads such that most of them locate to the fewest possible number of loci. In some concrete cases, this appears reasonable. For instance, imagine a read that maps equally well to two genome loci. One locus is a "read desert" with no other reads mapping nearby. The other locus is an rRNA gene that has thousands of reads mapping. In this case, it would seem reasonable to assume that the read should be mapped to the highly expressed rRNA locus. Some work has already been made toward overcoming these challenges. The tool SeqCluster first fuses reads that overlap in sequence in a tiled way, and subsequently maps the fused sequences to the genome (Pantano et al., 2011). These methods can resolve many, although not all, ambiguous mappings.

CROSS-MAPPING EVENTS

Even though next-generation sequencing quality has improved the last years, some nucleotides are inevitably called incorrectly. Similarly, sRNAs can undergo biological editing events or have untemplated nucleotides added to their 3′ ends. In these cases, an sRNA will no longer map perfectly to the genome position; it was originally transcribed from, but it may map perfectly to a distinct genome position (de Hoon et al., 2010). These wrongly mapped sRNAs will often be considered by miRNA prediction algorithms and may cause false positives. In one study, an explicit statistical model to correct these errors was developed, and numerous wrong mappings were corrected (de Hoon et al., 2010). However, this model has to our knowledge never been implemented as a user-friendly mapping tool. Ideally, such a model could be combined with a method to unambiguously trace sequenced RNAs to a single genome position (above). This would provide the sRNA community with a custom tool to handle some of the difficulties inherent in studying short sequences, and would provide an excellent platform for miRNA prediction.

REPEAT-DERIVED miRNAs

The most commonly used tools for miRNA prediction discards mature sequences that map to many genome loci. This is a practical step to reduce the number of genome loci investigated and thus the number of false positives. However, it is well established that miRNA hairpins can arise from repetitive sequences such as transposable elements (Smalheiser and Torvik, 2005; Berezikov, 2011), and these cannot be detected by current prediction methods, unless the hairpins have diverged in sequence from the consensus repeats. Since repeat-derived sRNAs have been shown to have important functions in, for instance, the mammalian germ line (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Lau et al., 2006; Watanabe et al., 2006, 2008; Tam et al., 2008), it would be interesting to investigate the prevalence and function of repeat-derived miRNAs. However, such a study could be complicated by multi-mapping problems (above) and would be much facilitated by the development of custom mapping and sequence analysis tools. Overall, the field of mapping sRNAs is understudied, and advances in this field could benefit the community.

REDUCING sRNA-seq BIASES

It is well established that library preparation introduces strong biases in sRNA-seq. One study has shown that artificial miRNAs introduced to a buffer in carefully controlled equal abundance give rise to numbers of reads that differ by orders of magnitude (Linsen et al., 2009). This means that some miRNAs give rise to disproportionate large numbers of reads, while others are difficult to detect and thus also more difficult to discover using sequencing. A recent study has traced these biases back to the ligase protein that joins the miRNA with sequencing adapters (Sorefan et al., 2012). miRNAs and adapters together form structures, some of which are easily ligated and some of which are difficult to ligate. In fact, since most sRNA-seq studies use the same ligase and the same adapters (from the Illumina small RNA TruSeq protocol), the miRBase database has been biased toward miRNAs that are easily ligated with this protocol. The researchers of this study has developed an alternative "high definition" protocol using pools of

adapters that even out the biases, giving a more even representation of miRNAs and facilitating identification of novel sequences (Sorefan et al., 2012). As this protocol becomes more widely used in miRNAs discovery efforts, the skew in the miRBase database will, for sure, be corrected.

UNDERSTANDING THE FEATURES THAT DETERMINE HAIRPIN BIOGENESIS

The human transcriptome contains more than 100,000 hairpin structures that resemble miRNA precursors (unpublished results). More than half of these are located in protein coding transcripts. Thus, while many mRNAs and miRNA primary transcripts resemble each other in being capped, poly-adenylated, and containing hairpin structures, the mRNAs are transported to the cytosol and translated, while the pri-miRNAs are cleaved into regulatory sRNAs. This mystery underlines our incomplete understanding of miRNA biogenesis: which features determine if a given hairpin is cleaved into miRNAs or left untouched? Does the presence of protein factors protect the hairpin or make it available for Drosha processing? Or does protein competition determine the hairpin fate? And which structural and sequence features of the hairpin determine which proteins are bound? Studies are unraveling these interactions (Auyeung et al., 2013) but it is clear that our understanding is still incomplete. If we would understand what hairpin features license biogenesis, we would be able to computationally predict from genome sequence, which hairpins are cleaved to miRNAs and which are left untouched.

ACKNOWLEDGMENTS

Wenjing Kang and Marc Riemer Friedländer acknowledge funding from the Strategic Research Area program of the Swedish Research Council through Stockholm University.

REFERENCES

Abelson, J. F., Kwan, K. Y., O'Roak, B. J., Baek, D. Y., Stillman, A. A., Morgan, T. M., et al. (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* 310, 317–320. doi:10.1126/science.1116502

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

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., et al. (2003). A uniform system for microRNA annotation. *RNA* 9, 277–279. doi:10.1261/rna.2183803

An, J., Lai, J., Lehman, M. L., and Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res.* 41, 727–737. doi:10.1093/nar/gks1187

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442, 203–207. doi:10.1038/nature04916

Auyeung, V. C., Ulitsky, I., Mcgeary, S. E., and Bartel, D. P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* 152, 844–858. doi:10.1016/j.cell.2013.01.031

Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small RNAs. *Genes Dev.* 22, 2773–2785. doi:10.1101/gad.1705308

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

Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. *FEBS Lett.* 579, 5904–5910. doi:10.1016/j.febslet.2005.09.040

Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.* 12, 846–860. doi:10.1038/nrg3079

Berezikov, E., Chung, W. J., Willis, J., Cuppen, E., and Lai, E. C. (2007). Mammalian mirtron genes. *Mol. Cell* 28, 328–336. doi:10.1016/j.molcel.2007.09.028

Berezikov, E., Guryev, V., Van De Belt, J., Wienholds, E., Plasterk, R. H., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120, 21–24. doi:10.1016/j.cell.2004.12.031

Berninger, P., Gaidatzis, D., Van Nimwegen, E., and Zavolan, M. (2008). Computational analysis of small RNA cloning data. *Methods* 44, 13–21. doi:10.1016/j.ymeth.2007.10.002

Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409, 363–366. doi:10.1038/35053110

Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., et al. (2003). Dicer is essential for mouse development. *Nat. Genet.* 35, 215–217. doi:10.1038/ng1253

Bushati, N., and Cohen, S. M. (2007). microRNA functions. *Annu. Rev. Cell Dev. Biol.* 23, 175–205. doi:10.1146/annurev.cellbio.23.090506.123406

Calabrese, J. M., Seila, A. C., Yeo, G. W., and Sharp, P. A. (2007). RNA sequence analysis defines dicer's role in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18097–18102. doi:10.1073/pnas.0709193104

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., et al. (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res.* 33, e179. doi:10.1093/nar/gni178

Chen, C. J., and Heard, E. (2013). Small RNAs derived from structural non-coding RNAs. *Methods* 63, 76–84. doi:10.1016/j.ymeth.2013.05.001

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.* 24, 992–1009. doi:10.1101/gad.1884710

Christensen, B. C., Moyer, B. J., Avissar, M., Ouellet, L. G., Plaza, S. L., Mcclean, M. D., et al. (2009). A let-7 microRNA-binding site polymorphism in the KRAS 3′ UTR is associated with reduced survival in oral cancers. *Carcinogenesis* 30, 1003–1007. doi:10.1093/carcin/bgp099

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., et al. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13944–13949. doi:10.1073/pnas.0506654102

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823. doi:10.1126/science.1231143

Davalos, V., Moutinho, C., Villanueva, A., Boque, R., Silva, P., Carneiro, F., et al. (2012). Dynamic epigenetic regulation of the microRNA-200 family mediates epithelial and mesenchymal transitions in human tumorigenesis. *Oncogene* 31, 2062–2074. doi:10.1038/onc.2011.383

de Hoon, M. J., Taft, R. J., Hashimoto, T., Kanamori-Katayama, M., Kawaji, H., Kawano, M., et al. (2010). Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res.* 20, 257–264. doi:10.1101/gr.095273.109

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., and Hannon, G. J. (2004). Processing of primary microRNAs by the microprocessor complex. *Nature* 432, 231–235. doi:10.1038/nature03049

Dezulian, T., Remmert, M., Palatnik, J. F., Weigel, D., and Huson, D. H. (2006). Identification of plant microRNA homologs. *Bioinformatics* 22, 359–360. doi:10.1093/bioinformatics/bti802

Donaszi-Ivanov, A., Mohorianu, I., Dalmay, T., and Powell, P. P. (2013). Small RNA analysis in Sindbis virus infected human HEK293 cells. *PLoS One* 8:e84070. doi:10.1371/journal.pone.0084070

Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W., et al. (2008). A human snoRNA with microRNA-like functions. *Mol. Cell* 32, 519–528. doi:10.1016/j.molcel.2008.10.017

Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874. doi:10.1038/nrg3074

Etebari, K., and Asgari, S. (2014). Accuracy of microRNA discovery pipelines in non-model organisms using closely related species genomes. *PLoS One* 9:e84747. doi:10.1371/journal.pone.0084747

Farazi, T. A., Juranek, S. A., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* 135, 1201–1214. doi:10.1242/dev.005629

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 9, 102–114. doi:10.1038/nrg2290

Friedländer, M. R., Adamidi, C., Han, T., Lebedeva, S., Isenbarger, T. A., Hirst, M., et al. (2009). High-resolution profiling and discovery of planarian small RNAs. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11546–11551. doi:10.1073/pnas.0905222106

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26, 407–415. doi:10.1038/nbt1394

Friedländer, M. R., Lizano, E., Houben, A. J., Bezdan, D., Banez-Coronel, M., Kudla, G., et al. (2014). Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.* 15, R57. doi:10.1186/gb-2014-15-4-r57

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40, 37–52. doi:10.1093/nar/gkr688

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 92–105. doi:10.1101/gr.082701.108

Ghildiyal, M., and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* 10, 94–108. doi:10.1038/nrg2504

Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., et al. (2005). microRNAs regulate brain morphogenesis in zebrafish. *Science* 308, 833–838. doi:10.1126/science.1109020

Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442, 199–202. doi:10.1038/nature04917

Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., et al. (2004). The microprocessor complex mediates the genesis of microRNAs. *Nature* 432, 235–240. doi:10.1038/nature03120

Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* 20, 1709–1714. doi:10.1101/gad.1434406

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., et al. (2014). Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol. Cell* 54, 1042–1054. doi:10.1016/j.molcel.2014.03.049

Guo, L., and Lu, Z. (2010). The fate of miRNA* strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule? *PLoS One* 5:e11387. doi:10.1371/journal.pone.0011387

Ha, M., and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* 15, 509–524. doi:10.1038/nrm3838

Hackenberg, M., Rodriguez-Ezpeleta, N., and Aransay, A. M. (2011). miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.* 39, W132–W138. doi:10.1093/nar/gkr247

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M., and Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* 37, W68–W76. doi:10.1093/nar/gkp347

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141. doi:10.1016/j.cell.2010.03.009

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* 18, 3016–3027. doi:10.1101/gad.1262504

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., et al. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* 125, 887–901. doi:10.1016/j.cell.2006.03.043

Hansen, T. B., Veno, M. T., Kjems, J., and Damgaard, C. K. (2014). miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs. *Nucleic Acids Res.* 42, e124. doi:10.1093/nar/gku598

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., et al. (2005). A microRNA polycistron as a potential human oncogene. *Nature* 435, 828–833. doi:10.1038/nature03552

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi:10.1016/j.cell.2013.03.043

Hendrix, D., Levine, M., and Shi, W. (2010). miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.* 11, R39. doi:10.1186/gb-2010-11-4-r39

Hill, D. A., Ivanovich, J., Priest, J. R., Gurnett, C. A., Dehner, L. P., Desruisseau, D., et al. (2009). DICER1 mutations in familial pleuropulmonary blastoma. *Science* 325, 965. doi:10.1126/science.1174334

Hu, H. Y., He, L., Fominykh, K., Yan, Z., Guo, S., Zhang, X., et al. (2012). Evolution of the human-specific microRNA miR-941. *Nat. Commun.* 3, 1145. doi:10.1038/ncomms2146

Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev. Genet.* 12, 99–110. doi:10.1038/nrg2936

Hutvagner, G., Mclachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T., and Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme dicer in the maturation of the let-7 small temporal RNA. *Science* 293, 834–838. doi:10.1126/science.1062961

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35, W339–W344. doi:10.1093/nar/gkm368

Johnston, R. J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* 426, 845–849. doi:10.1038/nature02255

Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., and Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* 15, 2654–2659. doi:10.1101/gad.927801

Kloosterman, W. P., and Plasterk, R. H. (2006). The diverse functions of microRNAs in animal development and disease. *Dev. Cell* 11, 441–450. doi:10.1016/j.devcel.2006.09.009

Knight, S. W., and Bass, B. L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science* 293, 2269–2271. doi:10.1126/science.1062039

Kohli, D., Joshi, G., Deokar, A. A., Bhardwaj, A. R., Agarwal, M., Katiyar-Agarwal, S., et al. (2014). Identification and characterization of wilt and salt stress-responsive microRNAs in chickpea through high-throughput sequencing. *PLoS One* 9:e108851. doi:10.1371/journal.pone.0108851

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152–D157. doi:10.1093/nar/gkq1027

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, D68–D73. doi:10.1093/nar/gkt1181

Kuenne, C., Preussner, J., Herzog, M., Braun, T., and Looso, M. (2014). MIR-PIPE: quantification of microRNAs in niche model organisms. *Bioinformatics* 30, 3412–3413. doi:10.1093/bioinformatics/btu573

Ladewig, E., Okamura, K., Flynt, A. S., Westholm, J. O., and Lai, E. C. (2012). Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res.* 22, 1634–1645. doi:10.1101/gr.133553.111

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* 294, 853–858. doi:10.1126/science.1064921

Lai, E. C. (2002). microRNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* 30, 363–364. doi:10.1038/ng865

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. (2003). Computational identification of *Drosophila* microRNA genes. *Genome Biol.* 4, R42. doi:10.1186/gb-2003-4-7-r42

Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and its *D. melanogaster* homolog are required for miRNA biogenesis. *Curr. Biol.* 14, 2162–2167. doi:10.1016/j.cub.2004.11.001

Langenberger, D., Pundhir, S., Ekstrom, C. T., Stadler, P. F., Hoffmann, S., and Gorodkin, J. (2012). deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics* 28, 17–24. doi:10.1093/bioinformatics/btr598

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi:10.1186/gb-2009-10-3-r25

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294, 858–862. doi:10.1126/science.1065062

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., et al. (2006). Characterization of the piRNA complex from rat testes. *Science* 313, 363–367. doi:10.1126/science.1130164

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294, 862–864. doi:10.1126/science.1065329

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, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q., and Shen, B. (2012). Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Res.* 40, 4298–4305. doi:10.1093/nar/gks043

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469. doi:10.1038/nature07488

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., et al. (2003). The microRNAs of *Caenorhabditis elegans. Genes Dev.* 17, 991–1008. doi:10.1101/gad.1074403

Linsen, S. E., De Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R. K., et al. (2009). Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods* 6, 474–476. doi:10.1038/nmeth0709-474

Lu, D. P., Read, R. L., Humphreys, D. T., Battah, F. M., Martin, D. I., and Rasko, J. E. (2005). PCR-based expression analysis and identification of microRNAs. *J. RNAi Gene Silencing* 1, 44–49.

Macias, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E., and Caceres, J. F. (2012). DGCR8 HITS-CLIP reveals novel functions for the microprocessor. *Nat. Struct. Mol. Biol.* 19, 760–766. doi:10.1038/nsmb.2344

Mackowiak, S. D. (2011). Identification of novel and known miRNAs in deep-sequencing data with miRDeep2. *Curr. Protoc. Bioinformatics* 12, 10. doi:10.1002/0471250953.bi1210s36

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., et al. (2013). RNA-guided human genome engineering via Cas9. *Science* 339, 823–826. doi:10.1126/science.1232033

Mapleson, D., Moxon, S., Dalmay, T., and Moulton, V. (2013). MirPlex: a tool for identifying miRNAs in high-throughput sRNA datasets without a genome. *J. Exp. Zool. B Mol. Dev. Evol.* 320, 47–56. doi:10.1002/jez.b.22483

Mathelier, A., and Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 26, 2226–2234. doi:10.1093/bioinformatics/btq329

Mayoral, J. G., Etebari, K., Hussain, M., Khromykh, A. A., and Asgari, S. (2014). Wolbachia infection modifies the profile, shuttling and structure of microRNAs in a mosquito cell line. *PLoS One* 9:e96107. doi:10.1371/journal.pone.0096107

Medina, P. P., and Slack, F. J. (2008). microRNAs and cancer: an overview. *Cell Cycle* 7, 2485–2492. doi:10.4161/cc.7.16.6453

Mencia, A., Modamio-Hoybjor, S., Redshaw, N., Morin, M., Mayo-Merino, F., Olavarrieta, L., et al. (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat. Genet.* 41, 609–613. doi:10.1038/ng.355

Metpally, R. P., Nasser, S., Malenica, I., Courtright, A., Carlson, E., Ghaffari, L., et al. (2013). Comparison of analysis tools for miRNA high throughput sequencing using nerve crush as a model. *Front. Genet.* 4:20. doi:10.3389/fgene.2013.00020

Mohorianu, I., Schwach, F., Jing, R., Lopez-Gomollon, S., Moxon, S., Szittya, G., et al. (2011). Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *Plant J.* 67, 232–246. doi:10.1111/j.1365-313X.2011.04586.x

Morita, S., Horii, T., Kimura, M., Goto, Y., Ochiya, T., and Hatada, I. (2007). One Argonaute family member, Eif2c2 (Ago2), is essential for development and appears not to be involved in DNA methylation. *Genomics* 89, 687–696. doi:10.1016/j.ygeno.2007.01.004

Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D. J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 24, 2252–2253. doi:10.1093/bioinformatics/btn428

Nam, J. W., Shin, K. R., Han, J., Lee, Y., Kim, V. N., and Zhang, B. T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* 33, 3570–3581. doi:10.1093/nar/gki668

Ohler, U., Yekta, S., Lim, L. P., Bartel, D. P., and Burge, C. B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* 10, 1309–1322. doi:10.1261/rna.5206304

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., and Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila. Cell* 130, 89–100. doi:10.1016/j.cell.2007.06.028

Okamura, K., Phillips, M. D., Tyler, D. M., Duan, H., Chou, Y. T., and Lai, E. C. (2008). The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat. Struct. Mol. Biol.* 15, 354–363. doi:10.1038/nsmb.1409

Pandey, R., Joshi, G., Bhardwaj, A. R., Agarwal, M., and Katiyar-Agarwal, S. (2014). A comprehensive genome-wide study on tissue-specific and abiotic stress-specific miRNAs in *Triticum aestivum. PLoS One* 9:e95800. doi:10.1371/journal.pone.0095800

Pantaleo, V., Szittya, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T., et al. (2010). Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. *Plant J.* 62, 960–976. doi:10.1111/j.0960-7412.2010.04208.x

Pantano, L., Estivill, X., and Marti, E. (2011). A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics* 27, 3202–3203. doi:10.1093/bioinformatics/btr527

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408, 86–89. doi:10.1038/35040556

Pundhir, S., and Gorodkin, J. (2013). microRNA discovery by similarity search to a database of RNA-seq profiles. *Front. Genet.* 4:133. doi:10.3389/fgene.2013.00133

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., et al. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans. Cell* 127, 1193–1207. doi:10.1016/j.cell.2006.10.040

Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature* 448, 83–86. doi:10.1038/nature05983

Sheng, Y., Engstrom, P. G., and Lenhard, B. (2007). Mammalian microRNA prediction through a support vector machine model of sequence and structure. *PLoS One* 2:e946. doi:10.1371/journal.pone.0000946

Shi, W., Hendrix, D., Levine, M., and Haley, B. (2009). A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.* 16, 183–189. doi:10.1038/nsmb.1536

Smalheiser, N. R., and Torvik, V. I. (2005). Mammalian microRNAs derived from genomic repeats. *Trends Genet.* 21, 322–326. doi:10.1016/j.tig.2005.04.008

Smith, L. K., Tandon, A., Shah, R. R., Mav, D., Scoltock, A. B., and Cidlowski, J. A. (2013). Deep sequencing identification of novel glucocorticoid-responsive miRNAs in apoptotic primary lymphocytes. *PLoS ONE* 8:e78316. doi:10.1371/journal.pone.0078316

Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., et al. (2012). Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* 3, 4. doi:10.1186/1758-907X-3-4

Stefani, G., and Slack, F. J. (2008). Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.* 9, 219–230. doi:10.1038/nrm2347

Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., et al. (2012). The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 28, 2059–2061. doi:10.1093/bioinformatics/bts311

Surridge, A. K., Lopez-Gomollon, S., Moxon, S., Maroja, L. S., Rathjen, T., Nadeau, N. J., et al. (2011). Characterisation and expression of microRNAs in developing wings of the neotropical butterfly *Heliconius melpomene. BMC Genomics* 12:62. doi:10.1186/1471-2164-12-62

Szittya, G., Moxon, S., Santos, D. M., Jing, R., Fevereiro, M. P., Moulton, V., et al. (2008). High-throughput sequencing of *Medicago truncatula* short RNAs identifies eight new miRNA families. *BMC Genomics* 9:593. doi:10.1186/1471-2164-9-593

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. *J. Pathol.* 220, 126–139. doi:10.1002/path.2638

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453, 534–538. doi:10.1038/nature06904

Wang, W. C., Lin, F. M., Chang, W. C., Lin, K. Y., Huang, H. D., and Lin, N. S. (2009). miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* 10:328. doi:10.1186/1471-2105-10-328

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., et al. (2005). microRNA identification based on sequence and structure alignment. *Bioinformatics* 21, 3610–3614. doi:10.1093/bioinformatics/bti562

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat. Genet.* 39, 380–385. doi:10.1038/ng1969

Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., et al. (2006). Identification and characterization of two novel classes of small RNAs in the

mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.* 20, 1732–1743. doi:10.1101/gad.1425706

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

Wen, M., Shen, Y., Shi, S., and Tang, T. (2012). miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. *BMC Bioinformatics* 13:140. doi:10.1186/1471-2105-13-140

Williamson, V., Kim, A., Xie, B., Mcmichael, G. O., Gao, Y., and Vladimirov, V. (2013). Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Brief. Bioinformatics* 14, 36–45. doi:10.1093/bib/bbs010

Wu, J., Liu, Q., Wang, X., Zheng, J., Wang, T., You, M., et al. (2013). mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol.* 10, 1087–1092. doi:10.4161/rna.25193

Yang, J. H., and Qu, L. H. (2012). DeepBase: annotation and discovery of microRNAs and other noncoding RNAs from deep-sequencing data. *Methods Mol. Biol.* 822, 233–248. doi:10.1007/978-1-61779-427-8_16

Yang, J. S., Phillips, M. D., Betel, D., Mu, P., Ventura, A., Siepel, A. C., et al. (2011). Widespread regulatory activity of vertebrate microRNA* species. *RNA* 17, 312–326. doi:10.1261/rna.2537911

Yang, X., and Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 27, 2614–2615. doi:10.1093/bioinformatics/btr430

Zeng, Y., and Cullen, B. R. (2003). Sequence requirements for micro RNA processing and function in human cells. *RNA* 9, 112–123. doi:10.1261/rna.2780503

Zhang, B. H., Pan, X. P., Cox, S. B., Cobb, G. P., and Anderson, T. A. (2006a). Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.* 63, 246–254. doi:10.1007/s00018-005-5467-7

Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M. S., Giannakakis, A., et al. (2006b). microRNAs exhibit high frequency genomic alterations in human cancer. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9136–9141. doi:10.1073/pnas.0508889103

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.

Received: 15 October 2014; accepted: 07 January 2015; published online: 26 January 2015.

*Citation: Kang W and Friedländer MR (2015) Computational prediction of miRNA genes from small RNA sequencing data. Front. Bioeng. Biotechnol. **3**:7. doi: 10.3389/fbioe.2015.00007*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

MINI REVIEW ARTICLE
published: 25 March 2015
doi: 10.3389/fbioe.2015.00037

A-to-I RNA editing: current knowledge sources and computational approaches with special emphasis on non-coding RNA molecules

Giovanni Nigita[1]*, Dario Veneziano[2] and Alfredo Ferro[2]

[1] *Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, Columbus, OH, USA*
[2] *Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy*

Edited by:
Christian M. Zmasek, Sanford-Burnham Medical Research Institute, USA

Reviewed by:
Subrata H. Mishra, Johns Hopkins University School of Medicine, USA
Yingqun Huang, Yale University School of Medicine, USA

***Correspondence:**
Giovanni Nigita, Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, 460 W 12th Avenue, Columbus, OH 43210, USA
e-mail: gianni.nigita@gmail.com

RNA editing is a dynamic mechanism for gene regulation attained through the alteration of the sequence of primary RNA transcripts. A-to-I (adenosine-to-inosine) RNA editing, which is catalyzed by members of the adenosine deaminase acting on RNA (ADAR) family of enzymes, is the most common post-transcriptional modification in humans. The ADARs bind double-stranded regions and deaminate adenosine (A) into inosine (I), which in turn is interpreted by the translation and splicing machineries as guanosine (G). In recent years, this modification has been discovered to occur not only in coding RNAs but also in non-coding RNAs (ncRNA), such as microRNAs, small interfering RNAs, transfer RNAs, and long non-coding RNAs. This may have several consequences, such as the creation or disruption of microRNA/mRNA binding sites, and thus affect the biogenesis, stability, and target recognition properties of ncRNAs. The malfunction of the editing machinery is not surprisingly associated with various human diseases, such as neurodegenerative, cardiovascular, and carcinogenic diseases. Despite the enormous efforts made so far, the real biological function of this phenomenon, as well as the features of the ADAR substrate, in particular in non-coding RNAs, has still not been fully understood. In this work, we focus on the current knowledge of RNA editing on ncRNA molecules and provide a few examples of computational approaches to elucidate its biological function.

Keywords: A-to-I RNA editing, ncRNA, microRNA, RNA-seq, ADARs, HTS

BACKGROUND

While in the past researchers mainly focused on DNA mutations in order to further elucidate molecular pathways involved in numerous cancers, in the last decade focus has shifted to the analysis of post-transcriptional modification events, such as RNA editing. Concurrently, it has been estimated that only 1% of mammalian genome codes for protein, while the vast majority of the transcriptome is composed of non-coding RNAs crucially involved in gene expression pathways, such as transcription, translation, and gene regulation (Cech and Steitz, 2014). The editing machinery, occurring both in coding and non-coding RNAs, has been implicated in various human diseases (Galeano et al., 2012; Tomaselli et al., 2014). Strong interest is thus growing toward understanding how and why RNA editing can influence non-coding RNA function.

RNA editing is a type of post-transcriptional modification that takes place in eukaryotes. Several forms of RNA editing have been discovered, but nowadays A-to-I RNA editing is considered the predominant one in mammals (Nishikura, 2010). Adenosine (A) deamination produces its conversion into inosine (I), which in turn is interpreted as guanosine (G) by both the translation and splicing machineries (Rueter et al., 1999). Enzymes members of the adenosine deaminase acting on RNA (ADAR) family catalyze this biological phenomenon which occurs only on dsRNA structures (Bass, 2002; Jepson and Reenan, 2008; Nishikura, 2010). Double-stranded RNAs are imperfect duplexes formed by base-pairing between residues in the region proximate to the editing site (usually overlapping a neighboring intron) and the exonic sequence containing the A. Such proximate region is termed *editing complementary sequence* (ECS), potentially located several hundred to several thousand nucleotides upstream or downstream of the edited A. This requires experimental validation and represents one critical issue with the detection of editing sites.

Three members of the ADAR gene family can be distinguished in humans, in particular, two isoforms of ADAR1 (ADAR1p150 and ADAR1p110) (Kim et al., 1994), ADAR2 (Lai et al., 1997), and ADAR3 (Chen et al., 2000). While ADAR1 and ADAR2 are widely expressed in tissues, ADAR3 is limited to brain tissues (Melcher et al., 1996). Interestingly, unlike ADAR1 and ADAR2, ADAR3 possesses a catalytically inactive (Chen et al., 2000) arginine-rich R domain, which allows the enzyme to bind single strand structures.

An RNA edited site neighborhood profiling was established for ADAR1-2. While for ADAR1, no 3′ neighbor preference has been identified, a 5′ nearest neighboring preference consisting of $U = A > C > G$ (Polson and Bass, 1994) can be observed. Like ADAR1, ADAR2 has a similar 5′ nearest neighboring preference ($U \approx A > C = G$) but, furthermore, it has a 3′ nearest neighboring preference ($U = G > C = A$) as well, creating a particular trinucleotide sequence with the adenosine at the center (U*A*U, A*A*G, U*A*G, A*A*U) (Lehmann and Bass, 2000). In addition, the ADARs

show selectivity based on both dsRNA length and the presence of mismatches, loops, and bulges that interrupt the base-pairing (Bass, 1997).

There are two kinds of A-to-I RNA editing: *specific* A-to-I editing occurs in short duplex regions interrupted by bulges and mismatches (Wahlstedt and Ohman, 2011); the *promiscuous* one occurs within longer stable duplexes of hundreds of nucleotides, mostly formed by repetitive elements, such as Alus, in which up to 50% of adenosines could be targeted by ADARs (Carmi et al., 2011; Bazak et al., 2014b).

Adenosine-to-inosine RNA editing has been discovered both in intronic and exonic regions, 5′ and 3′-UTRs as well. RNA editing events can take place in several cellular contexts: in the gene expression pathway (Bazak et al., 2014b), such as in translation (Nishikura, 2010) or in the creation and/or destruction of splicing sites (Rueter et al., 1999); during gene regulation through editing events in microRNA/mRNA binding regions (Nishikura, 2006; Borchert et al., 2009). Recent reports affirmed that RNA editing may occur in non-coding RNA molecules, particularly within precursor-tRNA (Su and Randau, 2011), pri-miRNA (Kawahara et al., 2008; Kawahara, 2012), and lncRNA (Mitra et al., 2012). It was estimated that 10–20% of miRNAs undergo A-to-I editing (Blow et al., 2006; Kawahara et al., 2008) at the pri-miRNA level (Yang et al., 2006). Editing can influence both the maturation process (Yang et al., 2006) and the recognition of binding sites on target mRNAs (Kawahara et al., 2007; Wu et al., 2011). Indeed, a single editing site in a miRNA seed region could drastically change its set of targets (Alon et al., 2012).

In the past decade, surprising results have been obtained in RNA editing site discovery, thanks initially to the application of bioinformatic approaches, subsequently fully replaced by RNAseq-based methods in recent years. The large amount of editing sites discovered by these methodologies has led to the creation of public databases (Kiran and Baranov, 2010; Kiran et al., 2013; Ramaswami and Li, 2014). As described below, all these resources containing very important information, such as editing level and genomic annotations, can help to functionally elucidate the RNA editing phenomenon.

This mini review summarizes both the current knowledge on RNA editing, as well as past and present approaches for discovery and analysis of editing sites, particularly emphasizing on RNA editing in non-coding RNA (ncRNA) molecules.

COMPUTATIONAL APPROACHES TO DISCOVER AND ANALYZE RNA EDITING EVENTS

THE ORIGINS OF THE ANALYSIS AND DETECTION OF RNA EDITING SITES – COMPUTATIONAL AND BIOCHEMICAL METHODS

In the early 2000s, the ADAR enzyme family was observed to play an important role during embryonic development (Higuchi et al., 2000; Wang et al., 2000), while also associating the alteration of the editing machinery to neurological diseases (Maas et al., 2001; Kawahara et al., 2004). At that time, only few RNA editing sites were discovered (Morse and Bass, 1999). Hoopengardner et al. (2003) using comparative genomics identified and experimentally validated 16 novel editing sites in fruit fly and one in human. Interestingly, they discovered that these editing sites are surrounded by highly conserved exonic regions which form a dsRNA structure as required for ADARs. Despite these efforts, most editing sites were detected by chance.

In 2004, unprecedented computational methods were designed in order to discover clustered A-to-I RNA editing sites in Alu repeats of the human transcriptome (Athanasiadis et al., 2004; Kim et al., 2004; Levanon et al., 2004), going from dozens to tens of thousands of editing sites. By aligning millions of publicly expressed sequence tags (EST) (Boguski et al., 1993) against a reference genome, it is indeed possible to identify A-to-G mismatches as putative candidates of A-to-I editing events. Unfortunately, without considering RNA editing, related features such as nearest neighbor preference sequence, this naïve approach produces a large amount of false positives due to sequencing errors originating from poor sequencing quality, somatic mutations, or single nucleotide polymorphisms (SNP). All of the above methods avoided this issue by taking into account cDNA-genome alignments along with clusters of mismatches in long and stable dsRNA structures and, finally, filtered known SNPs from the obtained candidates, reaching good accuracy.

A more quantitative and accurate analysis was later provided by Eggington et al. (2011)[1], who predicted editing sites in dsRNAs by assuming a multiplicative relationship between the coefficients (estimated by a non-linear regression model and dependent on the bases neighboring each site) used to determine the percentage of editing sites.

The bioinformatics methods for RNA editing detection comparing a cDNA sequence with a reference genome nevertheless present a significant problem: they are not able to distinguish a guanosine originating from an I-to-G replacement, from a guanosine as a product of noise, sequencing errors, or SNP. To overcome this limit, Sakurai et al. (2010) designed a biochemical method, called inosine chemical erasing (ICE), for the identification of inosine sites on RNA molecules by employing inosine-specific cyanoethylation with reverse transcription, PCR amplification, and direct sequencing. Without requiring changing profiles of cellular gene expression nor genomic DNA for reference, this method accurately and consistently identifies inosines in RNA strands. Recently, Sakurai et al. (2014) combined the ICE method with deep sequencing technology (ICE-seq) for an unbiased genome wide screening of novel A-to-I editing sites.

NEW *ERA* OF RNA EDITING DISCOVERY – HIGH-THROUGHPUT SEQUENCING APPROACHES

Despite the substantial results achieved with the approaches described above, some restrictions due to sequencing limitations remained. Before 2009, in fact, only a few dozen editing sites had been detected outside repetitive regions in humans due to the impossibility of designing a systematic method to discover editing events in ncRNA genes.

With the advent of high-throughput sequencing technology (HTS), things radically improved. In 2009, Li et al. (2009) developed the first HTS-based application which, through massively parallel target capture and DNA sequencing, identified 36,000 non-repetetive putative A-to-I editing events. Recently, several

[1]http://www.biochem.utah.edu/bass/index.html

HTS-based approaches for editing discovery have been developed (see **Table 1**). It was latterly hypothesized that there are more than 100 million editing sites in human Alu repeats, located mainly in genic regions (Bazak et al., 2014a). Despite the increased accuracy, these methods have limitations in terms of false positives produced (Kleinman and Majewski, 2012; Lin et al., 2012; Pickrell et al., 2012).

Table 1 depicts some of the most important studies on RNA editing detection by HTS. The majority was designed to identify RNA editing events in protein-coding RNA, while a few also focus on lncRNAs as well. In 2010, de Hoon et al. developed a strategy to correct cross-mapping of small RNA deep-sequencing libraries, applying it to analyze RNA editing in human mature miRNAs. They concluded that miRNA editing is rare in animals

Table 1 | Deep sequencing based approaches.

Focus	Year	# Editing sites (ES) discovered	Description	Reference
mRNAs	2009	239 A-to-I ES	Parallel target capturing and DNA sequencing	Li et al. (2009)
miRNAs	2010	10 (three A-to-I and two C-to-U)	Strategy to correct for cross-mapping in short RNA sequencing libraries	de Hoon et al. (2010)
mRNAs	2011	1,809 (1,096 A-to-I and 11 C-to-U)	Massively parallel DNA and RNA sequencing of 18 Korean individuals	Ju et al. (2011)
mRNAs	2012	9,636 (5,965 A-to-I)	Accurate mapping approach to distinguish single-nucleotide differences in one set of RNA-seq data	Bahn et al. (2012)
Coding, non-coding and small RNA genes	2012	22,588 (21,113 A-to-I)	Computational pipeline to identify RNA editing sites from genome and whole-transcriptome data of the same individual	Peng et al. (2012)
Alu and non-Alu regions	2012	150,865 (144,406 A-to-I) from GM12878 457,078 (423,377 A-to-I) from (Peng et al., 2012) data	Framework to robustly identify RNA editing sites using transcriptome and genome deep-sequencing data from the same individual	Ramaswami et al. (2013)
mRNAs	2012	61 A-to-I ES	Computational strategy based on two-step mapping procedure with only RNA-seq and without *a priori* RNA editing information	Picardi et al. (2012)
mRNAs	2012	5695 (5349 A-to-I)	A rigorous computational pipeline to identify RNA editing site in human polyA$^+$ ENCODE RNA-seq data from 14 cell types.	Park et al. (2012)
miRNAs	2012–2013	19 A-to-I ES	Protocol for the identification of RNA editing sites in mature miRNAs using deep sequencing data.	Alon et al. (2012) and Alon and Eisenberg (2013)
mRNAs	2013	>1 million of A-to-I ES in other human LCL and several tissues	Two methods (*separate* and *pooled* sample methods) to accurately identify RNA editing events by using RNA-seq data from multiple samples in a single species	Ramaswami et al. (2013)
mRNAs	2013	2,245 A-to-I ES	A strategy to accurately predict consecutive RNA editing events from human RNA-seq data in the absence of relevant genomic sequences	Zhu et al. (2013)
mRNAs	2013	223,490 A-to-I ES from (Ramaswami et al., 2013) data	Suite of python scripts to investigate RNA editing by using RNA-seq data	Picardi and Pesole (2013)
Alu elements	2014	1,586,270 A-to-I ES	Detection approach to analysis *Alu* editing by using large-scale RNA-seq data	Bazak et al. (2014a)
mRNAs	2014	29,843 A-to-I ES	Unbiased genome-wide screening of A-to-I editing events using the ICE-method combined with deep sequencing (ICE-seq)	Sakurai et al. (2014)
mRNAs	2014	455,014 A-to-I ES	Computational method to detect hyper-edited reads in RNA-seq data	Porath et al. (2014)

Some of the most important deep sequencing based approaches, developed in the last 5 years, to identify RNA editing sites in humans.

and addressed methodological problems in its analysis through RNAseq. Subsequently, Alon et al. (2012) systematically identified known editing events in mature miRNAs of human brain, in addition to 17 novel ones, 12 of which occur in the seed region (Alon and Eisenberg, 2013). They moreover identified sequence preference in the residues, both flanking and opposing the A-to-I editing site. As the authors suggested, this pipeline could identify editing sites in miRNAs from NGS data of different experimental set-ups. Currently, Alon's method is the only one able to accurately detect and quantify A-to-I RNA editing events in mature miRNAs by NGS. Together with the latest pipeline published by Picardi et al. (2014) for RNA editing detection in human lncRNAs from deep sequencing experiments.

CURRENT KNOWLEDGE OF RNA EDITING ON ncRNA MOLECULES

BIOLOGICAL DATABASES: DARNED AND RADAR

The birth of the first computational methods for the identification of RNA editing events (Athanasiadis et al., 2004; Kim et al., 2004; Levanon et al., 2004) caused a growing interest in the scientific community for RNA editing, as there was a strong need to collect in a centralized repository the tens of thousands of editing events discovered up to that point. For this reason, Kiran and Baranov designed DARNED[2] (DAtabase of RNa EDiting), the first public database of known editing sites in human (Kiran and Baranov, 2010). The first release of DARNED contained more than 40,000 predicted human editing sites, of which a few were experimentally validated (Ramaswami et al., 2013). The usefulness of the repository rests in the ability to retrieve information on RNA regions where editing events can occur, such as genome coordinates, cell/tissue/organ sources, and the number of ESTs supporting referenced and edited bases. According to the first release of DARNED, Laganà et al. (2012) built miR-EdiTar[3], a database of predicted miRNA binding sites that could be affected by A-to-I editing sites occurring in 3′UTRs.

In subsequent years, the advent of high-throughput RNA sequencing (RNAseq) and biochemically-based (Sakurai et al., 2010) techniques progressively led to the development of increasingly accurate transcriptome-wide methods for RNA editing detection. Furthermore, deep sequencing based approaches allowed to identify a large number of editing sites, up to two orders of magnitude higher than before. Two years later, a new release of DARNED recorded more than 330,000 editing sites in human (Kiran et al., 2013). This led to the design of tools to both visualize and annotate RNA-Seq data with known editing sites (Picardi et al., 2011; Distefano et al., 2013).

Although DARNED contains precious information regarding known editing sites, only a small portion of this have been later manually annotated, not providing any information about the spatiotemporal regulation of editing events through their editing level (Wahlstedt et al., 2009; Solomon et al., 2014). To improve this aspect, Ramaswami and Li built RADAR[4], a rigorously annotated database of A-to-I editing sites. Particularly, they have enriched RNA editing knowledge by including detailed manually curated information for each editing site, such as genomic coordinates, type of genomic region (intergenic region, 3′-or-5′ UTR, intron, or coding sequence if the editing site occurs in genic region), type of repetitive element (when the editing event occurs in Alu or not-Alu element), the conservation in other species (chimpanzee, rhesus, mouse), and the tissue-specific editing level when known. Currently, RADAR contains about 1.4 million editing sites as detected in Homo Sapiens (Ramaswami and Li, 2014). Among them, the editing sites that occur in human ncRNAs are only a small fraction, consisting of about 21,000 events, with only 1,219 editing sites in microRNAs. Despite being a relatively small percentage, amounting to about 1.6% of the total number of human editing sites, these miRNA editing events may very well posses significant importance as far as the editing phenomenon is concerned.

Without a doubt, continuous updating of the RADAR database gradually will become a precious resource for researchers in this field, leading to a better understanding of the editing phenomenon in coming years.

EFFECT OF RNA EDITING IN NON-CODING RNA MOLECULES

In the last decade, editing events have been discovered in ncRNA molecules, such as miRNAs, siRNAs, tRNAs, and lncRNAs. Although not fully demonstrated yet, these editing sites could alter the stability, the biogenesis, and target recognition of ncRNAs, as shown in **Figure 1**.

RNA editing in miRNAs and siRNAs

As seen above, many A-to-I editing sites in miRNAs have been discovered (Luciano et al., 2004; Kawahara et al., 2007; Alon et al., 2012), and these could influence miRNA-mediated gene regulation in several ways (Nishikura, 2010), although in some cases low percentage editing of mature miRNAs could be a low level of genomewide editing noise rather than possessing biological relevance (de Hoon et al., 2010). First, editing sites occurring in pri-miRNAs can suppress cleavage processing by Drosha and/or Dicer due to the presence of inosines, while in addition, highly edited dsRNAs could be rapidly degraded by Tudor-SN (TSN) (Yang et al., 2006). Second, some editing events in pri-miRNAs can produce edited pre-miRNAs, for which different scenarios can occur based on the location of the editing site. In particular, studies have demonstrated that A-to-I editing sites in miRNA seed regions can drastically change their target set (Kawahara et al., 2008; Alon et al., 2012), causing a functional transformation, but also affect the mRNA target selection and silencing processes (Kume et al., 2014).

Small interfereing RNAs, differently from miRNAs, originate from long double-strand RNAs exported to the cytoplasm, where they are cleaved by the Dicer-TRBP complex and successively loaded inside the RISC complex. It has been observed that ADAR1-p150, which acts in the cytoplasm, can bind to siRNAs preventing and thus overall reducing the cleavage process of the Dicer-TRBP complex (Yang et al., 2005; Kawahara et al., 2007).

Lately, a new role for ADAR1-p150 not associated to RNA editing was discovered, in which the enzyme forms an heterodimer complex with Dicer by protein–protein interaction

[2]http://darned.ucc.ie
[3]http://microrna.osumc.edu/mireditar
[4]http://RNAedit.com

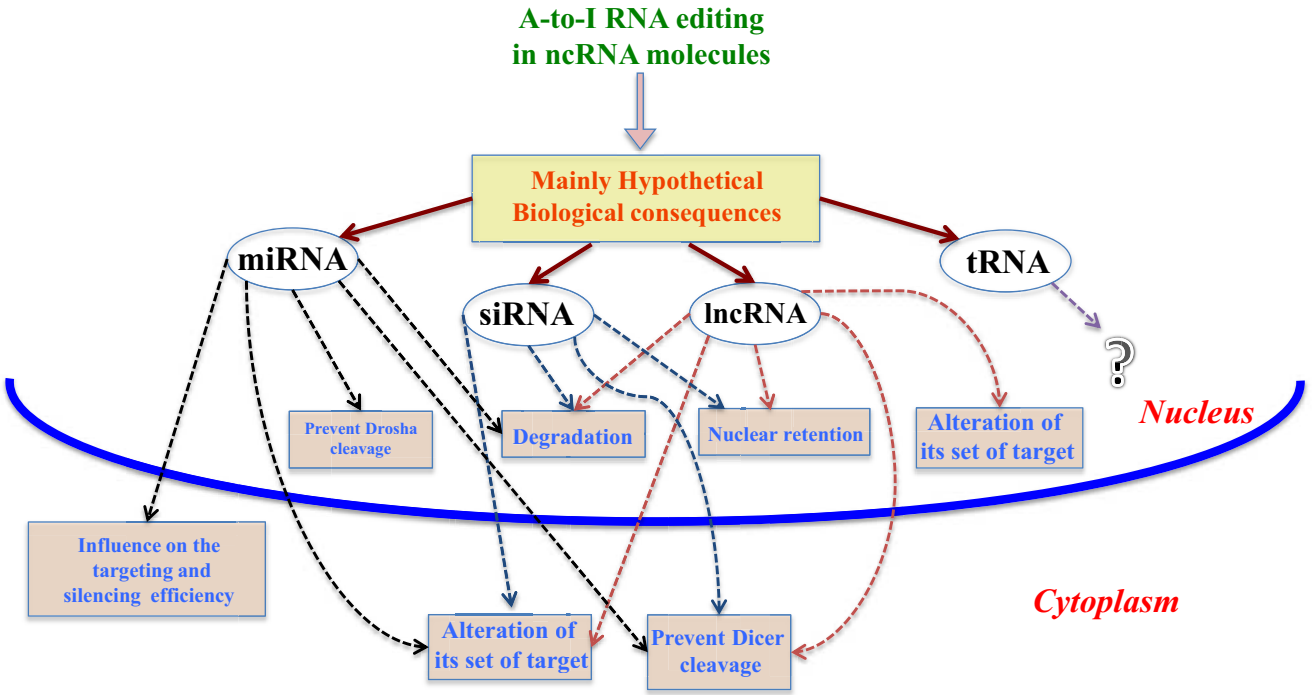


FIGURE 1 | Mainly hypothetical biological consequences. In this figure, we show some of the main biological consequences of A-to-I RNA editing in ncRNA molecules, both in nucleus and cytoplasm.

(PPI), increasing the rate of siRNA and miRNA processing and facilitating RISC loading and RNA silencing, instead of an antagonistic role in RNAi by an ADAR1–ADAR1 homodimer complex (Nishikura et al., 2013; Ota et al., 2013).

RNA editing in lncRNAs

Another category of ncRNAs is represented by long non-coding RNAs (lncRNAs). In recent years, HTS analyses have led to the identification of thousands of lncRNAs, many of which have revealed to be transcripts deriving from the antisense strand of protein coding genes. lncRNAs, due to their stable long double-strand regions, often originating from the presence of repetitive elements, such as Alus, can be affected by A-to-I RNA editing (Peng et al., 2012). The biological functions of A-to-I editing occurring in lncRNAs can be several.

Long non-coding RNAs can be retained in the nucleus as a consequence of the editing phenomenon until cleavage of the hyper-edited region takes place and the remaining lncRNA portion is exported to the cytoplasm (Prasanth et al., 2005). Nevertheless, as for miRNAs (Yang et al., 2006), edited lncRNAs could though be degraded through Tudor-SN. Considering the property lncRNAs possess to bind with RNA and DNA (Rinn and Chang, 2012; Mercer and Mattick, 2013), as well as RNA binding proteins (Hellwig and Bass, 2008), cases of editing sites in lncRNAs could clearly change their target set and RNP structures respectively, thus altering their intrinsic biological function (Geisler and Coller, 2013). Finally, a far more rare RNA editing phenomenon compared to the one caused by inverted repeat structures in mRNAs could occur for those transcripts which associate to antisense lncRNAs, providing a double strand RNA structure suitable for ADAR as suggested in (Geisler and Coller, 2013).

RNA editing in tRNAs

Differently from mRNAs and several categories of ncRNA molecules which undergo A-to-I editing primarily by ADARs, A-to-I editing events in mature transfer RNAs (tRNAs) in eukaryotes, can possibly be a result of adenosine deaminases acting on tRNA enzyme family (ADATs) (Su and Randau, 2011). A-to-I editing in these small ncRNAs is conserved in various species and occurs principally at positions 34, 37, and 57 of certain tRNAs (Torres et al., 2014). Despite this phenomenon being ubiquitously present in human tissues, the role of A-to-I tRNA editing remains still unknown.

CONCLUSION

As seen above, currently Alon's pipeline is the only HTS-based method to systematically identify A-to-I editing sites in pre- and mature microRNAs. There is a current and urgent necessity for new HTS-based methodologies to emerge in order to not only accurately identify and analyze editing events in other categories of ncRNA molecules, such as tRNAs, lncRNAs, and so on, but also to investigate through functional enrichment analysis, the biological outcomes that a single editing event can generate. Concurrently, it could be interesting to analyze how the editing phenomenon can influence a biological pathway within a temporally changing cellular condition, such as starvation or hypoxia, considering that

a single editing site in a ncRNA molecule could drastically modify its function.

ACKNOWLEDGMENTS

This work was supported by Italian Foundation for Cancer Research (NG 15046).

REFERENCES

Alon, S., and Eisenberg, E. (2013). Identifying RNA editing sites in miRNAs by deep sequencing. *Methods Mol. Biol.* 1038, 159–170. doi:10.1007/978-1-62703-514-9_9

Alon, S., Mor, E., Vigneault, F., Church, G. M., Locatelli, F., Galeano, F., et al. (2012). Systematic identification of edited microRNAs in the human brain. *Genome Res.* 22, 1533–1540. doi:10.1101/gr.131573.111

Athanasiadis, A., Rich, A., and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* 2:e391. doi:10.1371/journal.pbio.0020391

Bahn, J. H., Lee, J.-H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* 22, 142–150. doi:10.1101/gr.124107.111

Bass, B. L. (1997). RNA editing and hypermutation by adenosine deamination. *Trends Biochem. Sci.* 22, 157–162. doi:10.1016/S0968-0004(97)01035-9

Bass, B. L. (2002). RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* 71, 817–846. doi:10.1146/annurev.biochem.71.110601.135501

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., et al. (2014a). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* 24, 365–376. doi:10.1101/gr.164749.113

Bazak, L., Levanon, E. Y., and Eisenberg, E. (2014b). Genome-wide analysis of Alu editability. *Nucleic Acids Res.* 42, 6876–6884. doi:10.1093/nar/gku414

Blow, M. J., Grocock, R. J., Van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., et al. (2006). RNA editing of human microRNAs. *Genome Biol.* 7, R27. doi:10.1186/gb-2006-7-4-r27

Boguski, M. S., Lowe, T., and Tolstoshev, C. M. (1993). dbEST – database for "expressed sequence tags." *Nat. Genet.* 4, 332–333. doi:10.1038/ng0893-332

Borchert, G. M., Gilmore, B. L., Spengler, R. M., Xing, Y., Lanier, W., Bhattacharya, D., et al. (2009). Adenosine deamination in human transcripts generates novel microRNA binding sites. *Hum. Mol. Genet.* 18, 4801–4807. doi:10.1093/hmg/ddp443

Carmi, S., Borukhov, I., and Levanon, E. Y. (2011). Identification of widespread ultra-edited human RNAs. *PLoS Genet.* 7:e1002317. doi:10.1371/journal.pgen.1002317

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

Chen, C. X., Cho, D. S., Wang, Q., Lai, F., Carter, K. C., and Nishikura, K. (2000). A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA* 6, 755–767. doi:10.1017/S1355838200000170

de Hoon, M. J. L., Taft, R. J., Hashimoto, T., Kanamori-Katayama, M., Kawaji, H., Kawano, M., et al. (2010). Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res.* 20, 257–264. doi:10.1101/gr.095273.109

Distefano, R., Nigita, G., Macca, V., Laganà, A., Giugno, R., Pulvirenti, A., et al. (2013). VIRGO: visualization of A-to-I RNA editing sites in genomic sequences. *BMC Bioinformatics* 14(Suppl. 7):S5. doi:10.1186/1471-2105-14-S7-S5

Eggington, J. M., Greene, T., and Bass, B. L. (2011). Predicting sites of ADAR editing in double-stranded RNA. *Nat. Commun.* 2, 319. doi:10.1038/ncomms1324

Galeano, F., Tomaselli, S., Locatelli, F., and Gallo, A. (2012). A-to-I RNA editing: the "ADAR" side of human cancer. *Semin. Cell Dev. Biol.* 23, 244–250. doi:10.1016/j.semcdb.2011.09.003

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* 14, 699–712. doi:10.1038/nrm3679

Hellwig, S., and Bass, B. L. (2008). A starvation-induced noncoding RNA modulates expression of dicer-regulated genes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12897–12902. doi:10.1073/pnas.0805118105

Higuchi, M., Maas, S., Single, F. N., Hartner, J., and Rozov, A. (2000). Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* 406, 78–81. doi:10.1038/35017558

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. (2003). Nervous system targets of RNA editing identified by comparative genomics. *Science* 301, 832–836. doi:10.1126/science.1086763

Jepson, J. E. C., and Reenan, R. A. (2008). RNA editing in regulating gene expression in the brain. *Biochim. Biophys. Acta* 1779, 459–470. doi:10.1016/j.bbagrm.2007.11.009

Ju, Y. S., Kim, J.-I., Kim, S., Hong, D., Park, H., Shin, J.-Y., et al. (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat. Genet.* 43, 745–752. doi:10.1038/ng.872

Kawahara, Y. (2012). Quantification of adenosine-to-inosine editing of microRNAs using a conventional method. *Nat. Protoc.* 7, 1426–1437. doi:10.1038/nprot.2012.073

Kawahara, Y., Ito, K., Sun, H., Aizawa, H., Kanazawa, I., and Kwak, S. (2004). Glutamate receptors: RNA editing and death of motor neurons. *Nature* 427, 801–801. doi:10.1038/427801a

Kawahara, Y., Megraw, M., Kreider, E., Iizasa, H., Valente, L., Hatzigeorgiou, A. G., et al. (2008). Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res.* 36, 5270–5280. doi:10.1093/nar/gkn479

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. (2007). Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* 315, 1137–1140. doi:10.1126/science.1138050

Kim, D. D. Y., Kim, T. T. Y., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., et al. (2004). Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.* 14, 1719–1725. doi:10.1101/gr.2855504

Kim, U., Wang, Y., Sanford, T., Zeng, Y., and Nishikura, K. (1994). Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proc. Natl. Acad. Sci. U.S.A.* 91, 11457–11461. doi:10.1073/pnas.91.24.11457

Kiran, A., and Baranov, P. V. (2010). DARNED: a database of RNA editing in humans. *Bioinformatics* 26, 1772–1776. doi:10.1093/bioinformatics/btq285

Kiran, A. M., O'Mahony, J. J., Sanjeev, K., and Baranov, P. V. (2013). Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.* 41, D258–D261. doi:10.1093/nar/gks961

Kleinman, C. L., and Majewski, J. (2012). Comment on "widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335, 1302. doi:10.1126/science.1209658

Kume, H., Hino, K., Galipon, J., and Ui-Tei, K. (2014). A-to-I editing in the miRNA seed region regulates target mRNA selection and silencing efficiency. *Nucleic Acids Res.* 42, 10050–10060. doi:10.1093/nar/gku662

Laganà, A., Paone, A., Veneziano, D., Cascione, L., Gasparini, P., Carasi, S., et al. (2012). miR-EdiTar: a database of predicted A-to-I edited miRNA target sites. *Bioinformatics* 28, 3166–3168. doi:10.1093/bioinformatics/bts589

Lai, F., Chen, C. X., Carter, K. C., and Nishikura, K. (1997). Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. *Mol. Cell. Biol.* 17, 2413–2424.

Lehmann, K. A., and Bass, B. L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry* 39, 12875–12884. doi:10.1021/bi001383g

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

Li, J. B., Levanon, E. Y., Yoon, J.-K., Aach, J., Xie, B., Leproust, E., et al. (2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 324, 1210–1213. doi:10.1126/science.1170995

Lin, W., Piskol, R., Tan, M. H., and Li, J. B. (2012). Comment on "widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335, 1302. doi:10.1126/science.1210624

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

Maas, S., Patt, S., Schrey, M., and Rich, A. (2001). Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14687–14692. doi:10.1073/pnas.251531398

Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M., and Seeburg, P. H. (1996). RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J. Biol. Chem.* 271, 31795–31798. doi:10.1074/jbc.271.50.31795

Mercer, T. R., and Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi:10.1038/nsmb.2480

Mitra, S. A., Mitra, A. P., and Triche, T. J. (2012). A central role for long non-coding RNA in cancer. *Front. Genet.* 3:17. doi:10.3389/fgene.2012.00017

Morse, D. P., and Bass, B. L. (1999). Long RNA hairpins that contain inosine are present in *Caenorhabditis elegans* poly(A)+ RNA. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6048–6053. doi:10.1073/pnas.96.11.6048

Nishikura, K. (2006). Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat. Rev. Mol. Cell Biol.* 7, 919–931. doi:10.1038/nrm2061

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* 79, 321–349. doi:10.1146/annurev-biochem-060208-105251

Nishikura, K., Sakurai, M., Ariyoshi, K., and Ota, H. (2013). Antagonistic and stimulative roles of ADAR1 in RNA silencing. *RNA Biol.* 10, 1240–1247. doi:10.4161/rna.25947

Ota, H., Sakurai, M., Gupta, R., Valente, L., Wulff, B.-E. E., Ariyoshi, K., et al. (2013). ADAR1 forms a complex with Dicer to promote microRNA processing and RNA-induced gene silencing. *Cell* 153, 575–589. doi:10.1016/j.cell.2013.03.024

Park, E., Williams, B., Wold, B. J., and Mortazavi, A. (2012). RNA editing in the human ENCODE RNA-seq data. *Genome Res.* 22, 1626–1633. doi:10.1101/gr.134957.111

Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* 30, 253–260. doi:10.1038/nbt.2122

Picardi, E., D'Antonio, M., Carrabino, D., Castrignanò, T., and Pesole, G. (2011). ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments. *Bioinformatics* 27, 1311–1312. doi:10.1093/bioinformatics/btr117

Picardi, E., D'Erchia, A. M., Gallo, A., Montalvo, A., and Pesole, G. (2014). Uncovering RNA editing sites in long non-coding RNAs. *Front. Bioeng. Biotechnol.* 2:64. doi:10.3389/fbioe.2014.00064

Picardi, E., Gallo, A., Galeano, F., Tomaselli, S., and Pesole, G. (2012). A novel computational strategy to identify A-to-I RNA editing sites by RNA-Seq data: de novo detection in human spinal cord tissue. *PLoS ONE* 7:e44184. doi:10.1371/journal.pone.0044184

Picardi, E., and Pesole, G. (2013). REDItools: high-throughput RNA editing detection made easy. *Bioinformatics* 29, 1813–1814. doi:10.1093/bioinformatics/btt287

Pickrell, J. K., Gilad, Y., and Pritchard, J. K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335, 1302. doi:10.1126/science.1210484

Polson, A. G., and Bass, B. L. (1994). Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J.* 13, 5701–5711.

Porath, H. T., Carmi, S., and Levanon, E. Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* 5:4726. doi:10.1038/ncomms5726

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

Ramaswami, G., and Li, J. B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 42, D109–D113. doi:10.1093/nar/gkt996

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O'Connell, M. A. A., et al. (2013). Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods* 10, 128–132. doi:10.1038/nmeth.2330

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

Rueter, S. M., Dawson, T. R., and Emeson, R. B. (1999). Regulation of alternative splicing by RNA editing. *Nature* 399, 75–80. doi:10.1038/19992

Sakurai, M., Ueda, H., Yano, T., Okada, S., Terajima, H., Mitsuyama, T., et al. (2014). A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. *Genome Res.* 24, 522–534. doi:10.1101/gr.162537.113

Sakurai, M., Yano, T., Kawabata, H., Ueda, H., and Suzuki, T. (2010). Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nat. Chem. Biol.* 6, 733–740. doi:10.1038/nchembio.434

Solomon, O., Bazak, L., Levanon, E. Y., Amariglio, N., Unger, R., Rechavi, G., et al. (2014). Characterizing of functional human coding RNA editing from evolutionary, structural, and dynamic perspectives. *Proteins* 82, 3117–3131. doi:10.1002/prot.24672

Su, A. A. H., and Randau, L. (2011). A-to-I and C-to-U editing within transfer RNAs. *Biochemistry (Mosc)* 76, 932–937. doi:10.1134/S0006297911080098

Tomaselli, S., Locatelli, F., and Gallo, A. (2014). The RNA editing enzymes ADARs: mechanism of action and human disease. *Cell Tissue Res.* 356, 527–532. doi:10.1007/s00441-014-1863-3

Torres, A. G., Piñeyro, D., Filonava, L., Stracker, T. H., Batlle, E., and Ribas de Pouplana, L. (2014). A-to-I editing on tRNAs: biochemical, biological and evolutionary implications. *FEBS Lett.* 588, 4279–4286. doi:10.1016/j.febslet.2014.09.025

Wahlstedt, H., Luciano, D. J., Enstero, M., and Ohman, M. (2009). Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res.* 19, 978–986. doi:10.1101/gr.089409.108

Wahlstedt, H., and Ohman, M. (2011). Site-selective versus promiscuous A-to-I editing. *Wiley Interdiscip. Rev. RNA* 2, 761–771. doi:10.1002/wrna.89

Wang, Q., Khillan, J., Gadue, P., and Nishikura, K. (2000). Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* 290, 1765–1768. doi:10.1126/science.290.5497.1765

Wu, D., Lamm, A. T., and Fire, A. Z. (2011). Competition between ADAR and RNAi pathways for an extensive class of RNA targets. *Nat. Struct. Mol. Biol.* 18, 1094–1101. doi:10.1038/nsmb.2129

Yang, W., Chendrimada, T. P., Wang, Q., Higuchi, M., Seeburg, P. H., Shiekhattar, R., et al. (2006). Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat. Struct. Mol. Biol.* 13, 13–21. doi:10.1038/nsmb1041

Yang, W., Wang, Q., Howell, K. L., Lee, J. T., Cho, D.-S. C., Murray, J. M., et al. (2005). ADAR1 RNA deaminase limits short interfering RNA efficacy in mammalian cells. *J. Biol. Chem.* 280, 3946–3953. doi:10.1074/jbc.M407876200

Zhu, S., Xiang, J.-F., Chen, T., Chen, L.-L., and Yang, L. (2013). Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences. *BMC Genomics* 14:206. doi:10.1186/1471-2164-14-206

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.

Received: 31 October 2014; accepted: 09 March 2015; published online: 25 March 2015.
Citation: Nigita G, Veneziano D and Ferro A (2015) A-to-I RNA editing: current knowledge sources and computational approaches with special emphasis on non-coding RNA molecules. Front. Bioeng. Biotechnol. ***3****:37. doi: 10.3389/fbioe.2015.00037*
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

MINI REVIEW ARTICLE
published: 11 December 2014
doi: 10.3389/fbioe.2014.00065

Synthetic RNAs for gene regulation: design principles and computational tools

Alessandro Laganà[1]*, Dennis Shasha[2] and Carlo Maria Croce[1]

[1] *Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA*
[2] *Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*

Edited by:
Christian M. Zmasek, Sanford-Burnham Medical Research Institute, USA

Reviewed by:
Rui Alves, Universitat de Lleida, Spain
Yu Xue, Huazhong University of Science and Technology, China

****Correspondence:***
Alessandro Laganà, Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, 460 West 12th Avenue, Room 0995, Columbus, OH 43210, USA
e-mail: alessandro.lagana@osumc.edu

The use of synthetic non-coding RNAs for post-transcriptional regulation of gene expression has not only become a standard laboratory tool for gene functional studies but it has also opened up new perspectives in the design of new and potentially promising therapeutic strategies. Bioinformatics has provided researchers with a variety of tools for the design, the analysis, and the evaluation of RNAi agents such as small-interfering RNA (siRNA), short-hairpin RNA (shRNA), artificial microRNA (a-miR), and microRNA sponges. More recently, a new system for genome engineering based on the bacterial CRISPR-Cas9 system (Clustered Regularly Interspaced Short Palindromic Repeats), was shown to have the potential to also regulate gene expression at both transcriptional and post-transcriptional level in a more specific way. In this mini review, we present RNAi and CRISPRi design principles and discuss the advantages and limitations of the current design approaches.

Keywords: RNAi, siRNA, miRNA, a-miR, AntagomiR, Sponge, CRISPRi

INTRODUCTION

Natural regulatory RNAs are a heterogenous group of endogenous non-coding RNAs that modulate biological processes at many levels through different mechanisms. They have inspired the design of synthetic RNA molecules, such as riboswitches, sensors, and controllers, as key elements for programing cellular behaviors, as well as antisense-based approaches for specific gene expression regulation, which is the focus of this mini-review (Sharma et al., 2008; Culler et al., 2010; Liang et al., 2011).

RNA interference (RNAi) was discovered in 1998, when Andrew Fire and Craig C. Mello reported the capability of exogenous double-stranded RNAs (dsRNA) to silence genes in a specific manner in *C. elegans* (Fire et al., 1998). Central molecules in RNAi are microRNA (miRNA) and small-interfering RNA (siRNA).

miRNAs are small endogenous non-coding RNAs, typically 18–22 bp long, which derive from longer hairpin-shaped precursors called pre-miRNA (Bartel, 2004). A pre-miRNA can encode one or two different mature miRNAs, one from each arm (-5p and -3p). Pre-miRNAs come, in turn, from primary transcripts, called pri-miRNA, which are transcribed from miRNA genes. Mature miRNAs are incorporated into effector protein complexes called RISCs (RNA-induced silencing complex) and exert their regulatory function by binding specific target mRNAs through perfect or, more often, partial sequence complementarity, leading to the inhibition of their translation or promoting their degradation.

siRNAs are mostly exogenous dsRNA molecules derived from viral RNAs or artificially introduced into the cell (Chu and Rana, 2007).

The use of artificially designed siRNA has become a common and powerful strategy for the knock-down of gene expression yielding functional including therapeutic phenotypes (Gunsalus and Piano, 2005; Kim and Rossi, 2007). Several optimizations have been proposed in order to improve their efficacy and specificity (Liu et al., 2012b). Although research is focused on the development of selective delivery systems, a crucial factor is the presence of undesired off-target effects. siRNAs are designed to be perfectly complementary to their target sequences, ideally with few or no off-target genes. However, several studies have shown that a siRNA can bind mRNAs through partial complementarity, in a miRNA-like way, thus leading to undesirable and not easily predictable side effects (Birmingham et al., 2006; Jackson et al., 2006). In fact, despite the advances made in the recent past years, miRNA-target recognition has revealed itself to be a very dynamic mechanism influenced by many factors, which are only partially understood (Bartel, 2009; Thomas et al., 2010).

Along with specific gene silencing, the artificial repression of miRNAs can also provide a valuable tool for functional studies and have important therapeutic applications (Esquela-Kerscher and Slack, 2006; Garofalo et al., 2008; Croce, 2009). Two different strategies have been developed for the specific inhibition of miRNAs: antagomiRs and miRNA sponges (Krützfeldt et al., 2005; Ebert et al., 2007). The former consist of small RNAs exhibiting anti-complementarity to the miRNA to repress. The latter are longer RNA transcripts that act as attractors for miRNAs by distracting them from their original targets.

Finally, a novel methodology for artificial gene expression and miRNA regulation based on Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) has been recently proposed (Qi et al., 2013). CRISPR interference (CRISPRi) employs an engineered CRISPR/Cas system to control gene expression at the

transcriptional level through a catalytically inactive Cas9 protein. Recent studies have shown that the CRISPR/Cas system can also target RNA (Hale et al., 2009).

In this mini-review, we summarize RNAi and CRISPRi design principles and discuss the advantages and limitations of the current approaches.

siRNA DESIGN PRINCIPLES

siRNA are usually synthesized as double-stranded RNA duplexes or as hairpin-shaped molecules called shRNA. The siRNA design process consists of the identification of a functional binding site on a target mRNA sequence, which will correspond to the sense strand of the siRNA. The anti-sense sequence is obtained as the complement to the sense strand.

Many studies have been conducted to determine the features associated to functional siRNAs and have allowed to establish siRNA design rules. Elbashir et al. (2001b) suggest to choose the 23-nt sequence motif AA(N19)TT as binding site, where N19 means any combination of 19 nucleotides (nt) and corresponds to the sense strand of the siRNA. The complement to AA(N19) corresponds to the anti-sense strand (**Figures 1A–C**).

Symmetric 3′ dTdT overhangs are added to the siRNA duplex to improve its stability and facilitate RISC loading. Although other combinations of nucleotides are acceptable, dGdG overhangs should be avoided, as they appear to be associated to decreased siRNA activity (Elbashir et al., 2001a,b; Strapps et al., 2010). siRNA duplexes often have asymmetric loading of the anti-sense versus sense strands. The strand whose 5′ end is thermodynamically less stable is preferentially incorporated into the RISC (Khvorova et al., 2003).

siRNA design rules can be classified into sequence and structure rules (See Table S1 in Supplementary Material). Sequence rules concern the position of the binding site in the target transcript and its nucleotide composition. The target region should be chosen preferably 50-100 nt downstream of the start codon and should avoid the middle of the coding sequence (Elbashir et al., 2001a; Hsieh et al., 2004). The G/C content of the binding site, and consequently of the siRNA, is relevant to the silencing activity and should be in the range of 30–55%, although values as low as 25% or as high as 79% are still associated to functional siRNAs (Reynolds et al., 2004; Liu et al., 2012a).

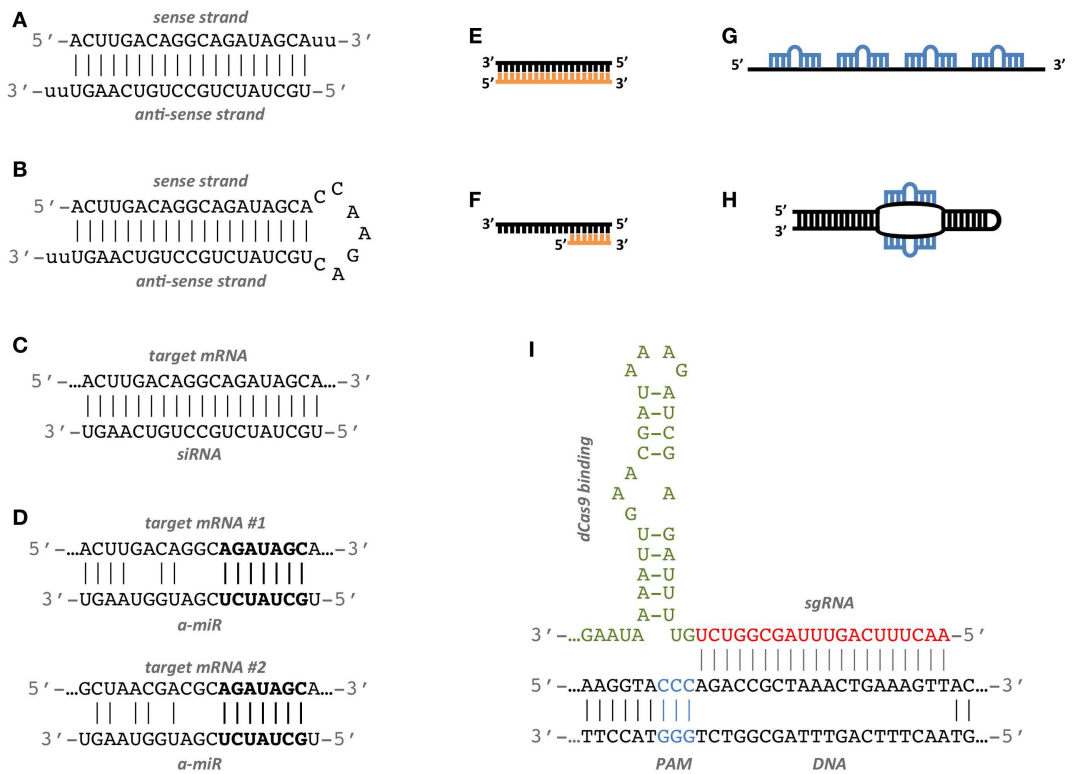


FIGURE 1 | Artificial RNA constructs for miRNA and gene regulation. **(A)** Standard double strand siRNA; the anti-sense strand is the active agent which binds the target site. **(B)** shRNA construct; it is produced inside the target cell from a DNA construct that has been delivered to the nucleus and it expresses the anti-sense active strand. **(C)** The siRNA anti-sense strand binds the target mRNA with perfect complementarity. **(D)** Example of an a-miR sequence targeting two different sites with partial complementarity. The seed sequence of the a-miR, highlighted in bold characters, matches perfectly the target sites. **(E)** The antagomiR sequence (orange) perfectly matches the sequence of the target miRNA (black). **(F)** The Tiny LNA sequence (orange) perfectly matches the seed sequence of the target miRNA (black). **(G)** miRNA sponge construct with four miRNA binding sites separated by spacers. **(H)** Synthetic TUD construct with two exposed miRNA binding sites. **(I)** Model of a CRISPR sgRNA sequence binding the target DNA region. The PAM sequence (blue) is a short DNA motif juxtaposed to the DNA complementary region. The base-pairing nucleotides of the sgRNA are shown in red, while the dCas9-binding hairpin is in green.

Table 1 | Computational tools for siRNA, a-miR and CRISPR design.

Tool	URL	Reference
siRNA Design Tools		
OptiRNAi 2.0	http://rnai.nci.nih.gov	Cui et al. (2004)
siDirect 2	http://sidirect2.rnai.jp	Naito et al. (2009)
siRNA Scales	http://gesteland.genetics.utah.edu/siRNA_scales	Matveeva et al. (2007)
siExplorer	http://rna.chem.t.u-tokyo.ac.jp/cgi/siexplorer.htm	Katoh and Suzuki (2007)
RFRCDB-siRNA	http://www.bioinf.seu.edu.cn/siRNA/index.htm	Jiang et al. (2007)
OligoWalk	http://rna.urmc.rochester.edu/cgi-bin/server_exe/oligowalk/oligowalk_form.cgi	Lu and Mathews (2008)
Sfold	http://sfold.wadsworth.org	Ding et al. (2004)
siMAX	http://www.operon.com/products/siRNA/sirna-overview.aspx	Schramm and Ramey (2005)
DSIR	http://biodev.cea.fr/DSIR/	Vert et al. (2006)
siRNA Scan	http://bioinfo2.noble.org/RNAiScan.htm	Xu et al. (2006)
RNAxs	http://rna.tbi.univie.ac.at/cgi-bin/RNAxs	Tafer et al. (2008)
i-Score	http://www.med.nagoya-u.ac.jp/neurogenetics/i_Score/i_score.html	Ichihara et al. (2007)
siVirus	http://sivirus.rnai.jp	Naito et al. (2006)
a-miR Design Tools		
miR-Synth	http://microrna.osumc.edu/mir-synth/	Lagana et al. (2014)
CRISPR Design Tools		
Cas9 Design	http://cas9.cbi.pku.edu.cn	Ma et al. (2013)
CRISPR Design	http://crispr.mit.edu	Hsu et al. (2013)
Broad Inst. sgRNA Designer	http://www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design	Doench et al. (2014)
sgRNAcas9	http://www.biootools.com	Xie et al. (2014)
CRISPR Genome Analyzer	http://crispr-ga.net	Guell et al. (2014)
CasOT	http://eendb.zfgenetics.org/casot	Xiao et al. (2014)
DNA 2.0 gRNA Design Tool	https://www.dna20.com/eCommerce/cas9/input	Cong et al. (2013); Ran et al. (2013)
E-CRISP	http://www.e-crisp.org/E-CRISP/	Heigwer et al. (2014)
ZiFiT	http://zifit.partners.org/ZiFiT/	Hwang et al. (2013)
CHOPCHOP	https://chopchop.rc.fas.harvard.edu	Montague et al. (2014)
CRISPRseek	http://www.bioconductor.org/packages/release/bioc/html/CRISPRseek.html	Zhu et al. (2014)
SSFinder	https://code.google.com/p/ssfinder/	Upadhyay and Sharma (2014)

URLs and references are given for each tool.

Numerous sequence rules regard the selection of nucleotides to prefer or avoid in specific positions of either the sense or the anti-sense strand of the duplex. For example, a higher content of A/U nucleotides in the 5′ end of the anti-sense strand of the siRNA yields higher silencing efficacy (Ui-Tei et al., 2004; Shabalina et al., 2006). Also, the 5′ half of the anti-sense strand dictates competition potency of siRNAs, which is a consequence of the RNAi machinery saturation followed by transfection of multiple siRNAs (Yoo et al., 2008).

Other relevant sequence features include the absence of internal repeats and the presence/absence of specific motifs (Reynolds et al., 2004).

Structure rules refer to the thermodynamics features of the siRNA/target duplex and are mostly expressed in terms of the nucleotide composition of the duplex itself or of the area surrounding the binding site (Chalk et al., 2004; Shabalina et al., 2006). Structure rules specify functional levels of binding energy at different positions of the duplex, and optimal energy difference between different positions of the duplex itself. Another important thermodynamic feature associated to siRNA efficacy is the structural accessibility of the target site. It has been demonstrated, in fact, that an mRNA stretch, which is not involved in a strict secondary structure exhibits a stronger binding affinity to a siRNA (or miRNA) molecule than one with a highly structured conformation (Tafer et al., 2008).

Several optimizations have been proposed to improve the activity of siRNA molecules, such as a more accurate prediction of the active strand of the duplex, design rules to avoid competition with endogenous miRNAs and vectors expressing multiple siRNAs at once (Cheng et al., 2009; Ma et al., 2014; Malefyt et al., 2014).

Many tools are available online for the design of siRNA and shRNA molecules (see **Table 1**).

OFF-TARGETS, MULTIPLE TARGETS, AND THE a-miR APPROACH

Although siRNAs and shRNAs are designed to specifically target a single gene through perfect complementarity to the binding site, several studies show that they can partially bind to many other transcripts in a way reminiscent of the endogenous miRNAs (Birmingham et al., 2006; Jackson et al., 2006). A single miRNA can potentially regulate hundreds of different mRNAs through partial sequence complementarity. In particular, perfect base pairing of

the 5′ end region of the miRNA, termed "seed," to a binding site located in the 3′ UTR of a mRNA, is usually sufficient to yield a significant repression of the target, while other recent studies also report functional centered site-mediated interactions (Shin et al., 2010; Helwak et al., 2013; Martin et al., 2014).

This represents a relevant drawback of single-target siRNAs, especially when pools of four or five siRNA duplexes per target gene are used to achieve stronger repression but also leading to widespread off-target effects.

One approach to the off-targeting problem consists of employing pools of siRNAs, at low concentrations, that target a single gene in multiple sites (Straka and Boese, 2010). The advantage of this approach lies in the fact that such pools are both effective on that one target, while the effects of a low concentrations siRNA on other potential targets should be negligible (Arvey et al., 2010; Larsson et al., 2010). Another study showed that siRNAs with a bulge at position 2 of the anti-sense strand were able to discriminate better between perfectly matched and mismatched targets (Dua et al., 2009; Li et al., 2010).

Targeting multiple genes can also be an intended choice, as there are many biological and biomedical applications in which it is important to regulate multiple genes at once while suffering as few side effects as possible. One way to achieve this goal is to exploit the multi-targeting properties of endogenous miRNAs by employing artificially designed miRNAs, or a-miRs. Two recent papers have shown that a-miRs can successfully repress at least two targets simultaneously by binding to one or more sites in their 3′ UTRs (**Figure 1D**) (Arroyo et al., 2014; Lagana et al., 2014). The employment of a single multi-target a-miR in place of a pair or a pool of single-target siRNAs is likely to yield significant repression of targets with few off-target effects.

SILENCING THE SILENCERS: ANTAGOMIRS AND miRNA SPONGES

While the inhibition of over-expressed genes has been the main goal of RNAi research for years, the de-repression of down-regulated miRNA targets has increasingly gained importance over time. The inhibition of endogenous miRNAs was first introduced in 2005 by Krützfeldt et al. (2005). They employed cholesterol-conjugated oligo-ribonucleotides, which they termed "antagomiRs," reproducing the anti-sense strand of the endogenous miRNA they inhibit. Their design is thus straightforward, as there is not much space for sequence variations (**Figure 1E**). Since then, a variety of chemical modifications have been proposed in order to increase binding affinity, improve nuclease resistance and facilitate *in vivo* delivery. They include locked nucleic acid (LNA), which possesses the highest affinity toward complementary RNA, Bifunctional oligodeoxynucleotide/antagomiR constructs, which ensure high transfection efficiency and prevention of unintended immune stimulation, and morpholino oligomers, which have been shown to be efficient inhibitors of both pri-miRNA and mature miRNA activity in zebrafish and *Xenopus laevis* (Summerton and Weller, 1997; Braasch and Corey, 2001; Petersen and Wengel, 2003; Ziegler et al., 2013). A further variant of antagomiRs is represented by short seed-targeting LNA oligonucleotides, called tiny LNAs. These molecules allow simultaneous inhibition of miRNAs within families sharing the same seed (**Figure 1F**) (Obad et al., 2011).

AntagomiRs represent one well-established tool for miRNA functional studies, and several works have also shown successful employment of antagomiRs as therapeutic agents able to restore disease-associated pathways altered by miRNA up-regulation. Like siRNAs, AntagomiRs can also have significant off-target effects, as they act like endogenous miRNAs and may hit complementary mRNA transcripts. However, experiments have showed no detectable effect on mRNAs with perfect tiny LNA complementary sites, not even at the proteomic level (Obad et al., 2011).

miRNA sponges are an alternative to antagomiRs. They act as competitive inhibitors that distract endogenous miRNAs from their natural targets. Many sponge variants have been described, such as miRNA-target mimics, miRNA decoys, and miRNA erasers, and they all consist of RNA constructs containing multiple binding sites for the miRNA to be sponged (**Figure 1G**) (Carè et al., 2007; Ebert et al., 2007; Franco-Zorrilla et al., 2007; Sayed et al., 2008).

A basic sponge consists of an RNA sequence exhibiting 4–10 miRNA binding sites separated by short spacers, usually 2–4 nt long. These sites can be either bulged or perfectly complementary to the miRNAs. In the first case, a bulge at positions 9–12 of the binding site is introduced in order to prevent cleavage and degradation of the sponge. Sponges with bulged binding sites produce stronger de-repressive effects than sponges with perfect binding sites (Ebert et al., 2007). Kluiver et al. (2012) developed a methodology for the rapid generation of miRNA sponges by making use of simple constructs with up to 20 perfect or bulged miRNA binding sites.

Structural optimizations have also been proposed. TuD RNAs (tough decoy RNAs) are efficient sponges with structurally accessible and indigestible miRNA binding sites (**Figure 1H**) (Haraguchi et al., 2009, 2012). The optimal TuD RNA consists of a bulged stem-loop structure where both sides of the bulge are miRNA binding sites which are perfectly complementary to the miRNA sequence and which do not form any base-pairing regions longer than 9nt.

CRISPRi: THE GENE SILENCING REVOLUTION

An exciting and promising advance in the field of artificial gene regulation comes from Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). CRISPR is a natural adaptive immune system used by archaea and bacteria against phage and plasmids (Jinek et al., 2012). This system is genomically encoded by the prokaryotic chromosome and consists of a series of short repeats separated by spacer sequences that match previously encountered foreign DNA. Thus, CRISPR arrays are transcribed and processed in order to produce mature crRNAs, which are loaded onto effector protein complexes and function as a guide to target recognition and degradation.

The CRISPR/Cas system has been engineered to function with synthetic small guide RNA (sgRNA) in order to perform genome editing in eukaryotes (Mali et al., 2013) (See Table S2 in Supplementary Material). The sgRNA consists of a 20 nt crRNA sequence complementary to the target region followed by a 42 nt Cas9-binding hairpin and a 40 nt transcription terminator. The target region must be of the form N20NGG that is any 21 nucleotides followed by GG. NGG is the 3′ protospacer-adjacent motif (PAM)

and is required for Cas9 binding. This particular PAM sequence is derived from *Streptococcus pyogenes*, but other functional PAM sequences have been characterized from other bacteria (Esvelt et al., 2013). In addition, if a U6 snRNA or T7 promoters are used to express the sgRNA, this must start with G or GG, respectively, in order to maintain transcript initiation. Thus, the target region must be of the form GN19NGG or GGN18NGG. Ultimately, the beginning of the sgRNA and the PAM sequence will depend on the specific promoters and Cas9 used (**Figure 1I**).

DNA breaks caused by Cas9 are repaired through either homologous recombination or non-homologous end joining (NHEJ) mechanisms, thus this system can be used to either disrupt or edit a gene (i.e., insertions and deletions). Many tools are currently available online for the design of sgRNAs (**Table 1**).

Besides genome-editing applications, the CRISPR/Cas9 system can be employed for gene expression regulation. The system, known as CRISPR interference (CRISPRi), is based on a catalytically dead Cas9 (dCas9) lacking endonuclease activity co-expressed with a sgRNA (Gilbert et al., 2013; Qi et al., 2013). Instead of generating DNA breaks, the recognition complex interferes with transcriptional elongation, RNA polymerase binding, or transcription factor binding, leading to efficient inhibition of gene expression. CRISPRi gene silencing is inducible and reversible and recognition of the targets depends solely on the sgRNA sequence (Qi et al., 2013).

A "seed" region has been identified as the 12nt region adjacent to the PAM site. Mismatches in the seed region can dramatically reduce the repression, while mismatches in the non-seed area can cause a mild effect. Design guidelines recommend using a length of 20–25 nt as the base-pairing region of the sgRNA (Larson et al., 2013) and provide specific design rules based on nucleotide preference for active sgRNA (Doench et al., 2014). A recent study aimed at the identification of features of effective sgRNA specific to CRISPRi, suggests that the target site should be chosen from −50 to +300 bp relative to the Transcription Start Site (TSS) of a gene (Gilbert et al., 2014). The authors observed that nucleotide homopolymers have a strongly negative effect on sgRNA activity and that the GC content of the sgRNA or the binding site is not correlated with sgRNA activity, although another study reports a decreased activity of sgRNA with low or high GC content (Doench et al., 2014). Moreover, CRISPRi activity seems to be highly sensitive to mismatches between the sgRNA and DNA sequence, thus the authors conclude that properly designed sgRNA will have minimal off-target effects. However, previous studies reported silencing activity with sgRNAs exhibiting mismatches to the target in the seed area (Cradick et al., 2013) and that off-targets might be cell type dependent and determined by various complicated factors in addition to primary DNA sequences (Duan et al., 2014). Thus, side effects still constitute a challenge, which needs to be properly addressed by further focused research.

Gilbert et al. also introduced the sunCas9 CRISPRa system, in which expression of a single sgRNA with one binding site is sufficient to turn on genes that are poorly expressed or that increase the expression of well-expressed genes (Gilbert et al., 2014; Tanenbaum et al., 2014).

CRISPRi can also be successfully employed to knock out miRNAs, by using a sgRNA/Cas9 complex targeting the pre-miRNA sequence (Zhao et al., 2014), and to study functional miRNA-target interactions *in vivo* by site-specific genome engineering (Bassett et al., 2014).

Finally, although current tools for CRISPRi are based on the DNA targeting approach described above, the discovery of other Cas proteins targeting RNA molecules, such as Cmr, suggests an alternative post-transcriptional methodology similar to RNAi (Hale et al., 2009; Zebec et al., 2014).

CONCLUSION

Both RNAi and CRISPRi represent valid approaches for artificial gene regulation and both can suffer from significant side effects which may result from factors beyond sequence match. One clear advantage of CRISPRi over RNAi is that being an exogenous system it does not compete with the endogenous machinery of miRNA processing. Nevertheless, both techniques require more work in terms of enhancing targeting efficiency and reducing side effects.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00065/abstract

REFERENCES

Arroyo, J. D., Gallichotte, E. N., and Tewari, M. (2014). Systematic design and functional analysis of artificial microRNAs. *Nucleic Acids Res.* 42, 6064–6077. doi:10.1093/nar/gku171

Arvey, A., Larsson, E., Sander, C., Leslie, C. S., and Marks, D. S. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.* 6, 1–7. doi:10.1038/msb.2010.24

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi:10.1016/S0092-8674(04)00045-5

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

Bassett, A. R., Azzam, G., Wheatley, L., Tibbit, C., Rajakumar, T., McGowan, S., et al. (2014). Understanding functional miRNA-target interactions in vivo by site-specific genome engineering. *Nat. Commun.* 5, 1–11. doi:10.1038/ncomms5640

Birmingham, A., Anderson, E. M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., et al. (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat. Meth.* 3, 199–204. doi:10.1038/nmeth854

Braasch, D. A., and Corey, D. R. (2001). Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. *Chem. Biol.* 8, 1–7. doi:10.1016/S1074-5521(00)00058-2

Carè, A., Catalucci, D., Felicetti, F., Bonci, D., Addario, A., Gallo, P., et al. (2007). MicroRNA-133 controls cardiac hypertrophy. *Nat. Med.* 13, 613–618. doi:10.1038/nm1582

Chalk, A. M., Wahlestedt, C., and Sonnhammer, E. L. L. (2004). Improved and automated prediction of effective siRNA. *Biochem. Biophys. Res. Commun.* 319, 264–274. doi:10.1016/j.bbrc.2004.04.181

Cheng, T. L., Teng, C. F., Tsai, W. H., Yeh, C. W., Wu, M. P., Hsu, H. C., et al. (2009). Multitarget therapy of malignant cancers by the head-to-tail tandem array multiple shRNAs expression system. *Cancer Gene Ther.* 16, 516–531. doi:10.1038/cgt.2008.102

Chu, C.-Y., and Rana, T. M. (2007). Small RNAs: regulators and guardians of the genome. *J. Cell. Physiol.* 213, 412–419. doi:10.1002/jcp.21230

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823. doi:10.1126/science.1231143

Cradick, T. J., Fine, E. J., Antico, C. J., and Bao, G. (2013). CRISPR/Cas9 systems targeting-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* 41, 9584–9592. doi:10.1093/nar/gkt714

Croce, C. M. (2009). Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.* 10, 704–714. doi:10.1038/nrg2634

Cui, W., Ning, J., Naik, U. P., and Duncan, M. K. (2004). OptiRNAi, an RNAi design tool. *Comput. Methods Programs Biomed.* 75, 67–73. doi:10.1016/j.cmpb.2003.09.002

Culler, S. J., Hoff, K. G., and Smolke, C. D. (2010). Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science* 330, 1251–1255. doi:10.1126/science.1192128

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* 32, W135–W141. doi:10.1093/nar/gkh449

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., et al. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* doi:10.1038/nbt.3026

Dua, P., Yoo, J. W., Kim, S., and Lee, D.-K. (2009). Modified siRNA structure with a single nucleotide bulge overcomes conventional siRNA-mediated off-target silencing. *Mol. Ther.* 19, 1676–1687. doi:10.1038/mt.2011.109

Duan, J., Lu, G., Xie, Z., Lou, M., Luo, J., Guo, L., et al. (2014). Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Res.* 24, 1009–1012. doi:10.1038/cr.2014.87

Ebert, M. S., Neilson, J. R., and Sharp, P. A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat. Meth.* 4, 721–726. doi:10.1038/nmeth1079

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001a). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411, 494–498. doi:10.1038/35078107

Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001b). Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* 20, 6877–6888. doi:10.1093/emboj/20.23.6877

Esquela-Kerscher, A., and Slack, F. J. (2006). Oncomirs – microRNAs with a role in cancer. *Nat. Rev. Cancer* 6, 259–269. doi:10.1038/nrc1840

Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J., and Church, G. M. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Meth.* 10, 1116–1121. doi:10.1038/nmeth.2681

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

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., et al. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* 39, 1033–1037. doi:10.1038/ng2079

Garofalo, M., Condorelli, G., and Croce, C. M. (2008). MicroRNAs in diseases and drug response. *Curr. Opin. Pharmacol.* 8, 661–667. doi:10.1016/j.coph.2008.06.005

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., et al. (2014). Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* 159, 647–661. doi:10.1016/j.cell.2014.09.029

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* 154, 442–451. doi:10.1016/j.cell.2013.06.044

Guell, M., Yang, L., and Church, G. M. (2014). Genome editing assessment using CRISPR genome analyzer (CRISPR-GA). *Bioinformatics* 30, 2968–2970. doi:10.1093/bioinformatics/btu427

Gunsalus, K. C., and Piano, F. (2005). RNAi as a tool to study cell biology: building the genome-phenome bridge. *Curr. Opin. Cell Biol.* 17, 3–8. doi:10.1016/j.ceb.2004.12.008

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., et al. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* 139, 945–956. doi:10.1016/j.cell.2009.07.040

Haraguchi, T., Nakano, H., Tagawa, T., Ohki, T., Ueno, Y., Yoshida, T., et al. (2012). A potent 2′-O-methylated RNA-based microRNA inhibitor with unique secondary structures. *Nucleic Acids Res.* 40, e58–e58. doi:10.1093/nar/gkr1317

Haraguchi, T., Ozaki, Y., and Iba, H. (2009). Vectors expressing efficient RNA decoys achieve the long-term suppression of specific microRNA activity in mammalian cells. *Nucleic Acids Res.* 37, e43–e43. doi:10.1093/nar/gkp040

Heigwer, F., Kerr, G., and Boutros, M. (2014). Correspondence. *Nat. Meth.* 11, 122–123. doi:10.1038/nmeth.2812

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi:10.1016/j.cell.2013.03.043

Hsieh, A., Bo, R., Manola, J., Vazquez, F., Bare, O., Khvorova, A., et al. (2004). A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic. Acids Res.* 32, 893–901. doi:10.1093/nar/gkh238

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* 31, 827–832. doi:10.1038/nbt.2647

Hwang, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., et al. (2013). Brief communications. *Nat. Biotechnol.* 31, 227–229. doi:10.1038/nbt.2501

Ichihara, M., Murakumo, Y., Masuda, A., Matsuura, T., Asai, N., Jijiwa, M., et al. (2007). Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. *Nucleic Acids Res.* 35, e123–e123. doi:10.1093/nar/gkm699

Jackson, A. L., Burchard, J., Schelter, J., Chau, B. N., Cleary, M., Lim, L., et al. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA* 12, 1179–1187. doi:10.1261/rna.25706

Jiang, P., Wu, H., Da, Y., Sang, F., Wei, J., Sun, X., et al. (2007). RFRCDB-siRNA: Improved design of siRNAs by random forest regression model coupled with database searching. *Comput. Methods Programs Biomed.* 87, 230–238. doi:10.1016/j.cmpb.2007.06.001

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821. doi:10.1126/science.1225829

Katoh, T., and Suzuki, T. (2007). Specific residues at every third position of siRNA shape its efficient RNAi activity. *Nucleic Acids Res.* 35, e27–e27. doi:10.1093/nar/gkl1120

Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115, 209–216. doi:10.1016/S0092-8674(03)00801-8

Kim, D. H., and Rossi, J. J. (2007). Strategies for silencing human disease using RNA interference. *Nat. Rev. Genet.* 8, 173–184. doi:10.1038/nrg2006

Kluiver, J., Gibcus, J. H., Hettinga, C., Adema, A., Richter, M. K. S., Halsema, N., et al. (2012). Rapid generation of MicroRNA sponges for microRNA inhibition. *PLoS ONE* 7:e29275. doi:10.1371/journal.pone.0029275

Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M., et al. (2005). Silencing of microRNAs in vivo with "antagomirs". *Nature* 438, 685–689. doi:10.1038/nature04303

Lagana, A., Acunzo, M., Romano, G., Pulvirenti, A., Veneziano, D., Cascione, L., et al. (2014). miR-Synth: a computational resource for the design of multi-site multi-target synthetic miRNAs. *Nucleic Acids Res.* 42, 5416–5425. doi:10.1093/nar/gku202

Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., and Qi, L. S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* 8, 2180–2196. doi:10.1038/nprot.2013.132

Larsson, E., Sander, C., and Marks, D. (2010). mRNA turnover rate limits siRNA and microRNA efficacy. *Mol. Syst. Biol.* 6, 1–9. doi:10.1038/msb.2010.89

Li, X., Yoo, J. W., Lee, J. H., Hahn, Y., Kim, S., and Lee, D.-K. (2010). Identification of sequence features that predict competition potency of siRNAs. *Biochem. Biophys. Res. Commun.* 398, 92–97. doi:10.1016/j.bbrc.2010.06.041

Liang, J. C., Bloom, R. J., and Smolke, C. D. (2011). Engineering biological systems with synthetic RNA molecules. *Mol. Cell* 43, 915–926. doi:10.1016/j.molcel.2011.08.023

Liu, Q., Zhou, H., Cui, J., Cao, Z., and Xu, Y. (2012a). Reconsideration of in-silico siRNA design based on feature selection: a cross-platform data integration perspective. *PLoS ONE* 7:e37879. doi:10.1371/journal.pone.0037879

Liu, Q., Zhou, H., Zhu, R., Xu, Y., and Cao, Z. (2012b). Reconsideration of in silico siRNA design from a perspective of heterogeneous data integration: problems and solutions. *Brief Bioinform* 15, 292–305. doi:10.1093/bib/bbs073

Lu, Z. J., and Mathews, D. H. (2008). OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. *Nucleic Acids Res.* 36, W104–W108. doi:10.1093/nar/gkn250

Ma, H., Zhang, J., and Wu, H. (2014). Designing Ago2-specific siRNA/shRNA to avoid competition with endogenous miRNAs. *Mol Ther Nucleic Acids* 3, e176. doi:10.1038/mtna.2014.27

Ma, M., Ye, A. Y., Zheng, W., and Kong, L. (2013). A guide RNA sequence design platform for the CRISPR/Cas9 system for model organism genomes. *Biomed Res. Int.* 2013, 1–4. doi:10.1155/2013/270805

Malefyt, A. P., Wu, M., Vocelle, D. B., Kappes, S. J., Lindeman, S. D., Chan, C., et al. (2014). Improved asymmetry prediction for short interfering RNAs. *FEBS J.* 281, 320–330. doi:10.1111/febs.12599

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., et al. (2013). RNA-guided human genome engineering via Cas9. *Science* 339, 823–826. doi:10.1126/science.1232033

Martin, H. C., Wani, S., Steptoe, A. L., Krishnan, K., Nones, K., Nourbakhsh, E., et al. (2014). Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. *Genome Biol.* 15, R51. doi:10.1186/gb-2014-15-3-r51

Matveeva, O., Nechipurenko, Y., Rossi, L., Moore, B., Saetrom, P., Ogurtsov, A. Y., et al. (2007). Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Nucleic Acids Res.* 35, e63–e63. doi:10.1093/nar/gkm088

Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., and Valen, E. (2014). CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* 42, W401–W407. doi:10.1093/nar/gku410

Naito, Y., Ui-Tei, K., Nishikawa, T., Takebe, Y., and Saigo, K. (2006). siVirus: web-based antiviral siRNA design software for highly divergent viral sequences. *Nucleic Acids Res.* 34, W448–W450. doi:10.1093/nar/gkl214

Naito, Y., Yoshimura, J., Morishita, S., and Ui-Tei, K. (2009). siDirect 2.0: updated software for designing functional siRNA with reduced seed-dependent off-target effect. *BMC Bioinformatics* 10:392. doi:10.1186/1471-2105-10-392

Obad, S., Santos dos, C. O., Petri, A., Heidenblad, M., Broom, O., Ruse, C., et al. (2011). Silencing of microRNA families by seed-targeting tiny LNAs. *Nat. Genet.* 43, 371–378. doi:10.1038/ng.786

Petersen, M., and Wengel, J. (2003). LNA: a versatile tool for therapeutics and genomics. *Trends Biotechnol.* 21, 74–81. doi:10.1016/S0167-7799(02)00038-0

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., et al. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173–1183. doi:10.1016/j.cell.2013.02.022

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154, 1380–1389. doi:10.1016/j.cell.2013.08.021

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S., and Khvorova, A. (2004). Rational siRNA design for RNA interference. *Nat. Biotechnol.* 22, 326–330. doi:10.1038/nbt936

Sayed, D., Rane, S., Lypowy, J., He, M., Chen, I.-Y., Vashistha, H., et al. (2008). MicroRNA-21 targets Sprouty2 and promotes cellular outgrowths. *Mol. Biol. Cell* 19, 3272–3282. doi:10.1091/mbc.E08-02-0159

Schramm, G., and Ramey, R. (2005). siRNA design including secondary structure target site prediction. *Nat. Meth.* 2. doi:10.1038/nmeth780

Shabalina, S. A., Spiridonov, A. N., and Ogurtsov, A. Y. (2006). Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics* 7:65. doi:10.1186/1471-2105-7-65

Sharma, V., Nomura, Y., and Yokobayashi, Y. (2008). Engineering complex Riboswitch regulation by dual genetic selection. *J. Am. Chem. Soc.* 130, 16310–16315. doi:10.1021/ja805203w

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., and Bartel, D. P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell* 38, 789–802. doi:10.1016/j.molcel.2010.06.005

Straka, M., and Boese, Q. (2010). Current topics in RNAi: why rational pooling of siRNAs is SMART. *Thermo Fisher Scientific Inc*

Strapps, W. R., Pickering, V., Muiru, G. T., Rice, J., Orsborn, S., Polisky, B. A., et al. (2010). The siRNA sequence and guide strand overhangs are determinants of in vivo duration of silencing. *Nucleic Acids Res.* 38, 4788–4797. doi:10.1093/nar/gkq206

Summerton, J., and Weller, D. (1997). Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev.* 7, 187–195. doi:10.1089/oli.1.1997.7.187

Tafer, H., Ameres, S. L., Obernosterer, G., Gebeshuber, C. A., Schroeder, R., Martinez, J., et al. (2008). The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.* 26, 578–583. doi:10.1038/nbt1404

Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S., and Vale, R. D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* 159, 635–646. doi:10.1016/j.cell.2014.09.039

Thomas, M., Lieberman, J., and Lal, A. (2010). Desperately seeking microRNA targets. *Nat. Struct. Mol. Biol.* 17, 1169–1174. doi:10.1038/nsmb.1921

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., et al. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.* 32, 936–948. doi:10.1093/nar/gkh247

Upadhyay, S. K., and Sharma, S. (2014). SSFinder: high throughput CRISPR-Cas target sites prediction tool. *Biomed Res. Int.* 2014, 1–4. doi:10.1155/2014/742482

Vert, J. P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* 7:520. doi:10.1186/1471-2105-7-520

Xiao, A., Cheng, Z., Kong, L., Zhu, Z., Lin, S., Gao, G., et al. (2014). CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics* 30, 1180–1182. doi:10.1093/bioinformatics/btt764

Xie, S., Shen, B., Zhang, C., Huang, X., and Zhang, Y. (2014). sgRNAcas9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PLoS ONE* 9:e100448. doi:10.1371/journal.pone.0100448

Xu, P., Zhang, Y., Kang, L., Roossinck, M. J., and Mysore, K. S. (2006). Computational estimation and experimental verification of off-target silencing during posttranscriptional gene silencing in plants. *Plant Physiol.* 142, 429–440. doi:10.1104/pp.106.083295

Yoo, J. W., Kim, S., and Lee, D.-K. (2008). Competition potency of siRNA is specified by the 5'-half sequence of the guide strand. *Biochem. Biophys. Res. Commun.* 367, 78–83. doi:10.1016/j.bbrc.2007.12.099

Zebec, Z., Manica, A., Zhang, J., White, M. F., and Schleper, C. (2014). CRISPR-mediated targeted mRNA degradation in the Archaeon *Sulfolobus solfataricus*. *Nucleic Acids Res.* 42, 5280–5288. doi:10.1093/nar/gku161

Zhao, Y., Dai, Z., Liang, Y., Yin, M., Ma, K., He, M., et al. (2014). Sequence-specific inhibition of microRNA via CRISPR/CRISPRi system. *Sci. Rep.* 4:3943. doi:10.1038/srep03943

Zhu, L. J., Holmes, B. R., Aronin, N., and Brodsky, M. H. (2014). CRISPRseek: a bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. *PLoS ONE* 9:e108424. doi:10.1371/journal.pone.0108424

Ziegler, S., Eberle, M. E., Wölfle, S. J., Heeg, K., and Bekeredjian-Ding, I. (2013). Bifunctional oligodeoxynucleotide/antagomir constructs: evaluation of a new tool for microRNA silencing. *Nucleic Acid Therapeut.* 23, 427–434. doi:10.1089/nat.2013.0447

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.

Received: 30 September 2014; accepted: 23 November 2014; published online: 11 December 2014.

*Citation: Laganà A, Shasha D and Croce CM (2014) Synthetic RNAs for gene regulation: design principles and computational tools. Front. Bioeng. Biotechnol. **2**:65. doi: 10.3389/fbioe.2014.00065*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 29 September 2014
doi: 10.3389/fbioe.2014.00038

IsomiRage: from functional classification to differential expression of miRNA isoforms

Heiko Muller†, Matteo Jacopo Marzi† and Francesco Nicassio*

Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Milan, Italy

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Tao Sun, Cornell University Weill Medical College, USA
Francesco Russo, National Research Council, Italy

***Correspondence:**
Francesco Nicassio, Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), c/o Campus IFOM-IEO, Via Adamello 16, Milan 20139, Italy
e-mail: francesco.nicassio@iit.it

†Heiko Muller and Matteo Jacopo Marzi have contributed equally to this work.

As more small RNA sequencing libraries are becoming available, it clearly emerges that microRNAs (miRNAs) are highly heterogeneous both in length and sequence. In comparison to canonical miRNAs, miRNA isoforms (termed as "isomiRs") might exhibit different biological properties, such as a different target repertoire, or enhanced/reduced stability. Nonetheless, this layer of information has remained largely unexplored due to the scarcity of small RNA NGS-datasets and the absence of proper analytical tools. Here, we present a workflow for the characterization and analysis of miRNAs and their variants in next-generation sequencing datasets. IsomiRs can originate from an alternative dicing event ("templated" forms) or from the addition of nucleotides through an enzymatic activity or target-dependent mechanisms ("non-templated" forms). Our pipeline allows distinguishing canonical miRNAs from templated and non-templated isomiRs by alignment to a custom database, which comprises all possible 3′-, 5′-, and trimmed variants. Functionally equivalent isomiRs can be grouped together according to the type of modification (e.g., uridylation, adenylation, trimming . . .) to assess which miRNAs are more intensively modified in a given biological context. When applied to the analysis of primary epithelial breast cancer cells, our methodology provided a 40% increase in the number of detected miRNA species and allowed to easily identify and classify more than 1000 variants. Most modifications were compatible with templated IsomiRs, as a consequence of imprecise Drosha or Dicer cleavage. However, some non-templated variants were consistently found either in the normal or in the cancer cells, with the 3′-end adenylation and uridylation as the most frequent events, suggesting that miRNA post-transcriptional modification frequently occurs. In conclusion, our analytical tool permits the deconvolution of miRNA heterogeneity and could be used to explore the functional role of miRNA isoforms.

Keywords: miRNA, isomiRs, next-generation sequencing, pipeline, alignment, cancer

INTRODUCTION

microRNAs (miRNAs), a small (18–25 nt long), evolutionarily conserved class of non-coding RNAs, are important regulators of transcriptional programs by silencing the expression of a multitude of target mRNAs at a post-transcriptional level (Bartel, 2009). The biogenesis of miRNAs typically requires a nuclear cleavage of the primary transcript by the Drosha/DGCR8 complex and a cytoplasmic cleavage of the hairpin-folded precursor miRNA (pre-miRNA) by Dicer [reviewed in Krol et al. (2010)]. The product of this cleavage is usually a mature 21/22 bp miRNA duplex, which is loaded onto the RNA-induced silencing complex (RISC) to function in the miRNA silencing mechanism (Gregory et al., 2005). Only one strand is retained in the RISC, usually the one with unstable base-pairing at its 5′-end, and it mediates target repression through base complementarity between the miRNA "seed region" (nucleotides 2–7) and the miRNA responsive elements (MRE), mostly located at the 3′ untranslated region (3′UTR) of target genes (Bartel, 2009).

Generally, each mature miRNA is annotated as a unique mature sequence (the reference or canonical miRNA sequence) and could derive from either the 5′ or 3′ arm of the same pre-miRNA hairpin (termed as -5p or -3p, respectively). However, the recent advent of next-generation sequencing has clearly shown that mature miRNAs can be present in several sequence variants or isoforms, named "isomiRs" [reviewed in Neilsen et al. (2012)]. Initially, isomiRs were considered as sequencing artifacts, but a growing body of evidence revealed that isomiRs are actual miRNA variants that can exert a biological activity. For instance, isomiRs are found associated with Argonaute proteins in the RISC complex as canonical miRNAs and could exert silencing of a specific target in *in vitro* luciferase assays (Lee et al., 2010; Cloonan et al., 2011). The generation of isomiRs is heterogeneous. In fact, they can originate from imprecise cleavage by Drosha or Dicer (the so-called "templated isomiRs"), which generates variants that show perfect sequence complementarity to their pre-miRNA. Alternatively, isomiRs could be generated by post-transcriptional modifications due to enzymatic activity, which could either add or remove specific nucleotides to miRNA ends. These miRNA variants are known as "non-templated isomiRs," with sequence imperfectly matching their pre-miRNA. Typically, non-templated modifications occur at the 3′end, while 5′end isomiRs are rare (Newman et al., 2011; Wyman et al., 2011). This is likely due to fact that a 5′-end

modification (templated or non-templated) actually modifies the target repertoire of the miRNA, which is dictated by the "seed" region [nucleotide 2–7 (Bartel, 2009)]. The expression profiles of miRNA variants are dynamic, with differences across tissues or cell lines (Landgraf et al., 2007). Nonetheless, the functional significance of isomiRs has remained elusive due to the limited number of tools available to specifically monitor their levels in sequencing experiments [e.g., isomiRex (Sablok et al., 2013), miRNA-MATE (Cloonan et al., 2011), miRAnalyzer (Hackenberg et al., 2009)]. In sporadic cases, it was shown that isomiRs could alter the target specificity (Azuma-Mukai et al., 2008), the efficiency of Ago loading (Burroughs et al., 2010), or the half-life (Katoh et al., 2009) of the cognate miRNA. Regardless of their biological activity, many isomiRs are highly expressed, even more than the corresponding canonical miRNAs. Thus, their annotation is particularly relevant in order to properly analyze expression profiles and eventually identify contexts where miRNA isoforms could be functional.

Here, we describe a pipeline that allows the identification and analysis of all miRNA variants (canonical miRNAs and "templated" or "non-templated" isomiRs) from small RNA sequencing experiments (Illumina). These variants could be grouped according to the site (5′-end or 3′-end) or the type of modification (trimming, adenylation, uridylation . . .) to assess the extent of miRNA modifications in a given biological context. As a proof-of-principle analysis, we applied our methodology to analyze miRNAs and isomiRs expression in human samples (i.e., primary normal and breast cancer cells), revealing that miRNA modifications frequently occur and may significantly affect global miRNA expression and regulation.

MATERIALS AND METHODS

SMALL RNA SAMPLES: CELL CULTURE, RNA ISOLATION, AND SMALL RNA SEQUENCING

The samples described in this work were prepared from a triple-negative breast cancer primary culture and its normal counterpart as described in Pece et al. (2004). The epithelial origin of the cultures was confirmed by immunofluorescence with an anti-Pan cytokeratin antibody (Sigma-Aldrich). All tissues were collected at the European Institute of Oncology via standardized operative procedures approved by the Institutional Ethical Board, and informed consent was obtained for all tissue specimens. Total RNA, including small species, was isolated through the miRNeasy mini kit (Qiagen). One microgram of total RNA was used to prepare Small RNA libraries following the Illumina TruSeq™ Small RNA Sample Preparation Guide, as by manufacturers' instructions. The libraries were sequenced at 50 bp single-read mode and 80 million read depth on an Illumina HiSeq 2000 platform. All the relevant steps of the *IsomiRage* analysis workflow are fully described in the text. Sequencing results are listed in Table S1 in Supplementary Material. Raw data together with detailed description of the procedures are available in GEO database (GSE21090).

QUANTITATIVE REAL TIME PCR

RT-qPCR reactions were performed in triplicate using the miScript RT system in conjunction with miScript primer assays (Qiagen), as by manufacturers' protocol. One microgram of total RNA was used to prepare cDNA. U6b was used as housekeeping.

STATISTICAL ANALYSIS

Microsoft Excel was used to generate bar graphs. Bivariate analyses, pie-chart, and statistics (Fisher's test, Student's *t*-test) were performed using JMP 10 (SAS) software.

IsomiRage JAVA TOOL

IsomiRage is a standalone desktop application written in the Java programing language. It was developed using NetBeans 7.3.1 Integrated Development Environment software. IsomiRage requires Java 1.6 to run and has been tested on Window 7 and MacOS 10 operating systems. The *IsoMirRagetool* (updated to the latest miRbase release, miRbase 21) is available at http://cru.genomics.iit.it/Isomirage/.

RESULTS

We present a pipeline, named as "*IsomiRage*," for profiling the miRNAs/isomiRs and corresponding differential expression patterns using Illumina next-generation sequencing datasets of small RNA. We discuss the application of *IsomiRage* to the analysis of small RNA sequencing data obtained by matched normal and tumor primary breast cell culture with the Illumina Hiseq 2000 sequencing system. The *IsomiRage* workflow has three main steps summarized in **Figure 1**: filtering of reads, alignment on a custom genome, and quantification and normalization of IsomiRs.

STEP 1 – FILTERING OF READS

Small RNA libraries are routinely prepared following the Illumina TruSeq™ Small RNA Sample Preparation guide, shown in **Figure 2**. Different biological samples are marked with specific 6 bp sequencing indices to allow multiplexing. According to our experience, up to 12 different small RNA libraries can be pooled in a single sequencing lane to obtain up to 18 million filtered reads from each library. Sequencing is performed in single-read mode with read length of 50 bp. De-multiplexing is carried out using CASAVA software to produce reads in FastQ format for each biological sample. Adapters used during library preparation are removed using The Flexible Adapter Remover software[1] (FAR version 2.15). As shown in **Figure 2**, adapter sequence may be found only at the 3′-end of the reads and corresponds to adapters RA3 and RPI, which have identical 5′-ends. FAR typically produces collections of reads whose lengths are distributed in a multi-modal distribution as shown in **Figure 3**. The largest mode is located at length 22, which corresponds to miRNAs. Minor modes can be observed at length 10, 34, 0, and 51. The first (length 10) marks the peak of range of read-lengths between 6 and 17 bases, which likely represent break-down products. The reads of the 34-bases mode show homology to tRNAs. The two last modes are at length 0 and length 51. The former corresponds to PCR fragments not containing any successfully cloned RNA molecules while the latter represents PCR fragments with RNA molecules longer than 50 bases or where adapter removal has failed for other reasons. Adapter removal may fail when the fraction of the adapter represented in the read is too short for being recognized as an adapter-derived sequence or when the adapter sequence contains errors. It is worth noting that adapter

[1] https://wiki.gacrc.uga.edu/wiki/FAR

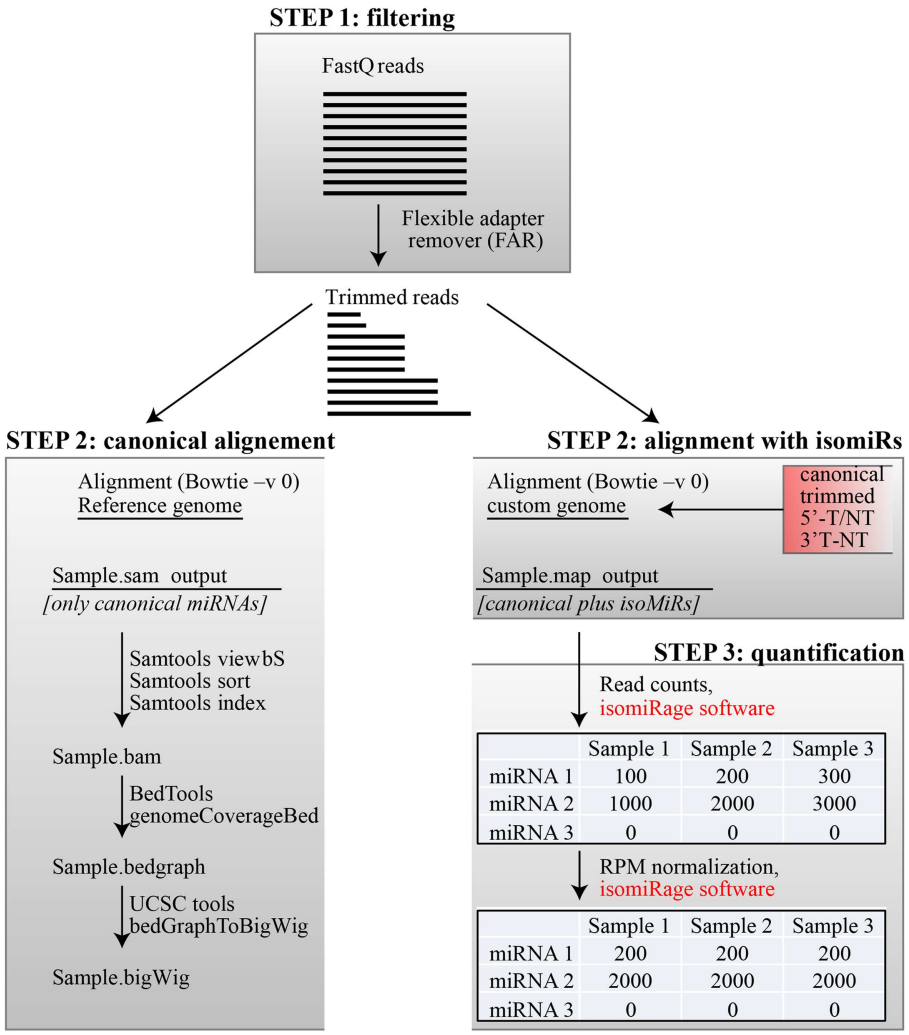


FIGURE 1 | IsomiRage pipeline. A schematic representation of the miRNA analysis workflow is shown, distinguishing the three main steps described in the text. Alignment of trimmed FastQ reads (Step 2) is carried out either using a reference genome for the production of browser viewable tracks or to custom sequence libraries for quantitative analyses of miRNA and their isoforms. The quantification of mapped reads in this latter case is performed through an *ad hoc* java application called *IsomiRage*, which also performs the read-per-million normalization and provides, for each miRNA, the different kinds of modifications found.

removal is essential for successful alignment of small RNA reads to a reference genome.

STEP 2 – ALIGNMENT

Filtered reads are aligned to a custom genome that includes the sequences of all canonical mature miRNAs[2] (2578 human and 1975 mouse miRNAs, according to the release 20 of miRBASE) and their related isomiRs (5′-end, 3′-end and trimmed variants, shown in Table S1 in Supplementary Material). The variant sequences were generated including all the possible combinations of one, two, or three bases extending the 5′- or the 3′-end of known miRNA sequences plus the sequences obtained by trimming canonical miRNA from their 3′-end down to a length of 18 bp (reads below 18 bp were not considered since the alignment would be unreliable). Each isomiR is associated to a series of feature, including the corresponding canonical miRNA, the site of modification (5′-end or 3′-end), the type of modification (trimming; addition of one, two, or three nucleotides; type of nucleotide/s added), and the origin of the isomiR ("templated" or "non-templated"). The latter definition was based on the alignment of each isomiR to the sequence of the pre-miRNA. A perfect pairing with the pre-miRNA sequence is associated to "templated" variants, while non-perfect pairing is associated to "non-templated" variants. It is

[2]http://www.mirbase.org

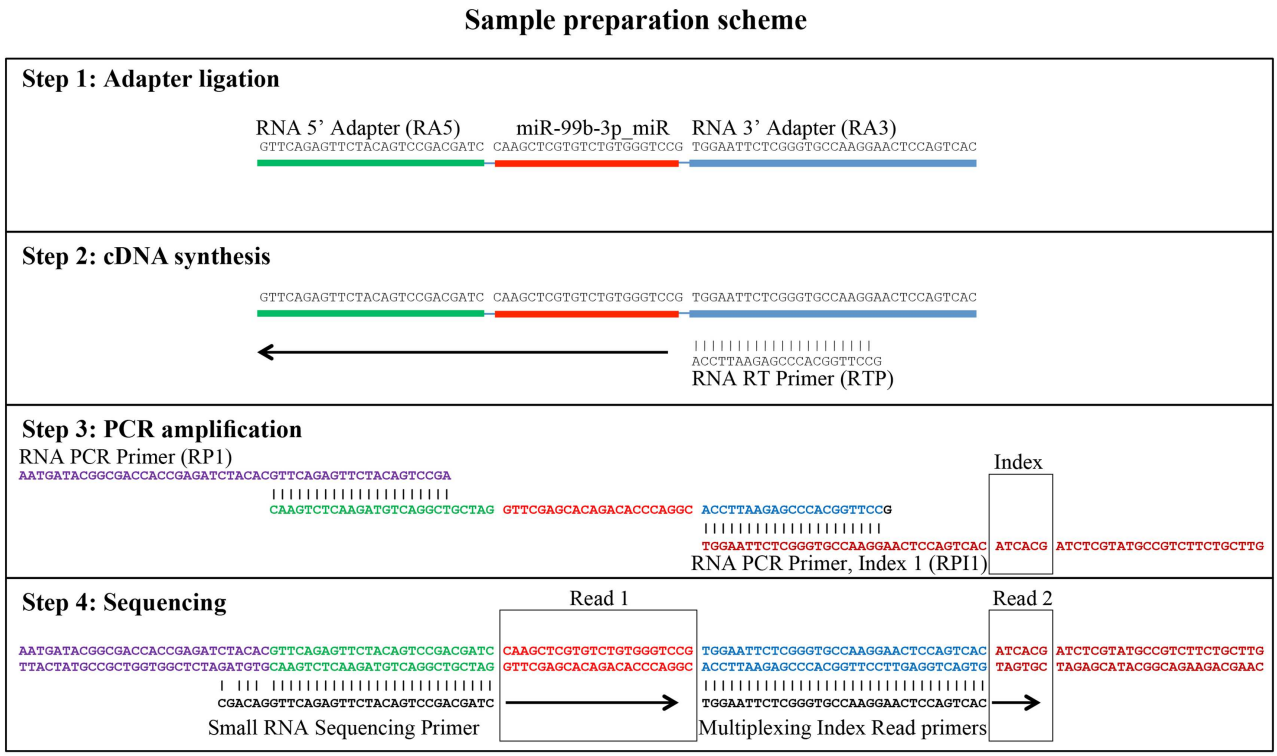


FIGURE 2 | Generation of small RNA sequencing libraries. The major steps from sample preparation to sequencing for small RNA molecules according to the Illumina protocol (Illumina TruSeq™ Small RNA Sample Preparation) are shown. The steps include adapter ligation, first strand cDNA synthesis, PCR-amplification, and sequencing. The regions of homology of the adapters involved are shown. The sequence of miR-99b-3p was chosen as an example. Read 1 indicates the sequencing cycles producing miRNA related sequence data. Read 2 indicates the production of index reads used for de-multiplexing of samples. Note that the miRNA related reads may contain adapter sequence at their 3′ end, which needs to be removed prior to downstream analyses.

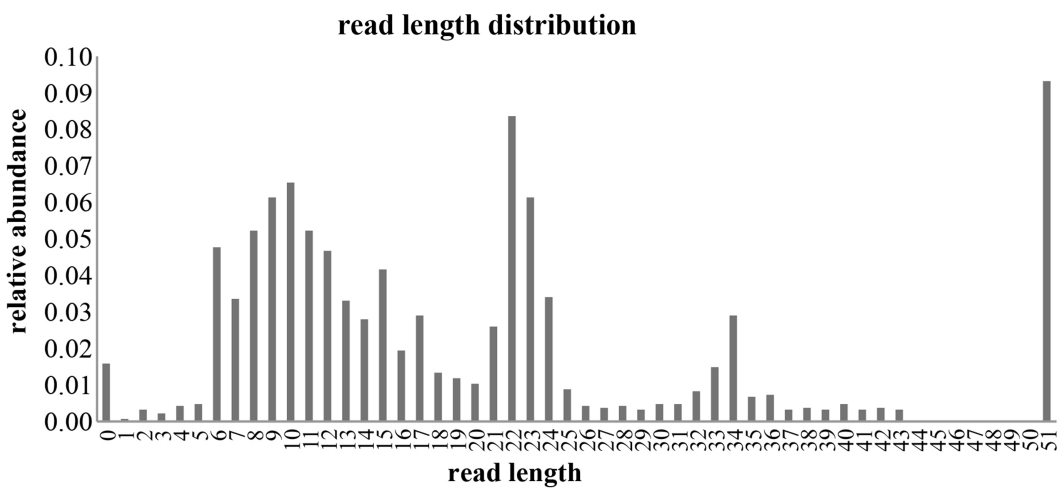


FIGURE 3 | Distribution of filtered reads. The distribution of read-lengths after adapter removal is shown. The distribution is multi-modal with peaks at 0, 10, 22, 34, and 51 bases. Mode 22 represents the miRNA related reads that are considered in this study.

worth mentioning that there is still a probability that a "templated" base variation might be a *de novo* modification, rather than an imprecise cleavage by Drosha or Dicer. Alignment is performed with the Bowtie ultrafast short-read aligner in the -v 0 alignment mode, which specifies that no mismatches are allowed (**Figure 1**). Only the best alignment is reported for each read. The Bowtie

output is stored in .map textual format and supplied as input to a custom software (i.e., *IsomiRage*) for downstream analyses. Filtered reads could also be aligned to the reference genome. In this case, the Bowtie output is stored in SAM format (Li et al., 2009) and contains data only for canonical miRNAs. The Bowtie output is processed further to produce browser viewable bam and bigwig files. These files can be used for qualitative analyses.

STEP 3 – QUANTIFICATION AND NORMALIZATION

To estimate the expression level of a given miRNA or IsomiR, the number of perfectly aligned Illumina reads are counted by *IsomiRage* JAVA tool (available at http://cru.genomics.iit.it/Isomirage/). The software reads the .map Bowtie output file and ensures that the read aligns perfectly to the chosen reference. Of note, this approach works only if the Bowtie output contains one and only one reported alignment for each Illumina read. This is achieved using the Bowtie -v 0 switch together with the best switch. The output is a table that lists the number of reads for each isoform in each biological condition (see **Figure 1**). We routinely obtain three to five million reads for each biological condition. To enable quantitative comparisons between samples, the read numbers must be normalized for sequencing depth. This step is carried out by standard read-per-million (RPM) normalization, providing a table of RPM-normalized read counts that can be used for comparisons of fold changes and other downstream analyses.

APPLICATION OF THE PIPELINE: SMALL RNA SEQUENCING OF NORMAL AND CANCER BREAST CELLS

Alignment and sequencing output

As a proof-of principle analysis, we applied our methodology to analyze miRNAs and isomiRs expression in real samples. We sequenced small RNAs from 1 μg of total RNA obtained from a matched normal/tumor primary culture pair of breast epithelium. We obtained about 18 million filtered reads for each sample, of which about 7 million were aligned to the custom genome (Table S1 in Supplementary Material; **Figure 4**). Of note, approximately 4 million reads could be mapped to canonical miRNAs, claiming that with our pipeline the sequencing output could be improved almost twofold (**Figure 4**). The improvement in the sequencing output has been similarly observed across multiple experiments and samples (not shown), regardless of the number of multiplexed samples (from 2 to 12). Considering the data from the normal and the tumor sample as a whole, we obtained at least 1 read for 1228 different miRNA species, of which 318 present with >100 counts in at least 1 sample (Table S1 in Supplementary Material). Of note, without considering IsomiRs, we would have identified only 876 miRNAs, 219 having >100 reads. Thus, our pipeline considerably expands the number of detected species and increases the number of mapped reads almost twofold.

Differential expression analysis

Having obtained the aligned data, we moved on to analyze the differential expression of canonical miRNAs in the tumor compared to the normal sample (**Figure 5**). To calculate fold changes, data where normalized to total read counts (RPM). We selected those miRNA robustly expressed (>100 reads) and identified 66 miRNAs differentially regulated ($|x| > 1$ $\log_2$ fold, **Figure 5A**). Among

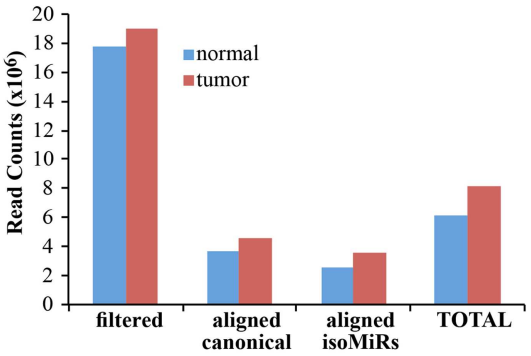


FIGURE 4 | Alignment of filtered reads. The cumulative amount of reads, after adapter removal, is shown for the two samples analyzed. "Filtered" reads refer to the sum of all the species reported in **Figure 3**. Of note, the alignment on the custom genome required a read to be at least 18 nucleotides in length. "Total" is the sum of canonical miRNAs and all their isoforms.

these, we selected randomly four upregulated and four downregulated miRNAs and measured their expression level by RT-qPCR (**Figure 5B**). For one miRNA, namely, miR-432-5p, RT-qPCR was not sensitive enough to detect the miRNA either in the normal or in the tumor sample (N.D., no data). Nonetheless, as shown in **Figure 5B**, 6/7 miRNAs were found concordantly regulated by the two methodologies (RNA-seq vs. RT-qPCR), both in qualitative and quantitative terms. Thus, our pipeline produces precise measurements of mature miRNA levels, with an 86% validation rate by an independent method.

Expression of isomiRs: the case of miR-92a

We next analyzed the expression of canonical miRNAs together with their isomiRs. **Figure 6** shows the locus of human pre-miR-92a-1 with the 5p- and 3p-arms and their related mature miRNA species found in the small RNA sequencing experiments. As expected, most of the variants (shown those with >10 counts) come from the 3p-arm, which is the usual processed arm (see miRbase 20 as reference), including 5′-end, 3′-end variants, and trimmed forms. For this miRNA, the canonical form is the prevalent one (>80% of all reads are the canonical hsa-miR-92a-3p; see **Figure 6**), followed by the trimmed and the 3′-end templated modifications. The 5′-end variants are poorly represented. There are a huge number of 3′end non-templated modifications, some with a robust level of expression (hundreds or even thousands of read counts, **Figure 6**). This very heterogeneous class could be grouped based on the type of the first nucleotide added to the mature miRNA, which should correspond to a different enzymatic activity (termed as A-, G-, C-, U-forms). To this regard, it is possible to distinguish "pure" forms (the same nucleotide added one, two, or three times; marked with a triangle in **Figure 6**) from "mixed" forms (with different nucleotides; marked with a circle in **Figure 6**). The nucleotide distribution appears very uneven, with the A-forms extremely abundant (pure and mixed equally distributed) followed by the U-forms. Modifications with the C or G bases are extremely rare.

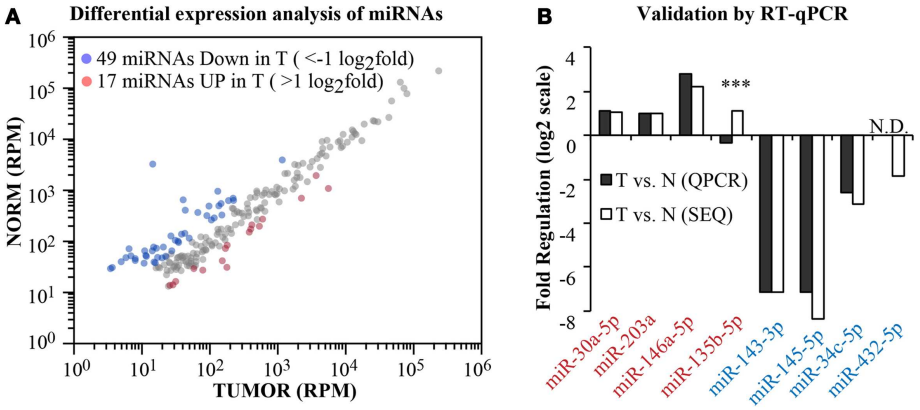


FIGURE 5 | Differential expression analysis. **(A)** A scatter plot shows the expression level of canonical miRNAs in the normal and tumor samples. Data have been normalized in reads-per-million (RPM) and shown in $\log_{10}$ scale. Blue and red dots identify those miRNAs with >1 $\log_2$ fold regulation between the two samples. **(B)** Comparison of the fold change as measured by RNA-seq or by RT-qPCR (miScript assay, Qiagen) for eight randomly selected miRNAs (four upregulated and four downregulated). Three asterisks mark non-concordant results. N.D., no PCR data.

Expression of isomiRs at genome-wide level

If we consider the expression of isomiRs at a global scale, we observed that about one-third of all detected species (>1 read count) are composed of canonical miRNAs, while the others come from miRNA variants, mostly at the 3′-end (**Figure 7A**). This uneven distribution is much more evident when considering read counts, with canonical miRNAs accounting for around 60% of all reads, followed by 3′-end modifications (30%) and 3′trimmed forms (10%) (**Figure 7B**). The 5′-end modifications only represented 0.4–0.5% of reads. Overall, templated modifications were roughly two-times more expressed than non-templated modifications. We observed little or no differences between the normal and the tumor sample, suggesting that, if any, the dynamic regulation of miRNA modifications is limited to specific isoforms (**Figure 7B**). Next, we analyzed the impact of variants on the total expression level of each miRNA (Table S1 in Supplementary Material; **Figure 7C**). Only miRNAs with >100 counts (canonical plus isomiRs) were considered. As expected, the canonical miRNA was the prevalent form (>50% of reads) in more than half of cases. Trimmed forms are well represented and constitute more than 20% of total reads for approximately 100 miRNAs (**Figure 7C**). Their distribution is similar to the one of 3′-end templated modifications. Indeed, trimmed variants can originate equally from active 3′ shortening (exonucleolytic cleavage) or alternative dicing during miRNA biogenesis. Conversely, 5′-end modifications (templated or non-templated) encompass a minority fraction for each and every miRNA (**Figure 7C**). Surprisingly, 3′-end non-templated variants, which are unambiguously a product of a post-biogenetic activity, constitute more than 10% of total reads for approximately 100 miRNAs (more than 20% for about 50 miRNAs; **Figure 7C**). No major differences were observed at a global level in the tumor sample compared to the normal. In fact, only 16 of the 258 (6.2%) miRNAs commonly expressed in the 2 samples showed a variation >10% in 3′ non-templated isomiRs, and 37/258 (14.3%) a variation >5% (Table S1 in Supplementary Material). These data, although coming from just two samples, confirm that the dynamicity of regulation of isomiRs is limited to selected species rather than a global effect.

Non-templated 3′ end modifications

Non-templated 3′-end isomiRs could originate from the activity of nucleotidyl-transferases (Neilsen et al., 2012). These enzymes usually catalyze the addition of uridyl and adenyl nucleotides at the 3′-end of miRNAs. Thus, we expect that uridylation and adenylation should be the prevalent modifications. As shown previously for miR-92a (see **Figure 6**), we classified 3′-end non-templated (3′-NT) isomiRs into "Iso-groups" (A-forms, C-forms, G-forms, U-forms) according to the nucleotide added at the 3′ end. To be rigorous in our definition, we focused only on the "pure" forms (those with the same nucleotide, e.g., A-forms include only -A, -AA, and -AAA modifications). As shown in **Figure 8**, adenylation was the most common modification (approximately 50% of the 3′NT modifications are A-forms, **Figure 8A**) and encompassed most of the reads (**Figure 8B**) followed by uridylation (50% of the 3′NT modifications and 20% of the reads; **Figures 8A,B**). Conversely, C- and G-forms accounted only for <5% of the 3′ non-templated modifications (**Figures 8A,B**). The frequency of adenylation was much higher than expected ($p < 0.0001$ Fisher's test), even when compared to the frequency of the last nucleotide of the 3′-end templated forms. A very similar trend has been observed in the tumor and in the normal sample (**Figures 8A,B**).

DISCUSSION

In the last few years, it became clear that the miRNome is far more complex than previously thought (Lee et al., 2010). The reference sequences reported in miRBase (canonical miRNAs) are usually the prevalent ones, but contemplating the alternative variants, called isomiRs, is crucial to completely understand the complexity of miRNA transcriptome. Here, we described a streamlined pipeline (termed *IsomiRage*) to identify and analyze miRNA isoforms from next-generation sequencing data. The pipeline has been developed for the analysis of data coming from Illumina

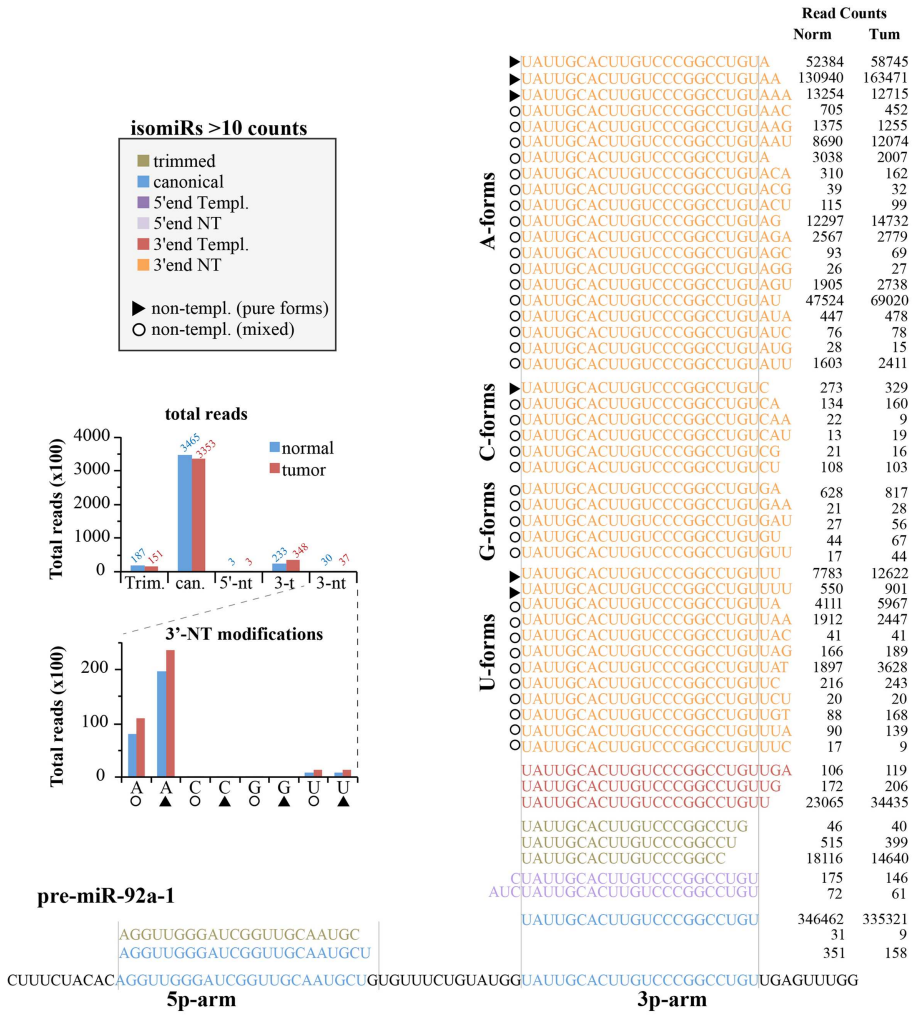


FIGURE 6 | IsomiRs from miR-92a-1 locus. Figure summarizes all the isoforms identified (>10 read counts) for the hsa-miR-92a-1 locus according to the *IsomiRage* pipeline. Isoforms are aligned over the precursor miRNA, shown at the bottom. Regions corresponding to the canonical hsa-mir-92a-5p and hsa-miR-92a-3p are highlighted in blue. Isoforms are grouped according to their type of modification (5′-end, 3′-end, trimmed forms, and canonical sequences). Isoforms perfectly pairing with the precursor miRNA (shown at the bottom) are designated as "templated" (templ.), otherwise we refer to them as "non-templated" (nt). Non-templated modifications at the 3′-end are further grouped according to the first non-templated nucleotide (A-forms, C-forms, G-forms, U-forms). As explained in the text, we could distinguish "mixed" forms (identified by circles), with different type of added nucleotides from "pure" forms (identified by triangles), which bears the same kind of nucleotide, likely as consequence of the same enzymatic activity. A bar graph summarizing the quantification of miR-92a isoforms is shown in the insert.

sequencing, but could be adapted to all the other sequencing methodologies. When applied to real samples (i.e., primary breast normal and cancer cells) *IsomiRage* almost doubled the number of aligned reads and considerably increased the number of detected miRNA species (approximately 40% more species), thus, revealing additional information "hidden" in sequencing datasets.

The identification of isomiRs is based on the alignment to a custom genome, which includes all the possible 3′- end, 5′-end, and trimmed variants for all annotated miRNAs (according to the latest miRBase release). By this approach, the pipeline is able to identify also the non-templated modifications, which are not completely matching with the pre-miRNA molecules and, therefore, missed by standard alignment procedures (that are based on perfect sequence complementarity of miRNAs to the genome or to the pre-miRNA sequence). In line with previous reports (Burroughs et al., 2010; Newman et al., 2011), a huge number of 3′-end non-templated modifications could be detected (>1 read count), several with robust expression (>100 read counts) and contributing to approximately 10–20% of the total reads of a given miRNAs. In extreme cases (11 miRNAs), the 3′-end non-templated isoform was the prevalent one. For instance, miRNAs such as miR-148b-3p, miR-152-3p, or miR-23b-3p displayed highly expressed (>1000 reads) 3′-end non-templated variants. If non-templated forms were not considered, these miRNAs might have been classified as poorly or not expressed. Given the high potential of miRNAs as molecular markers, useful in clinical studies [e.g., circulating

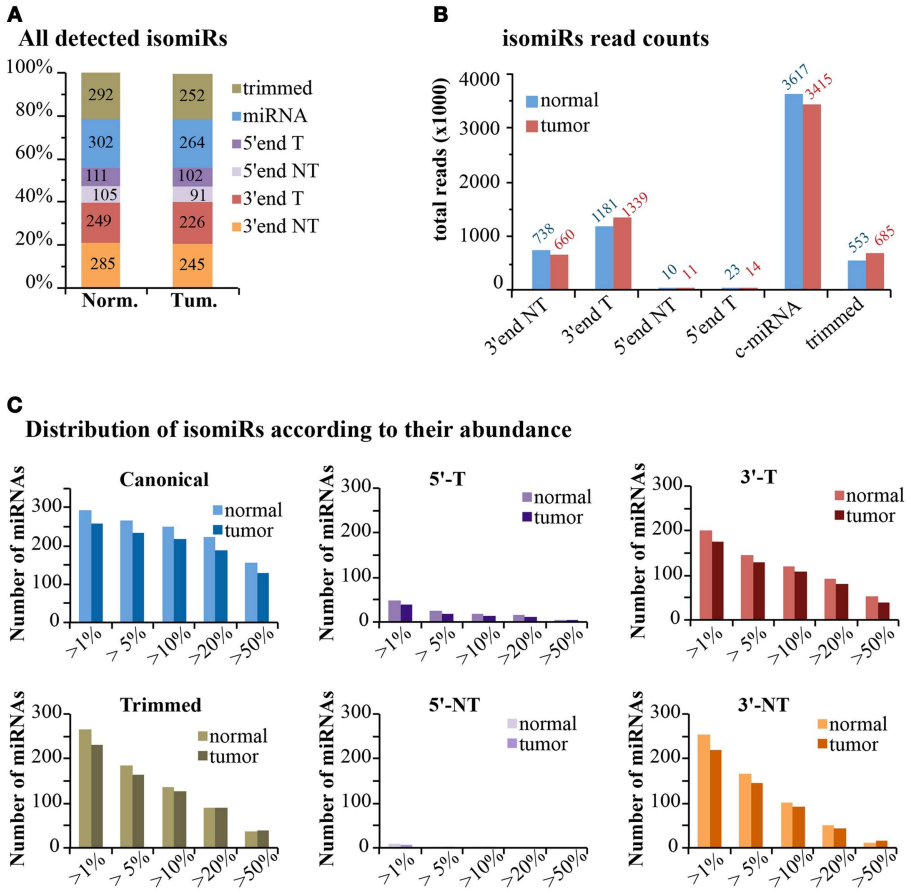


FIGURE 7 | IsomiRs distribution at genome-wide level. **(A)** The bar graph shows the percentage of detected IsomiRs (>1 read counts) in the normal and tumor samples, divided in classes, as in **Figure 6** (T, templated; NT, non-templated). The absolute number of species is also reported within the bar. **(B)** Bar graph shows the expression levels (total reads) of the each class. **(C)** Those miRNAs robustly detected (total read count for all isoforms >100 reads) were selected. Within each miRNA species, we calculated whether the selected isoform type contributes for at least a given percentage over the total reads of each miRNA.

miRNA and tumor diagnosis, reviewed in Kosaka et al. (2010)], it will be extremely relevant to consider the expression of both canonical and miRNA variants in these studies, thus, selecting the most expressed (or the most informative) variants as molecular markers.

One possible disadvantage of our approach is that we miss isomiRs that present simultaneously 5′- and 3′-end modifications or polymorphic isomiRs, which harbor substitutions in the internal nucleotide sequence due to genetic differences or epigenetic variations (i.e., editing). Since 5′-end modifications are rarely found, we could speculate that the frequency of concomitant 5′- and 3′-modifications is likely negligible. Similarly, internal variations are very sporadic, with A–I editing being the prevalent type of event and usually limited to specific miRNAs (Kawahara et al., 2007). Indeed, it is worth mentioning that it is always possible to update the custom genome, adding any other classes of modification to extend the coverage of the *IsomiRage* pipeline.

microRNA modifications are extremely heterogeneous and even a single miRNA can display a great number of similar variants (such as has-miR-92a-1, which expressed 43 different non-templated 3′-end isomiRs). Therefore, we propose grouping together functionally equivalent forms to analyze the distribution of non-templated variations at a global scale or at a miRNA-specific level. In the *IsomiRage* workflow, isomiRs are classified according to the site (5′-, 3′-end, or trimming), the origin of modification (templated or non-templated), and the nucleotide of modification (A, G, C, U). Since non-templated modification are believed to occur enzymatically through the activity of nucleotidyl-transferases (Neilsen et al., 2012), we preferred to distinguish those isoforms that bear the same type of added nucleotide ("pure" forms, likely derived from the same enzymatic activity) from those with different nucleotides ("mixed" forms, grouped on the basis of the first non-templated nucleotide).

At a global level, we found that adenylation was by far the most abundant and frequent non-templated modification, followed by uridylation (uridine is also the most frequent last nucleotide of any miRNA), in agreement with previous reports (Burroughs et al., 2010; Newman et al., 2011; Westholm et al., 2012). In plants and lower organisms, these modifications are linked to stabilization

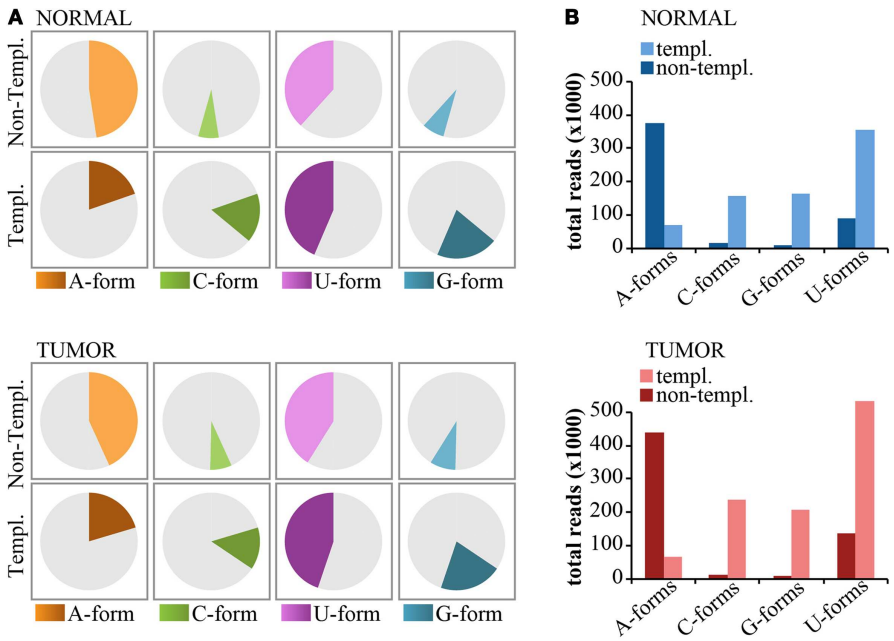


FIGURE 8 | Nucleotide distribution of 3′-end non-templated modifications. **(A)** Pie charts show the frequency of 3′ modifications, distinguished according the nucleotide type (A, G, C, U) subdivided in templated or non-templated 3′-end isomiRs. **(B)** The bar graphs show the expression levels (total read counts) of the same species as in **(A)**.

or destabilization of miRNAs, respectively (Ramachandran and Chen, 2008; Lu et al., 2009). In mammals, the functions of miRNA post-transcriptional modifications are largely unexplored, likely due to the lack of specific analytical tools. However, they could similarly have important regulatory functions, as shown for the adenylation-mediated stabilization of the liver specific miR-122 (Katoh et al., 2009). In our analysis, which was limited to just one matched tumor vs. normal sample, we did not score a global difference in the extent and the type of non-templated modification. However, when focusing on individual miRNAs, a few of them showed >5% fluctuation in the frequency of adenylated or uridylated forms in the comparison.

One relevant question is why cells have so many miRNA isoforms? As previously mentioned, most of isomiRs are templated variants, originated from imprecise processing of precursor molecules either at 5′- or at 3′-end by the processing enzymes, DGRC8 and Dicer1 (Ameres and Zamore, 2013). These variants are effectively loaded on AGO complexes and, thus, could function as canonical miRNAs (Ebhardt et al., 2009; Cloonan et al., 2011). We can just speculate on the potential usefulness of this "imprecise" machinery. One possibility is that the presence of multiple slightly different variants on the miRISC could improve miRNA functions by increasing the "on-target" to "off-target" ratio (Cloonan et al., 2011). Alternatively, variants could provide opportunity for the evolution of new miRNAs, with similar (3′-end) or different (5′-end) set of targets. For example, a change in 5′ usage might be subsequently fixed by gene duplication and by changes in the precursors miRNA transcript that affects the processing, favoring the so-called "IsomiR switching" (Wheeler et al., 2009; Tan et al., 2014).

In conclusion, using our methodology, it is possible to extend the analysis of small RNA sequencing datasets to reveal a large amount of information that lies unexplored and investigate miRNA post-transcriptional modifications. If applied to large sequencing datasets this approach could uncover the role of isomiRs in the regulation of miRNA expression and function in specific physiological and pathological contexts.

AUTHOR CONTRIBUTIONS

Heiko Muller, Matteo Jacopo Marzi, and Francesco Nicassio planned the pipeline. Heiko Muller developed the pipeline for data filtering and alignment together with the *IsomiRage* JAVA tool. Matteo Jacopo Marzi generated the custom genomes and together with Francesco Nicassio performed the analyses reported in the manuscript. Heiko Muller, Matteo Jacopo Marzi, and Francesco Nicassio wrote the manuscript.

ACKNOWLEDGMENTS

We thank Francesca Montani (European Institute of Oncology – IEO) for the production of small RNA libraries, Chiara Tordonato (European Institute of Oncology – IEO) for providing human primary samples, Luca Rotta, Thelma Capra, and Salvatore Bianchi from the Genomic Unit of IIT@SEMM for Illumina sequencing and members of the Computational Unit of IIT@SEMM for data processing. We also thank Francesco Ghini and Paola Bonetti for helpful discussion. This work was supported by grants from the Associazione Italiana per la Ricerca sul Cancro (AIRC-IG14085) and from the Umberto Veronesi Foundation (Investigator Grant 2012-2013) to Francesco Nicassio.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00038/abstract

REFERENCES

Ameres, S. L., and Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* 14, 475–488. doi:10.1038/nrm3611

Azuma-Mukai, A., Oguri, H., Mituyama, T., Qian, Z. R., Asai, K., Siomi, H., et al. (2008). Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proc. Natl. Acad. Sci. U.S.A.* 105, 7964–7969. doi:10.1073/pnas.0800334105

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

Burroughs, A. M., Ando, Y., De Hoon, M. J., Tomaru, Y., Nishibu, T., Ukekawa, R., et al. (2010). A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. *Genome Res.* 20, 1398–1410. doi:10.1101/gr.106054.110

Cloonan, N., Wani, S., Xu, Q., Gu, J., Lea, K., Heater, S., et al. (2011). microRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.* 12, R126. doi:10.1186/gb-2011-12-12-r126

Ebhardt, H. A., Tsang, H. H., Dai, D. C., Liu, Y., Bostan, B., and Fahlman, R. P. (2009). Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.* 37, 2461–2470. doi:10.1093/nar/gkp093

Gregory, R. I., Chendrimada, T. P., Cooch, N., and Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* 123, 631–640. doi:10.1016/j.cell.2005.10.022

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M., and Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* 37, W68–W76. doi:10.1093/nar/gkp347

Katoh, T., Sakaguchi, Y., Miyauchi, K., Suzuki, T., Kashiwabara, S., and Baba, T. (2009). Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes Dev.* 23, 433–438. doi:10.1101/gad.1761509

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. (2007). Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* 315, 1137–1140. doi:10.1126/science.1138050

Kosaka, N., Iguchi, H., and Ochiya, T. (2010). Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. *Cancer Sci.* 101, 2087–2092. doi:10.1111/j.1349-7006.2010.01650.x

Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* 11, 597–610. doi:10.1038/nrg2843

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., et al. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129, 1401–1414. doi:10.1016/j.cell.2007.04.040

Lee, L. W., Zhang, S., Etheridge, A., Ma, L., Martin, D., Galas, D., et al. (2010). Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA* 16, 2170–2180. doi:10.1261/rna.2225110

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Lu, S., Sun, Y. H., and Chiang, V. L. (2009). Adenylation of plant miRNAs. *Nucleic Acids Res.* 37, 1878–1885. doi:10.1093/nar/gkp031

Neilsen, C. T., Goodall, G. J., and Bracken, C. P. (2012). IsomiRs – the overlooked repertoire in the dynamic microRNAome. *Trends Genet.* 28, 544–549. doi:10.1016/j.tig.2012.07.005

Newman, M. A., Mani, V., and Hammond, S. M. (2011). Deep sequencing of microRNA precursors reveals extensive 3' end modification. *RNA* 17, 1795–1803. doi:10.1261/rna.2713611

Pece, S., Serresi, M., Santolini, E., Capra, M., Hulleman, E., Galimberti, V., et al. (2004). Loss of negative regulation by numb over notch is relevant to human breast carcinogenesis. *J. Cell Biol.* 167, 215–221. doi:10.1083/jcb.200406140

Ramachandran, V., and Chen, X. (2008). Degradation of microRNAs by a family of exoribonucleases in *Arabidopsis*. *Science* 321, 1490–1492. doi:10.1126/science.1163728

Sablok, G., Milev, I., Minkov, G., Minkov, I., Varotto, C., Yahubyan, G., et al. (2013). isomiRex: web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. *FEBS Lett.* 587, 2629–2634. doi:10.1016/j.febslet.2013.06.047

Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I. M., et al. (2014). 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.* 42, 9424–9435. doi:10.1093/nar/gku656

Westholm, J. O., Ladewig, E., Okamura, K., Robine, N., and Lai, E. C. (2012). Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs. *RNA* 18, 177–192. doi:10.1261/rna.030627.111

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

Wyman, S. K., Knouf, E. C., Parkin, R. K., Fritz, B. R., Lin, D. W., Dennis, L. M., et al. (2011). Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res.* 21, 1450–1461. doi:10.1101/gr.118059.110

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.

Received: 07 August 2014; paper pending published: 30 August 2014; accepted: 12 September 2014; published online: 29 September 2014.

*Citation: Muller H, Marzi MJ and Nicassio F (2014) IsomiRage: from functional classification to differential expression of miRNA isoforms. Front. Bioeng. Biotechnol. **2**:38. doi: 10.3389/fbioe.2014.00038*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

METHODS ARTICLE
published: 05 December 2014
doi: 10.3389/fbioe.2014.00064

Uncovering RNA editing sites in long non-coding RNAs

Ernesto Picardi [1,2] ****, Anna Maria D'Erchia*** [1,2] ***, Angela Gallo*** [3] ***, Antonio Montalvo*** [4,5] ***and Graziano Pesole*** [1,2] *******

[1] *Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Bari, Italy*
[2] *Institute of Biomembranes and Bioenergetics, Bari, Italy*
[3] *RNA Editing Laboratory, Oncohaematology Department, IRCCS Ospedale Pediatrico Bambino Gesù, Rome, Italy*
[4] *Department of Molecular Biology, Faculty of Medicine, University of Cantabria, Santander, Spain*
[5] *University Hospital Marqués de Valdecilla, Santander, Spain*

Edited by:
Alfredo Ferro, University of Catania, Italy

Reviewed by:
Thiruvarangan Ramaraj, National Center for Genome Resources, USA
Erez Levanon, Bar-Ilan University, Israel

****Correspondence:***
Ernesto Picardi and Graziano Pesole, Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Via Orabona 4, Bari 70126, Italy
e-mail: ernesto.picardi@uniba.it; graziano.pesole@uniba.it

RNA editing is an important co/post-transcriptional molecular process able to modify RNAs by nucleotide insertions/deletions or substitutions. In human, the most common RNA editing event involves the deamination of adenosine (A) into inosine (I) through the adenosine deaminase acting on RNA proteins. Although A-to-I editing can occur in both coding and non-coding RNAs, recent findings, based on RNA-seq experiments, have clearly demonstrated that a large fraction of RNA editing events alter non-coding RNAs sequences including untranslated regions of mRNAs, introns, long non-coding RNAs (lncRNAs), and low molecular weight RNAs (tRNA, miRNAs, and others). An accurate detection of A-to-I events occurring in non-coding RNAs is of utmost importance to clarify yet unknown functional roles of RNA editing in the context of gene expression regulation and maintenance of cell homeostasis. In the last few years, massive transcriptome sequencing has been employed to identify putative RNA editing changes at genome scale. Despite several efforts, the computational prediction of A-to-I sites in complete eukaryotic genomes is yet a challenging task. We have recently developed a software package, called REDItools, in order to simplify the detection of RNA editing events from deep sequencing data. In the present work, we show the potential of our tools in recovering A-to-I candidates from RNA-Seq experiments as well as guidelines to improve the RNA editing detection in non-coding RNAs, with specific attention to the lncRNAs.

Keywords: RNA editing, RNA-Seq, ncRNA, transcriptome, A-to-I editing, long non-coding RNA, lncRNA

INTRODUCTION

Massive transcriptome sequencing through high-throughput platforms has defiantly revealed that in mammals the vast majority of transcripts have little protein-coding potential (Djebali et al., 2012). Despite previous thoughts, large-scale projects like ENCODE have clearly demonstrated that more than 80% of mammalian genomes is transcribed and comprises numerous genes for non-coding RNAs (Consortium, 2012). These studies have shown that RNA is not only an essential intermediate in the flux of genetic information from DNA to proteins, but rather is a molecule involved in a plethora of fundamental cellular processes. Transfer RNAs (tRNAs) and ribosomal RNAs (rRNA), for instance, are essential components of translational machinery and highly abundant in all living cells. Small non-coding RNAs (sncRNAs) as small nuclear RNAs (snRNAs) or small nucleolar RNAs (snoRNAs) play relevant roles in alternative splicing and in guiding RNA chemical modifications (Jacquier, 2009). Additional sncRNAs as microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs) are highly conserved and associated with transcriptional and post-transcriptional gene silencing through specific base pairing with their target genes (Jacquier, 2009; Luteijn and Ketting, 2013).

Besides the different families of sncRNAs, a large proportion of the mammalian transcriptome includes RNA transcripts not coding for proteins, longer than 200 nucleotides, and defined as long non-coding RNAs (lncRNAs) (Fatica and Bozzoni, 2014). Such RNAs are poorly conserved, often polyadenylated, unstable, present in few copies and with biological roles not yet fully understood (Fatica and Bozzoni, 2014). Recent functional investigations, however, are shedding light on their functional activities and data on well-characterized lncRNAs have recently shown that such molecules have the ability to control the gene expression program at multiple levels (Wapinski and Chang, 2011). Of note, lncRNAs seem to be implicated in post-transcriptional gene regulation or in transcriptional gene silencing at epigenetic level through chromatin remodeling (Bernstein and Allis, 2005; Whitehead et al., 2009).

Virtually the entire collection of primary RNA transcripts, including the ncRNA fraction, can undergo post-transcriptional modifications as alternative splicing or RNA editing. In particular, RNA editing is widespread in the human transcriptome and involves mainly the deamination of adenosine (A) to inosine (I), recognized as guanosine (G) by all cell molecular machineries (Levanon et al., 2004). The family of adenosine deaminase acting on RNA (ADAR) proteins, characterized by the presence of double-stranded RNA binding domains (RBDs), is responsible for the deamination of specific or multiple adenosines depending on dsRNA secondary structures (Nishikura, 2010).

In human as well as in other mammals, RNA editing contributes to increase the transcriptome complexity expanding the repertoire of coding and non-coding RNAs with profound functional consequences. Indeed, RNA editing modifications may

alter codons and generate or destroy splice sites so modulating alternative splicing events and influence the dynamics of constitutive splice sites (Nishikura, 2010) with a final tuning of gene expression (Nishikura, 2010; Pullirsch and Jantsch, 2010). RNA editing is indispensable to preserve cell homeostasis and its deregulation in human has been linked to a variety of neurological/neurodegenerative disorders and cancer (Gallo and Locatelli, 2011).

In recent years, massive sequencing of RNA (RNA-Seq) has enabled the study of entire transcriptomes at single nucleotide resolution offering the unique opportunity to explore and investigate at large scale post/co-transcriptional modifications due to RNA editing (Picardi et al., 2010). Genome wide screenings in human have revealed that hundred thousands editing sites exist. Indeed the current specialized RADAR database (a comprehensive A-to-I RNA editing database) annotates over 1.4 million A-to-I changes (Ramaswami et al., 2012; Ramaswami and Li, 2014). Of these, the vast majority (~96%) is located in repetitive Alu elements (Ramaswami and Li, 2014) that comprise 11% of the human genome (having a copy number exceeding 1 million copies) and are transcribed and particularly abundant within introns and untranslated regions of mRNAs (UTRs) of RNA molecules. When located in opposite orientation, two Alu elements can fold into stable secondary structures which are a suitable target for ADAR activity (Savva et al., 2012).

Also lncRNAs are potential substrates for ADARs because of their ability to fold into specific secondary structures endowed of numerous functional properties as a consequence of their interaction with proteins or other RNAs. Indeed, lncRNAs secondary structures are quite versatile even though hard to predict by conventional computational tools. Consequently, the pattern of RNA editing could be largely dynamic making difficult investigations aimed to elucidate the final functional effects of A-to-I changes on lncRNAs.

The bioinformatic prediction of RNA editing changes by RNA-Seq data is tricky with several challenges as the discrimination of true RNA editing sites from genome-encoded SNPs and technical artifacts caused by reverse-transcription, sequencing, or read-mapping errors (Ramaswami et al., 2012). Indeed, reliable RNA editing candidates require DNA-Seq support from the same sample/individual from which RNA has been sequenced and the use of several stringent filters.

Recently, we have developed and released REDItools, a specialized bioinformatics package conceived to work with NGS data (RNA-Seq for deep RNA sequencing and DNA-Seq for massive genomic DNA sequencing) and implementing a variety of filters to provide reliable sets of RNA editing sites overcoming main sequencing biases (Picardi and Pesole, 2013). REDItools run on main unix/linux operating systems and can handle pre-aligned reads from whatever sequencing platform in the standard BAM format (they do not employ information from optional SAM/BAM fields).

In the present work, we describe a computational strategy to reliably detect A-to-I alterations in human lncRNAs through deep sequencing experiments. We apply our method to high-coverage public DNA-Seq and RNA-Seq dataset from human cell line GM12878 making use of REDItools and lncRNA transcript annotations from NON-CODEv4.1, one the most updated and comprehensive databases for lncRNAs (Xie et al., 2014).

MATERIALS AND METHODS

DATA SETS

Our workflow was tested on lymphoblastoid cell line GM12878 whose genome and RNA have been deeply sequenced. Pre-aligned DNA-Seq reads in BAM format were downloaded from the 1000 Genomes Project web page[1] and re-headed using the Picard ReplaceSamHeader.jar tool.

RNA-Seq reads, instead, were downloaded as FASTQ files from UCSC genome browser.[2] They consist of 499.4 million reads in two replicates.

QUALITY CHECK AND GENOME MAPPING OF RNA-Seq DATA

RNA-Seq quality was checked by FASTQC[3] and trimming of low quality read ends was performed by trim_galore[4] (phred cut-off was fixed to 20) excluding reads with a final length lower than 50 bases. A custom python script was used to remove reads containing low complexity regions or long stretches of unknown bases (Ns). STAR (Dobin et al., 2013) program with default parameters was used to identify reads mapping onto known rRNA annotations obtained from UCSC genome browser. Ribosomal reads were removed from next analysis step using an in house script (available upon request).

Cleaned RNA-Seq reads were aligned onto the human reference genome (hg19 assembly) using GSNAP program (main parameters were -s known-splicesites -E 1000 -n1 -Q -O --nofails -A sam --split-output = outputGsnap) providing a set of known splice sites from UCSC, RefSeq, Ensembl, and Gencode (Wu and Nacu, 2011). Unique and concordant paired-end alignments were converted to BAM format and used for downstream analyses. Duplicated reads were marked using the Picard MarkDuplicates.jar tool.

The REDItoolBlatCorrection.py script, included in the REDItools release, was applied to generate a list of reads mapping on multiple genome locations (default parameters were used).

RNA EDITING CALLING

RNA editing candidates in lncRNAs were detected using the REDItoolDnaRna.py script that is part of REDItools package (Picardi and Pesole, 2013). LncRNA transcript annotations were downloaded from NON-CODEv4 database (v4.1 including 145,331 entries) (Xie et al., 2014).

RESULTS

RNA-Seq is the *de facto* standard approach to investigate complex eukaryotic transcriptomes as well as co/post-transcriptional modifications occurring right inside. It is particularly helpful for comprehensively identifying RNA editing sites in combination with whole genome dataset to avoid false candidates due to single nucleotide polymorphisms (SNPs). Pre-aligned reads from

[1] http://www.1000genomes.org
[2] http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/
[3] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[4] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

DNA-Seq and RNA-Seq experiments constitute the input for our REDItools that implement extensive filters to mitigate sequencing biases, thus providing reliable lists of A-to-I RNA editing candidates.

WORKFLOW FOR RNA EDITING DETECTION

The main critical issue in the detection of RNA editing sites by NGS data is the mapping of RNA-Seq and DNA-Seq reads onto the reference genome that, in turn, relies on the type and quality of input data. Indeed, low quality reads lead to numerous non-canonical RNA editing sites while very short reads (<50 nucleotides) are prone to misalignments (Oshlack and Wakefield, 2009).

Before the alignment onto the reference genome, RNA-Seq reads are checked using the FASTQC program[3] that provides basic statistics about the global quality of the experiment and allows the discovery of sequencing anomalies. For example, standard RNA-Seq libraries show altered nucleotide composition (the first 6–10 read positions) due to the use of random hexamers in the library preparation. Also, RNA-Seq reads could include over-represented sequences due to adaptors, contaminants, or rRNAs not completely depleted. In addition, RNA-Seq experiments from degraded RNA may lead to high read duplication rates (Adiconis et al., 2013).

As depicted in **Figure 1**, our workflow starts with a FASTQC run to carefully check the quality of input experiments and design the next trimming step through the *trim_galore* utility[4]. Independently of FASTQC results, we removed low quality regions at 3′ ends of reads using a phred cut-off value of at least 20 and we excluded reads containing low complexity regions or long stretches of unknown nucleotides (Ns). Optionally, we add a quick step to eliminate reads showing high similarity to rRNAs by means of STAR program (Dobin et al., 2013) and custom scripts (available upon request).

After the quality assessment and an accurate data preprocessing, RNA-Seq reads are aligned onto the reference genome using GSNAP (Wu and Nacu, 2011), providing a non-redundant collection of known splice sites extracted from well-established databases as UCSC, RefSeq, Ensembl, and Gencode (Harrow et al., 2012). Although a plethora of mapping tools have been released, we preferred to use GSNAP since resulted one of the best performing aligners in a recent systematic evaluation of spliced alignment programs for RNA-Seq data (Engstrom et al., 2013). In addition, we demonstrated that realignment of RNA-Seq reads by GSNAP increased the detectability of RNA editing sites (Picardi and Pesole, 2013).

Following the mapping, GSNAP generates nine separate output files in the standard SAM format, one for each alignment type (concordant, halfmapping, paired and unpaired, etc.). Only unique and concordant alignments (in case of paired-end reads) are retained and used for downstream RNA editing calling.

An accurate detection of A-to-I editing events relies also on the type of input RNA-Seq reads. Optimal results are expected from experiments generating ultra-deep paired and stranded reads of at least 75 nucleotides. The type of RNA-Seq reads is particularly important for lncRNAs since many of them are natural antisense transcripts or produced from intronic regions of protein coding genes either in the sense or antisense direction. In addition, RNA-Seq libraries should be sequenced at high coverage since lncRNAs are generally expressed at low levels.

Although GSNAP works accurately, misalignment errors may occur. The mismapping effect can be mitigated realigning reads carrying mismatches by the classical Blat algorithm through an *ad hoc* script included in REDItools (REDItoolBlatCorrection.py). Such script identifies reads prone to mismapping and collects them in specific lists, ready to be inspected by main REDItools programs.

RNA EDITING CALLING BY REDItools

Uncovering RNA editing in lncRNAs is based on the REDItoolDnaRNA.py script in which single RNA editing modifications are identified by comparing pre-aligned RNA-Seq and DNA-Seq reads from the same sample/individual. Briefly, the script explores genomic positions site by site and applies several filters taking into account the coverage depth, the base quality score, the mapping quality, the bases supporting the variation, the type of substitution and its frequency, and changes in homopolymeric regions (≥5 bases) or in intronic sequences surrounding known splice sites. If stranded RNA-Seq data are provided, the script can infer the strand for each position mitigating biases due to antisense transcription

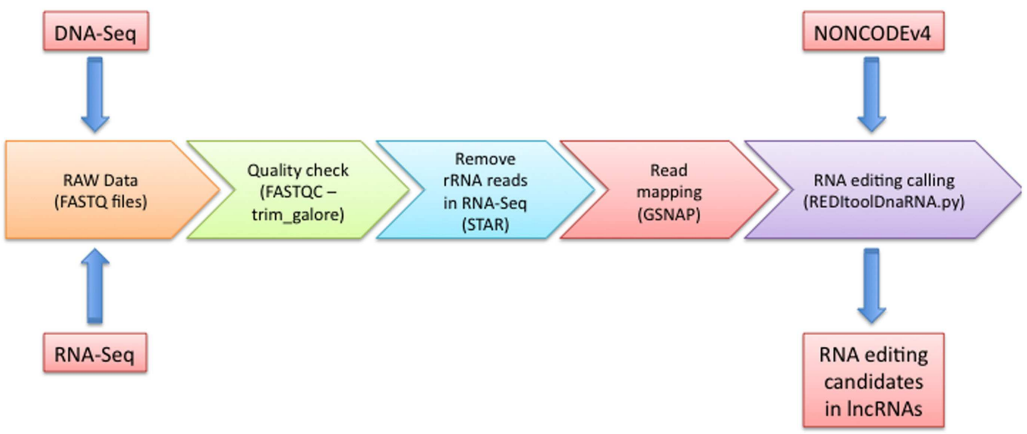


FIGURE 1 | Graphical overview of our computational methodology. In this figure, we show all steps that should be followed to predict potential RNA editing sites in human lncRNA transcripts. Details are discussed in the main text.

or mapping errors and can facilitate the A-to-I detection in lncRNAs. In the meantime, REDItoolDnaRNA.py interrogates also DNA-Seq alignments to exclude potential genomic SNPs. In addition, the script can work on specific genomic regions providing a valid set of coordinates in the GTF format.

RNA EDITING IN HUMAN lncRNAs

The above-described workflow has been applied to publicly available DNA-Seq and RNA-Seq data from human lymphoblastoid cell line GM12878. RNA-Seq data (polyA+) were obtained from the ENCODE project. Libraries were strand-specific and deeply sequenced with Illumina HiSeq2000 in two biological replicates, resulting in 235.8 and 263.7 million paired-end 76-base sequencing reads, respectively[2]. DNA-Seq data, instead, were provided as pre-aligned reads by the 1000 Genomes Project in BAM format[1]. The genomic DNA of GM12878 was sequenced at 44× coverage, allowing accurate genotype calls.

All transcriptomic reads were mapped onto the complete human genome using GSNAP in combination with a large repertoire of known splice sites. Resulting unique and concordant paired-end alignments were submitted to REDItoolDnaRNA.py as well as lncRNA transcript annotations from NON-CODE (v4.1, 145,331 entries), an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs) (Xie et al., 2014).

On the whole, we identified 11,726 potential RNA editing events supported by at least 10 DNA-Seq reads in the NON-CODE lncRNA transcript collection. Of these, we discarded only 227 positions annotated as genomic SNPs in dbSNP (release 138). The remaining 11,499 sites were annotated using the RepeatMask table from UCSC and NON-CODE transcripts (the complete list is available as Supplementary Material).

Our screen for RNA editing in lncRNAs achieved high specificity (**Figure 2**). Indeed, 97.45% of all detected changes were A-to-G mismatches while the second most frequent nucleotide substitution was T-to-C, with only 0.92% of the total number of editing sites (106/11,499). However, in 91 out of 106 T-to-C modifications the REDItoolDnaRNA.py script was not able to correctly infer the strand, most likely due to sequencing errors or concomitant expression of both strands at comparable levels. We think that several of these T-to-C changes may be genuine RNA editing events.

The majority of A-to-I modifications (86% – 9,682/11,206 unique A-to-G changes) were identified in Alu repeat regions while 1,140 resided in repetitive non-Alu regions (mostly long and short interspersed elements and long terminal repeats) and only 384 in non-repetitive regions. These findings are in accordance with other genome-wide computational screens in which a large fraction of RNA editing sites (> 90%) is located in Alu repetitive elements (Ramaswami et al., 2012; Bazak et al., 2014). The observed RNA editing pattern suggests that also in lncRNAs, Alu base pairing is predominant even though its functional role is yet elusive.

The distribution of RNA editing levels is shown in **Figure 3**. Like other previous studies, the vast majority of detected A-to-I changes showed low RNA editing levels (<0.5).

Almost all edited Alus were in intronic regions of lncRNAs while only 1913 A-to-I changes were located in exons. Excluding Alu elements, very few positions (104 sites) were found in non-repetitive regions of lncRNAs. In this reduced pool of sites, we observed several RNA editing clusters that may indicate the presence of secondary RNA structures. Such RNA editing sites may have important functional roles altering the secondary structure of lncRNAs preventing or promoting interactions with proteins or other RNAs.

The 11,206 unique A-to-G changes fell in 1649 lncRNA gene loci (3374 lncRNA transcripts) and, of these, a substantial number occurred in intervening sequences. According to the NON-CODE

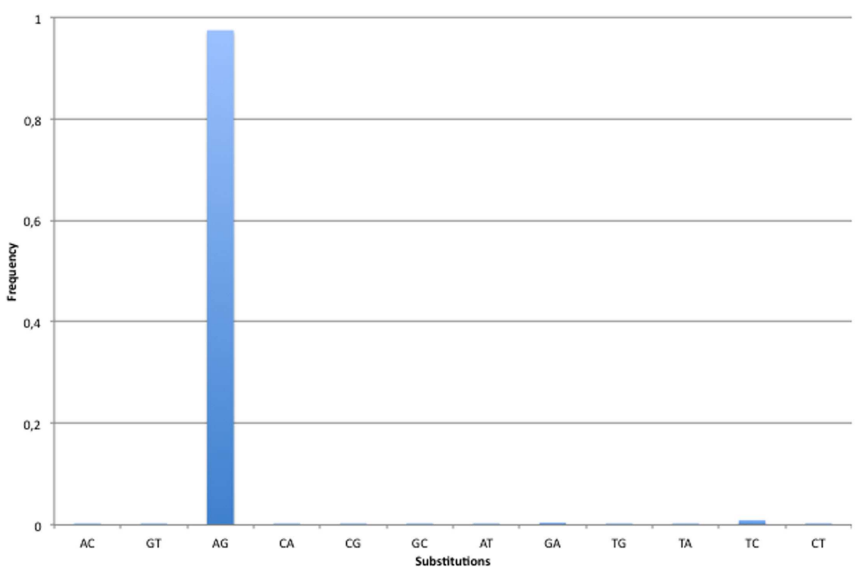


FIGURE 2 | Base substitutions observed in human lncRNAs. The specificity of our methodology has been valuated looking at base substitutions in the set of predicted RNA editing events. Since A-to-I is the most frequent RNA editing event in human and I is commonly interpreted as G by cellular molecular machineries, the A-to-G change is expected to be the prominent substitution. As shown in figure, 97% of base changes in the predicted set of RNA editing events are A-to-G substitutions. All other changes have substitution frequencies lower than 1%.

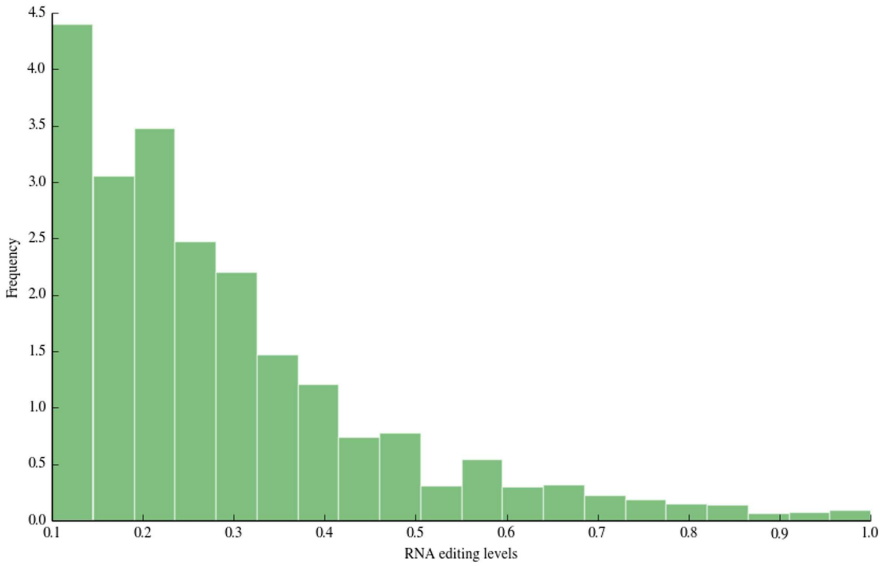


FIGURE 3 | RNA editing levels. In this figure, we depict the distribution of RNA editing levels. The vast majority of detected sites show low editing levels (<0.5), in accordance with previous large-scale studies.

database in which lncRNA genes are classified into four categories depending on their genomic location in relation to protein-coding genes (antisense, intergenic, sense exonic, and sense non-exonic), we valuated the distribution of edited lncRNA genes among these four categories. A consistent amount (62%) of lncRNA genes was predominantly in the sense exonic category and only 51 (3%) belonged to sense non-exonic grouping. The number of lncRNA genes cataloged as antisense and intergenic was roughly equivalent, being 267 and 289, respectively.

We finally compared our list of RNA editing changes with that identified in a previous study based on the same NGS dataset by using a slighting different methodology (Ramaswami et al., 2012). We found overlap for 10,898 sites (97%) indicating high specificity of our computational strategy and improved sensitivity over past bioinformatic methods.

DISCUSSION

Large-scale projects, such as the ENCODE (Encyclopedia of DNA Elements), have markedly revealed the pervasiveness of genome transcription (Consortium, 2012). Nearly 60% of human genome encodes transcripts that lack protein-coding capacity but with a potential role in multiple biological processes (Djebali et al., 2012). Among them, a particular attention has focused on a class of transcripts indicated as lncRNAs, generally defined as RNAs longer than 200 nucleotides (Fatica and Bozzoni, 2014). Although lncRNAs are poorly conserved, unstable, and present in few copies, they have been implicated in transcriptional regulation of protein-coding gene (Fatica and Bozzoni, 2014).

In addition to transcriptional complexity of eukaryotic genomes, the transcriptome landscape is further complicated by co/post-transcriptional mechanisms as alternative splicing and RNA editing (Djebali et al., 2012; Bazak et al., 2014). In particular, RNA editing may play relevant biological roles also at level of lncRNAs (Mallela and Nishikura, 2012). In human, the majority of RNA editing modifications is constituted by A-to-I conversions carried out by the ADAR enzymes (Ramaswami and Li, 2014). These proteins have the ability to target secondary RNA structures and deaminate specific adenosines located inside (Nishikura, 2010). Due to their secondary structures, lncRNAs are expected to be potential targets of ADARs with specific functional effects such as preventing or promoting interactions with proteins or other RNAs. The importance of studying RNA editing modifications in lncRNAs is mainly justified in pathological conditions in which editing events may be connected with alteration of lncRNA expression/function.

Nowadays lncRNAs and RNA editing can be profiled at single nucleotide resolution through NGS technologies (Picardi et al., 2010; Ramaswami et al., 2012; Ding et al., 2014). The massive transcriptome sequencing, indeed, facilitates the identification of lncRNAs as well as the detection of putative RNA editing events (Picardi et al., 2010). However, the computational prediction of RNA editing changes by RNA-Seq is not trivial due to technical artifacts (sequencing or read-mapping errors) and genomic information from same samples/individuals is required to discriminate true RNA editing sites from SNPs (Ramaswami et al., 2012).

To uncover the RNA editing landscape using NGS data, we have recently developed the package REDItools that includes specific scripts to investigate RNA editing starting from matched RNA-Seq and DNA-Seq data or RNA-Seq data alone (Picardi and Pesole, 2013). In the present work, we introduce a computational methodology devoted to the detection of RNA editing events in human lncRNAs, demonstrating in the meantime the suitability of our REDItools as a versatile package for screening RNA editing candidates in NGS data.

We tested our pipeline on DNA-Seq and RNA-Seq data from human lymphoblastoid cell line GM12878 using 145,331

lncRNA transcripts from NON-CODE database (Xie et al., 2014). Compared with previous computational pipelines (Ramaswami et al., 2012), our methodology achieved high specificity and improved sensitivity, as already shown in Picardi and Pesole (2013). Indeed, more than 97% of detected RNA editing changes were A-to-G mismatches mainly distributed in Alu repeated regions.

The majority of edited lncRNA genes were in the sense exonic category meaning that RNA editing target lncRNA genes were in overlap with known protein coding genes and in the same orientation. In such cases, since lncRNAs and overlapping protein coding transcripts share the same strand, the assessment of RNA editing membership, lncRNA or coding transcript, is very hard. Further checks taking into account the expression levels of involved genes and transcripts are extremely needed before claiming novel discoveries.

Although the computational detection of RNA editing events in NGS data is not yet completely optimized, our REDItools are the only available software to explore the RNA editing landscape in complete transcriptomes. Given the explosion of NGS technologies in genomic research, REDItools and derived methodologies, as the one described in this work, will be indispensable to characterize RNA editing in novel experimental conditions as well as in human disorders.

ACKNOWLEDGMENTS

This work was supported by the Italian Ministero dell'Istruzione, Università e Ricerca (MIUR): PRIN 2009 and 2010; Consiglio Nazionale delle Ricerche: Flagship Project Epigen, Medicina Personalizzata, and Aging Program 2012–2014. This work was also supported by the Italian Ministry for Foreign Affairs (Italy–Israel actions) to Ernesto Picardi and AIRC to Angela Gallo.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00064/abstract

REFERENCES

Adiconis, X., Borges-Rivera, D., Satija, R., Deluca, D. S., Busby, M. A., Berlin, A. M., et al. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* 10, 623–629. doi:10.1038/nmeth.2483

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., et al. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* 24, 365–376. doi:10.1101/gr.164749.113

Bernstein, E, and Allis, C. D. (2005). RNA meets chromatin. *Genes Dev* 19, 1635–1655. doi:10.1101/gad.1324305

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Ding, X., Zhu, L., Ji, T., Zhang, X., Wang, F., Gan, S., et al. (2014). Long intergenic non-coding RNAs (LincRNAs) identified by RNA-seq in breast cancer. *PLoS ONE* 9:e103270. doi:10.1371/journal.pone.0103270

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi:10.1038/nature11233

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Engstrom, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Alioto, T., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191. doi:10.1038/nmeth.2722

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi:10.1038/nrg3606

Gallo, A., and Locatelli, F. (2011). ADARs: allies or enemies? The importance of A-to-I RNA editing in human disease: from cancer to HIV-1. *Biol. Rev. Camb. Philos. Soc.* 87, 95–110. doi:10.1111/j.1469-185X.2011.00186.x

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* 22, 1760–1774. doi:10.1101/gr.135350.111

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* 10, 833–844. doi:10.1038/nrg2683

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

Luteijn, M. J., and Ketting, R. F. (2013). PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nat. Rev. Genet.* 14, 523–534. doi:10.1038/nrg3495

Mallela, A., and Nishikura, K. (2012). A-to-I editing of protein coding and non-coding RNAs. *Crit. Rev. Biochem. Mol. Biol.* 47, 493–501. doi:10.3109/10409238.2012.714350

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* 79, 321–349. doi:10.1146/annurev-biochem-060208-105251

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4, 14. doi:10.1186/1745-6150-4-14

Picardi, E., Horner, D. S., Chiara, M., Schiavon, R., Valle, G., and Pesole, G. (2010). Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. *Nucleic Acids Res.* 38, 4755–4767. doi:10.1093/nar/gkq202

Picardi, E., and Pesole, G. (2013). REDItools: high-throughput RNA editing detection made easy. *Bioinformatics* 29, 1813–1814. doi:10.1093/bioinformatics/btt287

Pullirsch, D., and Jantsch, M. F. (2010). Proteome diversification by adenosine to inosine RNA editing. *RNA Biol.* 7, 205–212. doi:10.4161/rna.7.2.11286

Ramaswami, G., and Li, J. B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 42, D109–D113. doi:10.1093/nar/gkt996

Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C., and Li, J. B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods* 9, 579–581. doi:10.1038/nmeth.1982

Savva, Y. A., Rieder, L. E., and Reenan, R. A. (2012). The ADAR protein family. *Genome Biol.* 13, 252. doi:10.1186/gb-2012-13-12-252

Wapinski, O., and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends Cell Biol* 21, 354–361. doi:10.1016/j.tcb.2011.04.001

Whitehead, J., Pandey, G. K., and Kanduri, C. (2009). Regulation of the mammalian epigenome by long noncoding RNAs. *Biochim Biophys Acta* 1790, 936–947. doi:10.1016/j.bbagen.2008.10.007

Wu, T. D., and Nacu, S. (2011). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26, 873–881. doi:10.1093/bioinformatics/btq057

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., et al. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.* 42, D98–D103. doi:10.1093/nar/gkt1222

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.

Received: 10 October 2014; paper pending published: 27 October 2014; accepted: 21 November 2014; published online: 05 December 2014.

*Citation: Picardi E, D'Erchia AM, Gallo A, Montalvo A and Pesole G (2014) Uncovering RNA editing sites in long non-coding RNAs. Front. Bioeng. Biotechnol. **2**:64. doi: 10.3389/fbioe.2014.00064*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

METHODS ARTICLE
published: 10 December 2014
doi: 10.3389/fbioe.2014.00069

Comprehensive reconstruction and visualization of non-coding regulatory networks in human

Vincenzo Bonnici[1], Francesco Russo[2,3], Nicola Bombieri[1], Alfredo Pulvirenti[4]† and Rosalba Giugno[4]*†*

[1] *Department of Computer Science, University of Verona, Verona, Italy*
[2] *Laboratory of Integrative Systems Medicine (LISM), Institute of Informatics and Telematics (IIT) and Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy*
[3] *Department of Computer Science, University of Pisa, Pisa, Italy*
[4] *Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy*

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Thomas Triplet, Wajam, Canada
Zhi-Ping Liu, Shandong University, China

****Correspondence:***
Alfredo Pulvirenti and Rosalba Giugno, Department of Clinical and Molecular Biomedicine, University of Catania, Viale Andrea Doria 6, Catania 95125, Italy
e-mail: apulvirenti@dmi.unict.it; giugno@dmi.unict.it

†Alfredo Pulvirenti and Rosalba Giugno have contributed equally to this work.

Research attention has been powered to understand the functional roles of non-coding RNAs (ncRNAs). Many studies have demonstrated their deregulation in cancer and other human disorders. ncRNAs are also present in extracellular human body fluids such as serum and plasma, giving them a great potential as non-invasive biomarkers. However, non-coding RNAs have been relatively recently discovered and a comprehensive database including all of them is still missing. Reconstructing and visualizing the network of ncRNAs interactions are important steps to understand their regulatory mechanism in complex systems. This work presents *ncRNA-DB*, a NoSQL database that integrates ncRNAs data interactions from a large number of well established on-line repositories. The interactions involve RNA, DNA, proteins, and diseases. ncRNA-DB is available at http://ncrnadb.scienze.univr.it/ncrnadb/. It is equipped with three interfaces: web based, command-line, and a Cytoscape app called ncINetView. By accessing only one resource, users can search for ncRNAs and their interactions, build a network annotated with all known ncRNAs and associated diseases, and use all visual and mining features available in Cytoscape.

Keywords: microRNAs, lncRNAs, non-coding RNAs, networks, cytoscape, gene expression

1. INTRODUCTION

After the sequencing of the human genome, it became evident that only 20,000 genes are protein-coding, while over 98% of all genes are untranslated non-protein-coding RNAs (ncRNAs) (ENCODE Project Consortium, 2012). During the last years, thousands of ncRNAs have been identified in the eukaryotic transcriptome (Khalil et al., 2009; Bu et al., 2011). Usually, ncRNAs are divided into two groups according to their length: short ncRNAs, consisting of <200 nucleotides, and long non-coding RNAs (lncRNAs), whose size ranges from 200 nucleotides up to 100 kb (Mattick, 2001).

The microRNAs (miRNAs) family is the best known class of short ncRNAs. They regulate gene expression and contribute to development, differentiation and are responsible of carcinogenesis. The aberrant expression or alteration of miRNAs also contributes to many of human pathologies, including cancer (Lu et al., 2005). Moreover, a significant amount of miRNAs has been found in extracellular human body fluids (Mitchell et al., 2008; Hanke et al., 2010) and some circulating miRNAs in the blood have been successfully revealed as biomarkers for several diseases including cardiovascular malfunctions (Gupta et al., 2010b) and cancer (Mitchell et al., 2008).

An emerging class of ncRNAs consists of lncRNAs (Fatica and Bozzoni, 2014) They are both nuclear and cytoplasmic. Nuclear lncRNAs function by guiding chromatin modifiers to specific genomic loci (Rinn and Chang, 2012; Batista and Chang, 2013; Guttman and Rinn, 2012; Khalil et al., 2009; Tay et al., 2011) while many others have been identified in the cytoplasm (Batista and Chang, 2013). These lncRNAs are involved in gene regulation and often show sequence complementarity with transcripts that originate from either the same chromosomal locus or independent loci.

One of the most recently discovered and not yet functionally characterized class is the circular RNA (circRNAs) (Memczak et al., 2013) Numerous circRNAs form by head-to-tail splicing of exons, suggesting previously unrecognized regulatory potential of coding sequences. Recent results (Memczak et al., 2013) have shown that thousands of well-expressed stable circRNAs have both tissue and developmental-stage specific expression. Moreover, human circRNAs are bound by miRNAs such as the miR-7 showing a potential role of circRNAs as post-transcriptional regulators.

Understanding the complex system derived from the interactions of regulators and possible targets gives a clue on the dynamics and causes of disorders (Couzin, 2007). In this direction, platforms to visualize networks such as Cytoscape (Shannon et al., 2003) together with tools to visualize and analyze them are becoming crucial in systems biology studies.

miRScape (Ferro et al., 2009) is one of the first Cytoscape plug-in visualizing protein–protein interaction networks annotated with miRNAs. It uses a web knowledge base (Laganà et al., 2009) to infer associations between genes and phenotypes though miRNAs. CyTargetLinker (Kutmon et al., 2013) is a recent Cytoscape app that builds biological networks annotated with miRNAs, transcription factors, and drugs.

Several methodologies are designed to analyze the regulatory effect of miRNAs and transcription factors in protein-coding genes (Liu et al., 2009, 2014; Sales et al., 2010; Huang et al., 2011; Laczny et al., 2012; Le et al., 2013; Guo et al., 2014). Some of them export the results also in a Cytoscape network format. For example, Magia (Sales et al., 2010) allows to perform statistical analysis on miRNAs and gene expressions. TSmir (Guo et al., 2014) browses regulatory network of tissue-specific miRNAs with transcriptor factors. mirConnX (Huang et al., 2011), given a network of genes, transcriptor factors, and miRNAs, extends it with further TF and miRNA–gene intersections inferred by user expression data. miRTrail (Laczny et al., 2012) analyzes the role of miRNAs and genes deregulated in a disease by using a miRNA-gene networks and expression data.

In this work, we have imported and integrated associations among non-coding RNAs (miRNAs, circulating miRNAs, lncRNAs, and other non-coding), genes, RNAs, and associated diseases from 10 on-line databases. The database, named non-coding RNA Human Interaction Data Base (ncRNA-DB), is built on top of the NoSQL platform OrientDB. It is kept updated by common semi-automated procedures. The interaction data of ncRNA-DB can be simply searched and visualized by a web based or a command-line interface. The database is accessible through a Cytoscape app, called ncINetView, which allows to: (i) build a network annotated with all known ncRNAs and associated diseases by accessing to only one database, and (ii) use all visual and mining features available in Cytoscape app store to analyze it. At http://ncrnadb.scienze.univr.it/ncrnadb/, users can search in ncRNA-DB, export the results in text format, download the command-line interface, Java API, the app ncINetView, and use ncRNA-DB as server for third party client applications.

2. CONSTRUCTION AND CONTENT

2.1. DATA SOURCE

Non-coding RNA human interaction data base integrates data from several state of the art non-coding databases. We selected sources that cover the majority of non-coding RNAs information with high quality and updated data. Moreover, this first version of ncRNA-DB focuses on databases of known interactions between non-coding RNAs and mRNAs. We discarded non-coding RNAs with unknown interactions such as piRNAs (RNA Piwi-interacting). In the following subsections, we give an overview of data sources in ncRNA-DB. **Table 1** summarizes the numbers of integrated data and how many are shared among datasources.

2.1.1. Nomenclature of non-coding RNAs

In ncRNA-DB, we used The HUGO Gene Nomenclature Committee (HGNC) as official database of approved names and aliases. HGNC is responsible for approving unique symbols and names for human loci, including protein-coding genes, ncRNA genes, and pseudogenes, to allow unambiguous scientific communication (Gray et al., 2012)[1].

2.1.2. Long non-coding RNAs databases

In this work, we selected several lncRNAs databases that provide a central repository of known lncRNAs, their aliases, and published characteristics. lncRNAdb (Amaral et al., 2011) is one of them and it is available online at http://www.lncrnadb.org.

Another database is The LncRNADisease (Chen et al., 2013)[2]. It is a resource for the experimentally supported LncRNA-disease association data. The platform integrates also tools for predicting novel LncRNA-disease associations. Moreover, LncRNADisease contains lncRNA interactions at various levels, including proteins, RNAs, miRNAs, and DNA.

We also included general non-coding databases such as NONCODE[3], which is a database of all kinds of non-coding RNAs (except tRNAs and rRNAs) containing 210,831 lncRNAs of several species (Bu et al., 2011).

Table 1 | The number of imported elements from external resources and how many among them are present at least in another datasource.

DataSource	Number of entities	Shared
CIRC2TRAITS	83,432	326
HMDD.2	8,040	282
LNCRNADISEASE	1,505	244
MIRANDOLA.1.6 2246	98	
NPINTER.2.0	138,328	440
MIRTARBASE	40,532	218
STARBASE.V2.0	31,463	8

This representation of shared notation is dictated by the fact that the number of elements shared in three or more datasources is approximately close to 0.

2.1.3. Circular RNAs database

Circ2Traits[4] is a comprehensive database for circRNA potentially associated with diseases and traits (Ghosal et al., 2013) circRNAs, formed by covalent linkage of the ends of a single RNA molecule, are newly discovered RNAs that sponge miRNAs to block their function (Memczak et al., 2013). Circ2Traits uses the circRNA dataset from Memczak et al. (2013). This dataset consists of 1,953 predicted human circRNAs along with their genomic coordinates, annotation, and predicted miRNA seed matches. The disease related miRNA data are taken from miR2disease (Jiang et al., 2009). The authors collect the miRNA–mRNA interaction data predicted by miRanda (Betel et al., 2008), TargetScan (Lewis et al., 2005), PiTA (Kertesz et al., 2007), PicTar (Krek et al., 2005), and RNA22 (Loher and Rigoutsos, 2012). Moreover, a dataset of predicted miRNA and lncRNA interaction pairs is collected from the miRCode database (Jeggari et al., 2012).

2.1.4. microRNA databases

Non-coding RNA human interaction data base includes The Human microRNA Disease Database (HMDD) (Li et al., 2013), a database of curated experiment-supported evidence for human miRNAs and disease associations[5]. The database contains detailed

[1]http://genenames.org

[2]http://210.73.221.6/lncrnadisease
[3]http://www.noncode.org/
[4]http://gyanxet-beta.com/circdb/
[5]http://www.cuilab.cn/hmdd

and comprehensive annotations of human miRNA-disease associations, including those from the evidence of genetics, epigenetics, circulating miRNAs, and miRNA-target interactions.

Another important resource is the miRandola database (Russo et al., 2012, 2014)[6]. It is a manually curated database of extracellular circulating miRNAs. It is a comprehensive classification of different extracellular miRNA types and a collection of non-invasive biomarkers for several diseases (e.g., cancer and cardiovascular diseases).

2.1.5. Interaction databases

We included several sources for non-coding RNAs interactions. The miRTarBase database (Hsu et al., 2014)[7] provides experimentally validated miRNA-target interactions.

NPInter (Wu et al., 2006)[8] reports functional interactions between non-coding RNAs (except tRNAs and rRNAs) and biomolecules (proteins, RNAs, and DNA), which are experimentally verified. The authors collected primarily physical interactions, although several interactions of other forms are also included. Interactions are manually collected from publications, followed by an annotation process that uses known databases including NONCODE (Bu et al., 2011), miRBase (the miRNA registry) (Kozomara and Griffiths-Jones, 2013), and UniProt (the database of proteins) (UniProt Consortium, 2013).

starBase (Li et al., 2014)[9] reports RNARNA and proteinRNA interactions from 108 CLIP-Seq (PAR-CLIP, HITS-CLIP, iCLIP, and CLASH) about 37 independent studies. The database contains about 9,000 miRNA–circRNA, 16,000 miRNA–pseudogene, and 285,000 protein–RNA relations. It also contains predicted miRNA–mRNA and miRNA–lncRNA interactions.

2.2. DATA SCHEMA

2.2.1. ncRNA-DB identifier

Public databases catalog biological entities (e.g., ncRNAs) via nomenclatures. They can be human readable names or alphanumeric identifiers. For example, genes are classified by their names, their symbols, or database-specific identifiers. For example, the *breast cancer 1* gene can be identified by its assigned symbols BRCA1, BRCC1, and PPP1R53, or by its specific database identifiers like HGNG:1100, Entrez Gene 672, and UCSC uc002ict.3.

Non-coding RNAs have been relatively recently discovered and a comprehensive database including all of them is still missing. The non-coding RNAs knowledge is spread among several databases and ambiguity on the identifiers exists. Moreover, new discovered entities are named with internal identifiers and they are not reported in any other databases. This is the case for example of NONCODE v4, the largest collection of ncRNAs available online, where most of the reported ncRNAs can be only mapped to NONCODE.

In ncRNA-DB, we use a generic resource identifier system (named RID) together with a unique system-scope identifier assigned by OrientDB (called ORID).

The RID is composed by three parts (or levels) *EntityType: DataSource: Alias.name*. The EntityType indicates the biological classification of the element such as ncRNA, RNA (not including ncRNA), Gene, Disease, and Others for all other cases including entities with unspecified type in the original data source. The DataSource reports the name of the external data source from where we got the data together with its version (e.g., HMDD_2). The Alias.name represents the nomenclature used in the data source.

2.2.2. Graph database schema for ncRNA-DB

A set of biological entities (genes, ncRNAs, RNAs, and diseases) and their relations (physical interactions, functional relations, and so on) can be modeled as a graph, a mathematical object composed by nodes (entities) and edges (relations).

Relational database management systems (RDMS) are widely used to store biological data. However, new rising models, grouped under the name of NoSQL (Not only SQL) databases (Stonebraker, 2010; Han et al., 2011), are becoming quite popular for web and biological applications. They can provide schema-less representation for non-structured data and can be easily implemented in a distributed fashion resulting effective for Big Data problems (Cattell, 2011).

NoSQL system can be classified into four classes, even if some of them belongs to more than one class: (i) column model, where data are represented by tuples, (ii) document-oriented databases for storing, retrieving, and managing document-oriented information, also known as semi-structured data, (iii) key-value store, where data are stored as a collection of key-value pairs stored using associative arrays, maps, symbol tables, or dictionaries, and (iv) graph databases, where data are modeled using a graph structure. These often implement the object-oriented model by modeling concepts like classes, instances, inheritance, and polymorphism.

Non-coding RNA human interaction data base is implemented in OrientDB (Tesoriero, 2013) OrientDB is both a graph model and an object-oriented model, on top of a document model. We chose OrientDB since it is a graph database and its object-oriented concepts are suitable to model the ncRNA-DB data. Furthermore, the use of OrientDB allows to give public accesses to our server, effective management of user privileges, use graph traversal procedures, and language bindings among a large choice. It offers a SQL-like interface in addition to several language specific interfaces. It is developed in Java and provides native Java API (Application Programing Interface) for accessing the database, which is suitable for developing Cytoscape applications.

Figure 1 depicts the schema of ncRNA-DB. The abstract class *BioEntity* represents biological entities and it is specialized in the five sub-classes: ncRNA, RNA, Gene, Disease, and Others. Aliases are represented by the abstract class *Alias*, which is specialized in five different sub-classes related to the five entity types. DataSource is a class containing the external resource name and version from where the data are got or equivalently the official repository of the entity (e.g., NONCODE v4). An instance of a class is a particular value (e.g., realization, element, and data). In a graph model, instances of classes and sub-classes are nodes.

Class inheritance happens when a class is a specialization of the other one (**Figure 1** mark 1). The naming of a biological entity by an alias is represented by adding an edge between the

[6]http://atlas.dmi.unict.it/mirandola/
[7]http://mirtarbase.mbc.nctu.edu.tw/
[8]http://www.bioinfo.org/NPInter/
[9]http://starbase.sysu.edu.cn/

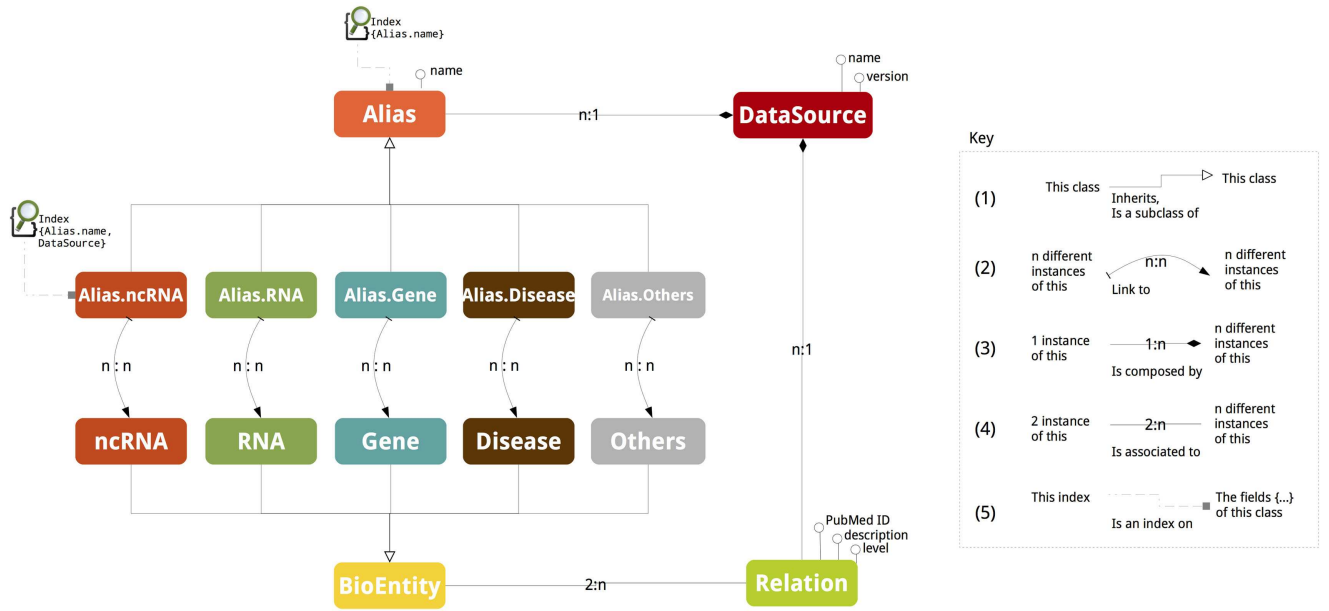


FIGURE 1 | The database schema. The picture reports all the stored information together with their associations.

corresponding graph nodes. Due to the ambiguity of nomenclatures, these edges are *n*:*n* cardinality (**Figure 1** mark 2). This means that, for example, an ncRNA can have different aliases and the same alias can refer to different ncRNAs.

Interactions among entities are modeled through a class called *Relation* associated to the class *BioEntity* (**Figure 1** mark 3). The cardinality of the association is *n*:2, since an entity participates at more than one relation and a relation involves exactly two entities.

The attributes of *Relation* are the *PubMed ID* containing the reference of article reporting such relation, the *description* with the support sentences and *level* to store the interaction level. The interaction level indicates the molecular strata where the interaction is realized. This is represented by a pair of strings $(a - b)$ with a and b belonging to (RNA, DNA, Protein, TF). For example, RNA–TF specifies that the ncRNA is interacting with the transcription factor of the gene; (RNA–DNA) indicates that the ncRNA is interacting with the coding genomic region of the gene; (RNA–Protein) describes that the ncRNA is interacting with the protein structure; and (RNA–RNA) tells that ncRNA is interacting with the transcript RNA. If the same relation is stored in two (or more) distinct data sources, two (or more) interaction edges are stored into our system. This choice is motivated by reporting for each interaction specific information such as the support sentence. The *level* is the string NA when this detail in the resource is not given.

When a class contains as field values of another class we indicate that a composition relation exists (**Figure 1** mark 4). For example, a data source name and version is part of a RID, which represents an Alias. The *Relation* has a composition association with *DataSource* to external databases reporting it. The cardinality of composition relation is *n*:1 since an alias or a relation is reported in a data source and a data source contains more than one relation or alias.

Table 2 | The total number of aliases associated with the imported elements from external resources and how many among them are present at least in another datasource.

DataSource	Number of aliases	Shared
HGNC	436,361	19,368
NONCODE.V4	327,099	5,671
LNCRNADB	218	115
CIRC2TRAITS	16,730	1,076
HMDD.2	1,376	1,376
LNCRNADISEASE	1,366	285
MIRANDOLA.1.6	1,231	1,231
NPINTER.2.0	7,678	4,857
MIRTARBASE	62,207	12,998
STARBASE.V2.0	5,298	3,747

Aliases act as access points to the data and they are indexed (**Figure 1** mark 5). The abstract class *Alias* is indexed by a single field not-unique map on the element nomenclature (the third field of the RID, Alias.name). This is used when the search is performed by giving only the nomenclature. The *Alias.type* subclasses are indexed by a composite key dictionary working on the second and third field of the RID, DataSource, and Alias.name. This index works when both the EntityType and the nomenclature are specified.

2.2.3. Data import

Here, we give details on the imported data from each resource. ncRNA-DB integrates data concerning only *Homo sapiens*.

- HGNC: we imported a list of non-coding RNAs and their approved aliases used by other datasources, protein-coding genes, pseudogenes, and phenotypes (considered as diseases).
- lncRNAdb: we imported a list of non-coding RNAs and their aliases.
- circ2traits: we imported a set of interacting lncRNAs, circRNAs, and messenger RNAs together with the associated diseases and the PubMed IDs of articles where the interactions are reported.
- HMDD: we imported a list of diseases, the set of genes that interact with ncRNAs, PubMed IDs of articles together with the support sentences. Here, interactions are listed as ncRNA-disease or ncRNA-gene-disease. We split the multi-relation ncRNA-gene-disease into two distinct relations ncRNA-gene and ncRNA-disease.
- LncRNAdisease: we imported a list of lncRNAs, their aliases, associated diseases, interaction levels, PubMed IDs of articles supporting the interactions, and sentences describing details such as the type of dysfunction.
- Mirandola: we imported a set of miRNAs, their aliases, PubMed IDs of articles together with the support sentences.
- miRTarBase: we imported a set of miRNAs, their validated targets, and their aliases, PubMed IDs of articles together with the support sentences.
- NONCODE: we imported a list of non-coding RNAs, their aliases and a mapping of NONCODE into external identifiers.
- NPInter: we imported a set of ncRNAs, their interactions, interaction levels, PubMed IDs of referencing articles, and supporting sentences.

Table 3 | For each biological entity type we report the number of entries present in ncRNA-DB.

Entity	Total	In relation
ncRNA	193,440	25,463
RNA	4,962	4,962
Gene	19,271	12,265
Disease	1,330	735
Others	6,700	5,517

*We report also the number of entities having relations with some other entities (details are given in **Table 4**).*

Table 4 | The number of ncRNAs interacting with other ncRNA-DB biological entities.

Relation	Total
ncRNA-ncRNA	77,982
ncRNA-RNA	36,369
ncRNA-gene	52,611
ncRNA-disease	16,662
ncRNA-others	132,663

From the integrated data source files, we extracted the following fields: source, target, and interaction details such as interaction levels, reference papers, and support sentences. The main issues about importing data from several resources are aliases disambiguation and the missing of entity type classification. In a first phase, we extracted and combined from HGNC, NONCODE, and LNCRNADB the sets of aliases for each bioentity. At the end of this step, each bioentity will have some aliases uniquely assigned to it, and some others shared with other entities. In a second phase, for each entity we merge its aliases with those taken from all other datasets integrated in ncRNA-DB. **Table 2** summarizes the number of aliases taken from the integrated datasources and how many are shared among them.

When the entity type of interaction actors are not provided, but only the entity levels (i.e., RNA-protein), we first searched the elements in the ncRNA-DB. If they were not present in any

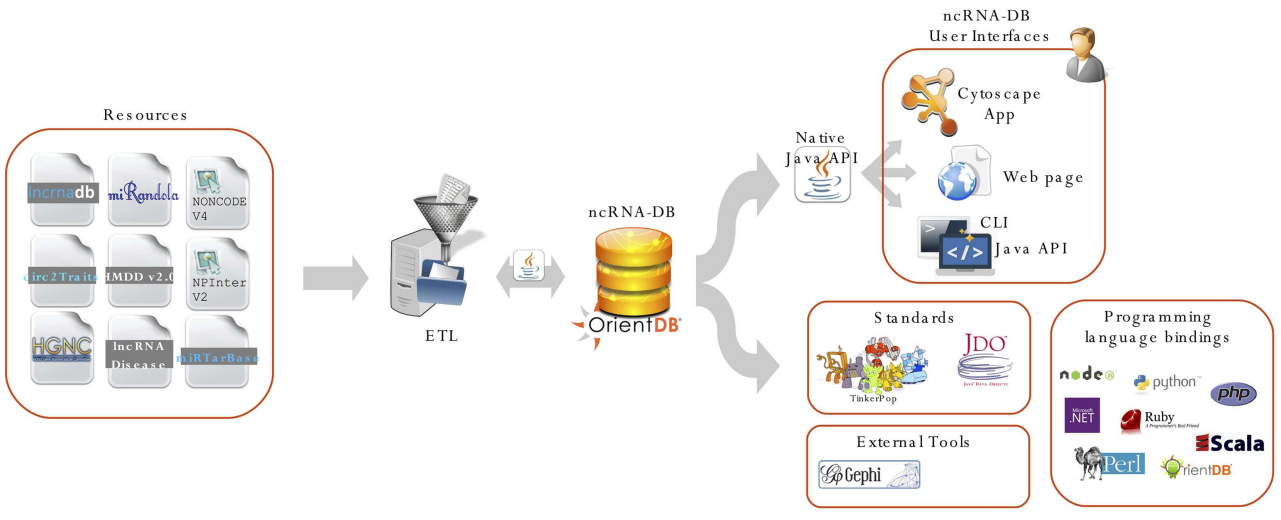


FIGURE 2 | We report the main architecture of our system. The resource is integrated through an import procedure and stored into OrientDB. All the data are exposed using three different user interfaces: (i) the nclNetView; Cytoscape app; (ii) the web app; and (iii) the command-line interface. All the data can be also queried using languages APIs and third party applications. ncRNA-DB is designed to be a server for client applications, thanks to the features offered by OrientDB.

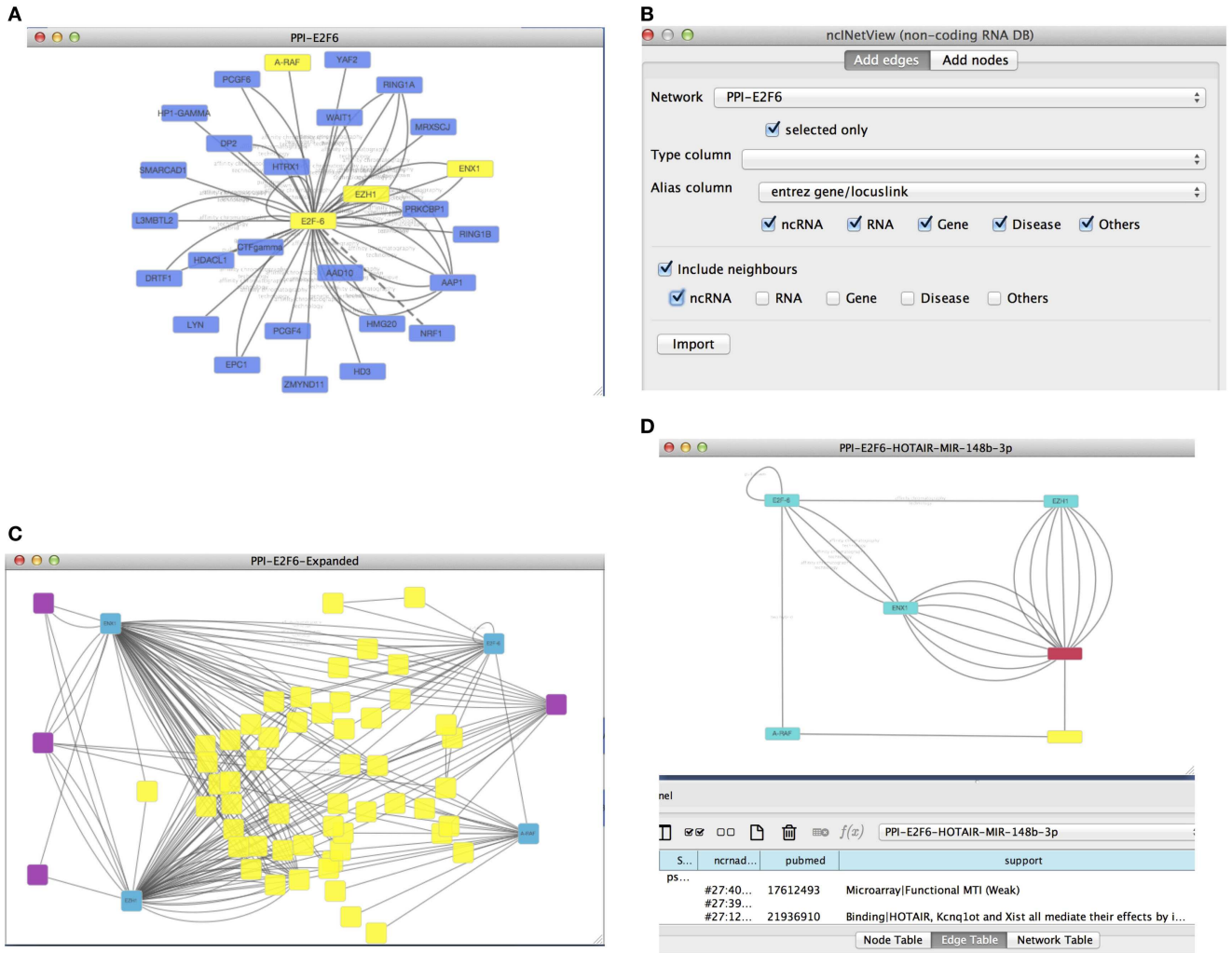


FIGURE 3 | (A) The picture reports the retrieved Biogrid network (with the E2F6 query) with four nodes (E2F6, EZH1, EZH2, and ARAF) selected by the user; **(B)** the first panel of ncINetView allows the annotation of selected nodes with all ncRNA neighbors and all their interactions with actors of the network present in ncRNA-DB; **(C)** the resulted annotated network; **(D)** an extracted circuit by the user involving the long non-coding RNA HOTAIR, the selected genes and the hsa-mir-148p-3b. The presence of more than one edge connecting nodes is motivated by reporting in the table panel for each interaction specific information such as the support sentence.

sub-classes (*ncRNA*, *ncRNA*, *RNA*, *Gene*, or *Disease*), we labeled them as Others.

At the end of the described ETL (Extract, Transform, and Load) procedure, we had: 853,543 alias, a total of 222,970 biological entities, 889,675 edges connecting *Alias* and *BioEntity* classes, and 238,524 entity relations.

Table 3 gives the total number of imported biological entities, grouped by type, and how many of them are actually involved in relations. **Table 4** reports the number of ncRNAs interacting with other ncRNA-DB biological entities.

3. UTILITY

OrientDB is supported by several language connectors, beside the native Java API. The user can query the system through programming language binding, or by using the OrientDB SQL-like console. It also implements technology standard like HTTP REST/JSON, TinkerPop Blueprints (for graph computing), and JDO (Java Data Object for object persistence). The user can develop software as client connected to the ncRNA-DB database.

Non-coding RNA human interaction data base is equipped with three alternative interfaces: (i) a CytoScape (version 3) app for importing data in a network visualization environment; (ii) a web interface; and (iii) a command-line interface for raw resource queries. Entities are specified by using their alias, through full or partial ncRNA-DB identifiers (RID or ORID).

The CytoScape plug-in and the command-line applications can be downloaded from the ncRNA-DB website at http://ncrnadb.scienze.univr.it/ncrnadb/. The documentation is also provided.

Figure 2 shows a complete schema of the proposed system, from the import phase to the user interfaces.

3.1. CYTOSCAPE INTERFACE

The CytoScape app interface, ncINetView, allows users to: (i) annotate an existing network with the ncRNA-DB relations; and (ii) search ncRNA-DB relations of specific elements and to add them to a user network or to create a new network. The source code of the Cytoscape interface, ncINetView, is available at https://code.google.com/p/ncrnadb/.

3.1.1. Add edges

The `Add Edges` takes an user network as input and annotates it with ncRNA-DB relations among its nodes.

The user selects the name of a network to be annotated by clicking on `Network`. The network needs to be already imported into the Network View of Cytoscape. In order to expand a subset of such a network, the user selects the relative nodes in the Network View Section and checks the `Selected only` option. The network table may have two columns specifying the biological entity type of each node together with the set of known aliases. The user assigns such columns to `Type column` and `Alias column`. The app maps each node of the network with the entities of the ncRNA-DB having the associated aliases.

The type column is optional. If missing, ncINetView creates one and associates the types in ncRNA-DB of the matched entities. The type of a vertex may be *ncRNA*, *RNA*, *Gene*, *Disease*, and *Others*. When it is labeled as Others, the user may assign a miscellaneous of entity types to the corresponding table entries or may leave it empty. Even in this case, the app tries to map all the matching aliases entities to the node. This behavior allows the user to specify nodes representing entities groups and to do disambiguation at a network data representation level.

The user can decide whenever some of the above entity types have to be excluded from the mapping. This can be done by unflagging the corresponding entity type check-boxes.

Once the user clicks on `Import`, the application retrieves from ncRNA-DB the matching biological entities and their relations. Then the user maps them into the network nodes and adds all found relations among them.

If the `Include neighbors` check-box is flagged, then the application retrieves all the ncRNA-DB neighbors of the matching entities and adds them to the mapped nodes, as well as relations among them and the other retrieved entities.

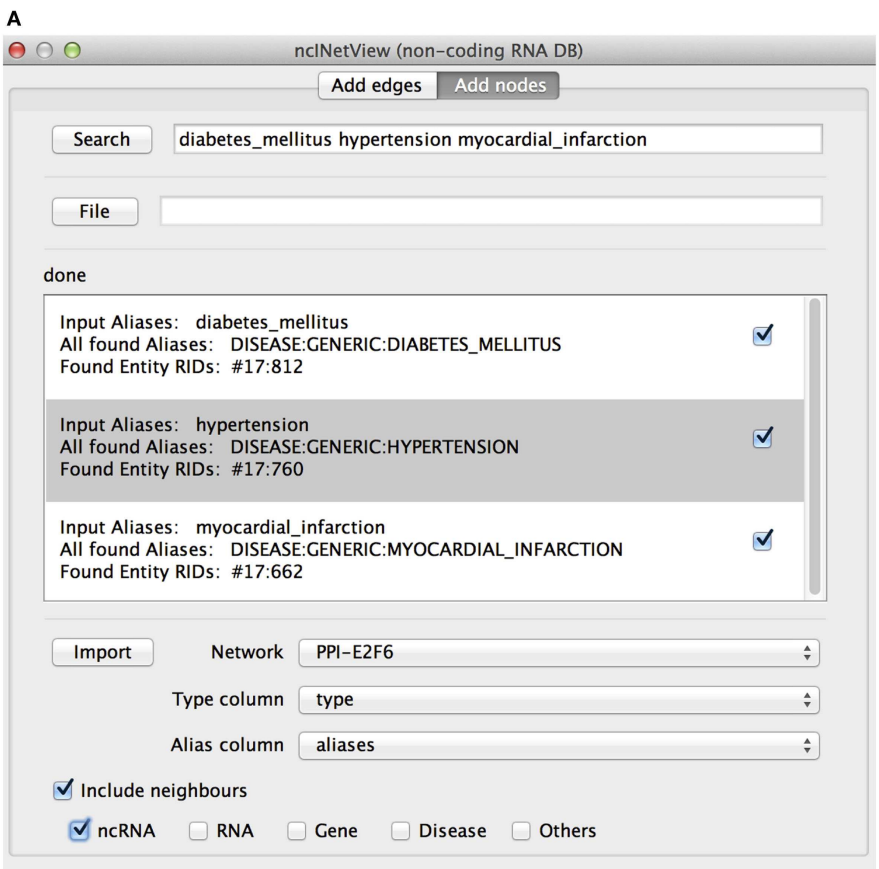


FIGURE 4 | The add nodes panel of ncINetView. **(A)** The user performs a query specifying: diabetes mellitus, hypertension, and myocardial infarction. All elements found in ncRNA-DB are reported in the text area with the associated aliases. A check box is used to include the elements in the network generated by clicking in `Import`. User selects in `Network` among those present in the cytoscape network panel, which network must be annotated; together with the name of table columns containing the aliases and the type of each node (the last is optional).

(Continued)

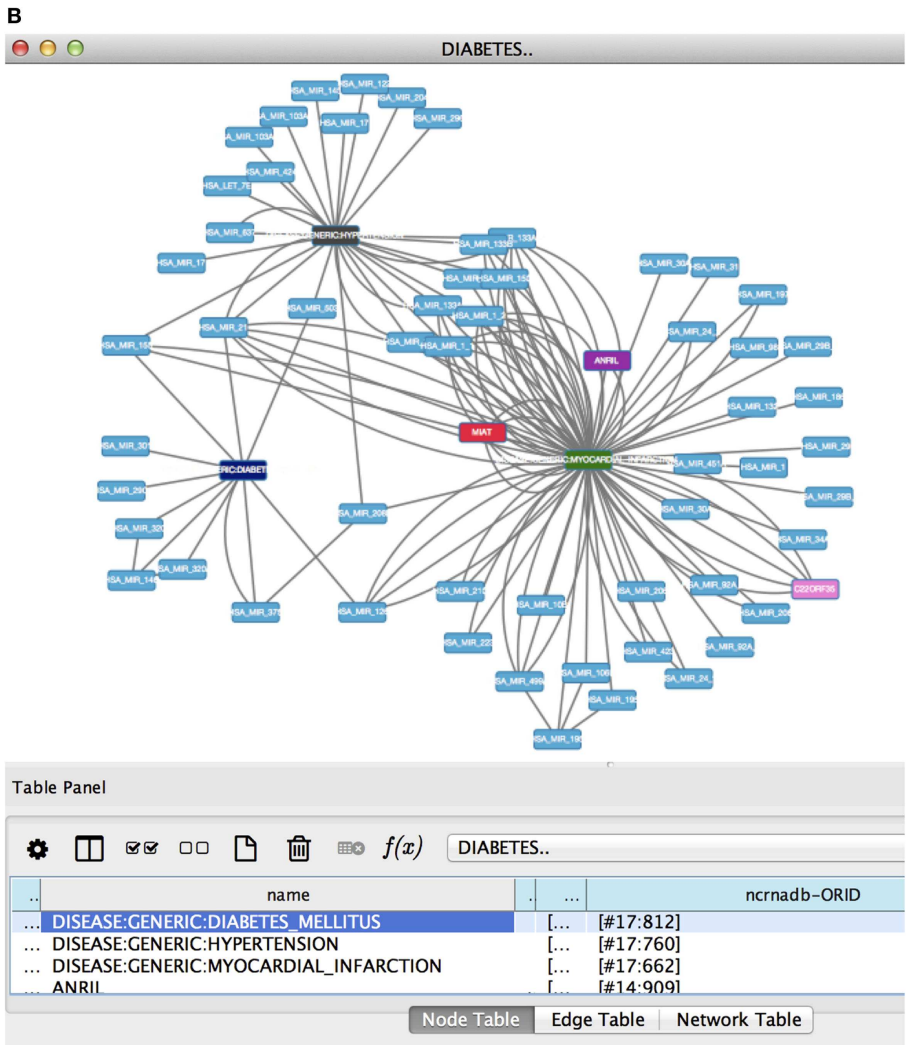


FIGURE 4 | Continued
(B) The corresponding network view is generated.

As an example, we can retrieve a protein–protein interaction network from Biogrid using the proper Cytoscape option. We searched for the protein E2F6 and we retrieved all the known experimental validated interactions stored in Biogrid. To uncover potential novel important interactions, we focused on a subnetwork by selecting some protein-coding genes: E2F6, EZH1, EZH2, and ARAF (see **Figure 3A**). Next, we used our app to extend the network with non-coding RNAs (e.g., lncRNAs and miRNAs). This yielded a new network (see **Figures 3B,C**).

From all retrieved interactions, we analyzed those involving one lncRNA (HOTAIR, Gupta et al., 2010a) and one miRNA (miR-148b-3p) (see **Figure 3D**). The hypothesis for this kind of interactions could be the following: (1) the regulation of cell cycle and (2) the role of this circuit in the chromatin remodeling. In fact, it is well known that EZH1 and EZH2 (also called ENX1) are involved in the chromatin remodeling (Margueron et al., 2008). Moreover, these genes are up-regulated in several cancers and in particular EZH2 interacts with E2F6 contributing to cellular proliferation and cell cycle progression (Attwooll et al., 2005). Interestingly, the long non-coding HOTAIR is also involved in the chromatin remodeling, carcinogenesis and metastasis (Gupta et al., 2010a). HOTAIR over-expression is associated with the reprograming of the Polycomb complex PRC2 function in breast cancer (Gupta et al., 2010a) and colorectal cancer (Kogo et al., 2011). Furthermore, its up-regulation may be a critical element in metastatic progression. In this context, the miR-148b-3p is considered a tumor suppressor miRNA, and it is down-regulated in several cancers such as the colorectal cancer (Song et al., 2012). Moreover, it has been reported that the over-expression of miR-148b could inhibit cell proliferation *in vitro* and suppress tumorigenicity *in vivo* (Song et al., 2012). A possible mechanism of the tumorigenesis in colorectal cancer and other cancers, could act through the above molecules in a circuit, which involves the up-regulation of the cited proteins and the down-regulation of miR-148b-3p mediated

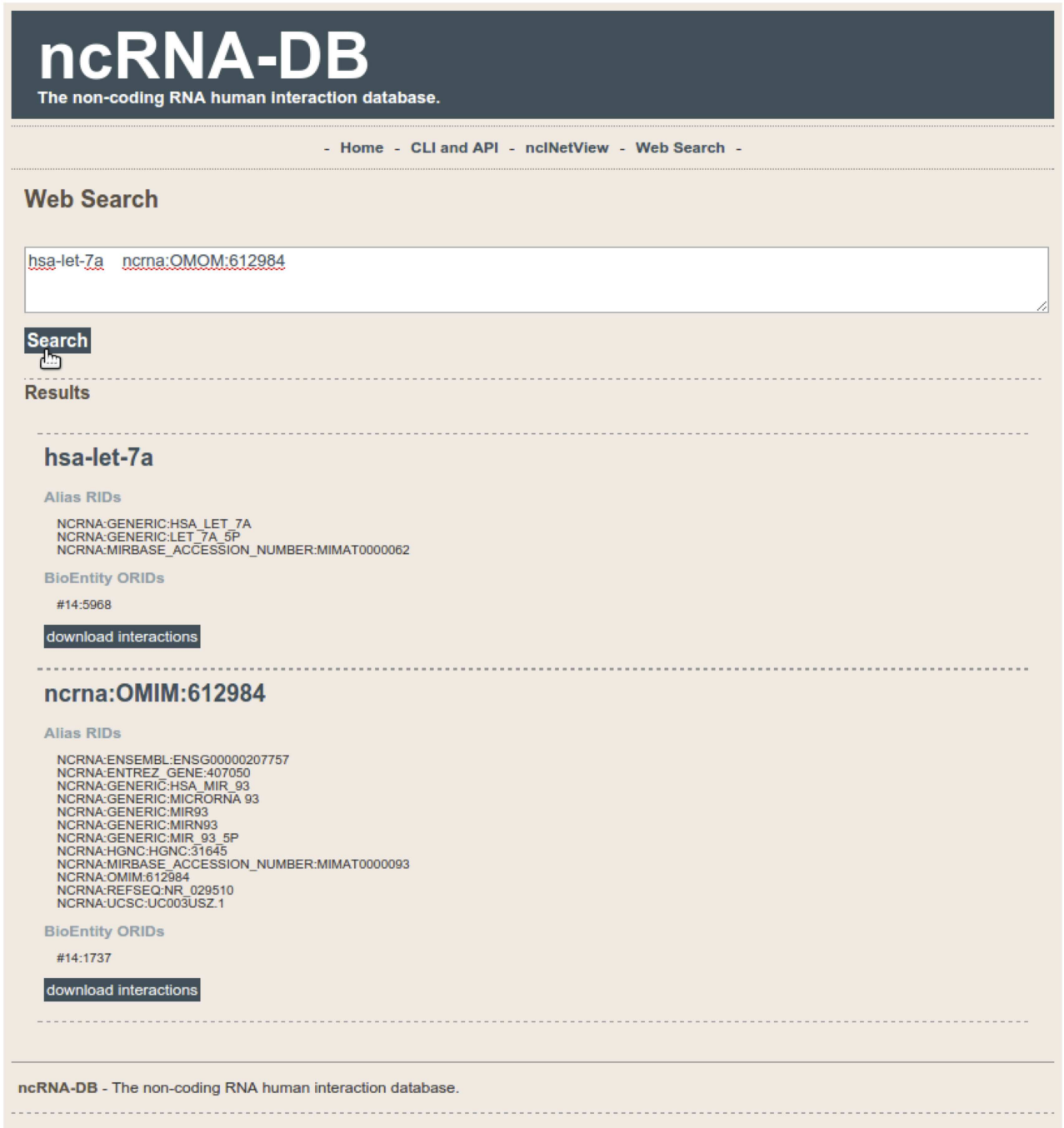


FIGURE 5 | The search section of the web interface of ncRNA-DB.

by the lncRNA HOTAIR. In this case, HOTAIR may function as competing endogenous RNAs (ceRNAs) to sponge miR-148b-3p, thereby modulating the de-repression of its targets (e.g., ARAF, a proto-oncogene may involved in cell proliferation).

3.1.2. Add nodes

`Add Nodes` allows users to search for biological entities by specifying their aliases.

In `Search`, the user specifies the entity nomenclatures to be searched separated by space (see **Figure 4A**). The app creates a node for each retrieved element. Aliases can be loaded also from file (`File`). The file has one or more aliases per row and each row corresponds to a node. If a row contains more elements than the node is a group node (i.e., a miscellaneous of entity types).

The app retrieves the matching entities and shows them in the `Results` panel (see **Figure 4B**). For each entity, the list

of corresponding aliases and their biological types are shown. Users can select the entities to be imported in the network (`Import`).

In `Network`, the user selects the name of the network to be annotated among those available in the Cytoscape Network View. Furthermore, the user specifies which column of the network table should be assigned to `Type column` and `Alias column` that contains the entity type of the nodes and their aliases. The network can be also empty.

If the `Include neighbors` check-box is flagged, then the application retrieves all the ncRNA-DB neighbors of the matching entities and adds them to the mapped nodes. The user can decide whenever some neighbor types have to be excluded from the mapping. This can be done by unflagging the corresponding entity type check-boxes.

For example, let's search for the diabetes mellitus, hypertension, myocardial infarction, and let's get all non-coding RNAs associated with them (see **Figures 4A,B**). Several ncRNAs are associated to one, two, or all three diseases.

3.2. WEB INTERFACE

We developed a web app for querying our database[10] . Users can search through a text area by putting a list of elements. The system will show the matching ncRNA-db entities and their neighbors (see **Figure 5**). Results can be saved in text format.

3.3. COMMAND-LINE INTERFACE

We developed a command-line interface to ncRNA-DB for entity searching and relation retrieval. It is released as a Java package to be platform independent and it does not require any external dependency. It provides two commands for accessing data. The `search` command takes a list of aliases as input and returns the matching biological entities stored in ncRNA-DB. This command is also useful to verify whether an identifier is included in the database and to retrieve all its alternative nomenclatures. The second command, `relations`, receives a list of entities as input, and returns the relations between them stored in ncRNA-DB and their support information. The released package also provides Java API implementing the functionality described above. The documentation is provided as JavaDoc at ncRNA-DB web site. Alternately, users may adopt the GraphAPI of OrientDB. The source code of CLI interface is available at https://code.google.com/p/ncrnadb/.

4. CONCLUSION

In this paper, we have presented ncRNA-DB, an integrated database storing knowledge concerning ncRNAs, genes, and associated diseases. The system has been implemented within the NoSQL database OrientDB. It stores data coming from several leading resources such as HGNC, lncRNAdb, circ2Traits, HMDD, lncRNADiseases, miRandola, miRTarBase, NON-CODE, and NPInter. ncRNA-DB can be queried trough three interfaces. A Cytoscape App, named ncINetView, allows to annotate biological networks with ncRNA knowledge. A web app and a command-line interface, which allows users to query the ncRNA-DB and to extract information in a text format. The aim of the proposed system is to give a comprehensive access to all the knowledge available in the literature concerning ncRNAs and associated diseases. As a key characteristics, the integrated data aim to reduce the problem of different nomenclatures used by different sources. The ncRNA-DB is available at http://ncrnadb.scienze.univr.it/ncrnadb/.

[10]http://ncrnadb.scienze.univr.it/ncrnadb/

ACKNOWLEDGMENTS

Funding: Francesco Russo has been supported by a fellowship sponsored by "Progetto Istituto Toscano Tumori Grant 2012 Prot.A00GRT."

REFERENCES

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. (2011). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39(Suppl. 1), D146–D151. doi:10.1093/nar/gkq1138

Attwooll, C., Oddi, S., Cartwright, P., Prosperini, E., Agger, K., Steensgaard, P., et al. (2005). A novel repressive E2F6 complex containing the polycomb group protein, EPC1, that interacts with EZH2 in a proliferation-specific manner. *J Biol Chem* 280, 1199–1208. doi:10.1074/jbc.M412509200

Batista, P. J., and Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell* 152, 1298–1307. doi:10.1016/j.cell.2013.02.012

Betel, D., Wilson, M., Gabow, A., Marks, D. S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res.* 36(Suppl. 1), D149–D153. doi:10.1093/nar/gkm995

Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., et al. (2011). NONCODE v3. 0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.* 40, D210–D215. doi:10.1093/nar/gkr1175

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record* 39, 12–27. doi:10.1145/1978915.1978919

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi:10.1093/nar/gks1099

Couzin, J. (2007). Erasing microRNAs reveals their powerful punch. *Science* 316, 5824. doi:10.1126/science.316.5824.530

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi:10.1038/nrg3606

Ferro, A., Giugno, R., Laganà, A., Mongioví, M., Pigola, G., Pulvirenti, A., et al. (2009). "miRScape: a cytoscape plugin to annotate biological networks with microRNAs," in *Network Tools and Applications in Biology (NETTAB), Focused on Technologies, Tools and Applications for Collaborative and Social Bioinformatics Research and Development*. ed. C. Romano (Catania: Libero di Scrivere).

Ghosal, S., Das, S., Sen, R., Basak, P., and Chakrabarti, J. (2013). Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* 4:283. doi:10.3389/fgene.2013.00283

Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., and Bruford, E. A. (2012). Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.* 41, D545–D552. doi:10.1093/nar/gks1066

Guo, Z., Maki, M., Ding, R., Yang, Y., Zhang, B., and Xiong, L. (2014). Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci. Rep.* 4, 5150. doi:10.1038/srep05150

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010a). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076. doi:10.1038/nature08975

Gupta, S. K., Bang, C., and Thum, T. (2010b). Circulating microRNAs as biomarkers and potential paracrine mediators of cardiovascular disease. *Circ. Cardiovasc. Genet.* 3, 484–488. doi:10.1161/CIRCGENETICS.110.958363

Guttman, M., and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* 482, 339–346. doi:10.1038/nature10887

Han, J., Haihong, E., Le, G., and Du, J. (2011). *Survey on NoSQL database. In Pervasive Computing and Applications (ICPCA)*. Port Elizabeth: IEEE, 363–366.

Hanke, M., Hoefig, K., Merz, H., Feller, A. C., Kausch, I., Jocham, D., et al. (2010). A robust methodology to study urine microRNA as tumor marker: microRNA-126

and microRNA-182 are related to urinary bladder cancer. *Urol. Oncol.* 28, 655–661. doi:10.1016/j.urolonc.2009.01.027

Hsu, S.-D., Tseng, Y.-T., Shrestha, S., Lin, Y.-L., Khaleel, A., Chou, C.-H., et al. (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* 42, D78–D85. doi:10.1093/nar/gkt1266

Huang, G., Athanassiou, C., and Benos, P. (2011). mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res.* 39, W416–W423. doi:10.1093/nar/gkr276

Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 28, 2062–2063. doi:10.1093/bioinformatics/bts344

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37(Suppl. 1), D98–D104. doi:10.1093/nar/gkn714

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39, 1278–1284. doi:10.1038/ng2135

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D. R., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672. doi:10.1073/pnas.0904715106

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., et al. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71, 6320–6326. doi:10.1158/0008-5472.CAN-11-1021

Kozomara, A., and Griffiths-Jones, S. (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, D68–D73. doi:10.1093/nar/gkt1181

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., et al. (2005). Combinatorial microRNA target predictions. *Nat. Genet.* 37, 495–500. doi:10.1038/ng1536

Kutmon, M., Kelde, T., Mandaviya, P., Evelo, C. T., and Coort, S. L. (2013). CyTargetLinker: a cytoscape app to integrate regulatory interactions in network analysis. *PLoS ONE* 8:e82160. doi:10.1371/journal.pone.0082160

Laczny, C., Leidinger, P., Haas, J., Ludwig, N., Backes, C., Gerasch, A., et al. (2012). miRTrail – a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC Bioinformatics* 13:36. doi:10.1186/1471-2105-13-36

Laganà, A., Forte, S., Giudice, A., Arena, M., Puglisi, P., Giugno, R., et al. (2009). miRò: a miRNA knowledge base. *Database* 2009, bap008. doi:10.1093/database/bap008

Le, T., Liu, L., Liu, B., Tsykin, A., Goodall, G., Satou, K., et al. (2013). Inferring microRNA and transcription factor regulatory networks in heterogeneous data. *BMC Bioinformatics* 14:92. doi:10.1186/1471-2105-14-92

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi:10.1016/j.cell.2004.12.035

Li, J., Liu, S., Zhou, H., Qu, L., and Yang, J. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2013). HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi:10.1093/nar/gkt1023

Liu, B., Li, J., Tsykin, A., Liu, L., Gaur, A., and Goodall, G. (2009). Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting averaging strategy. *BMC Bioinformatics* 10:408. doi:10.1186/1471-2105-10-408

Liu, Z.-P., Wu, H., Zhu, J., and Miao, H. (2014). Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. *BMC Bioinformatics* 15:336. doi:10.1186/1471-2105-15-336

Loher, P., and Rigoutsos, I. (2012). Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics* 28, 3322–3323. doi:10.1093/bioinformatics/bts615

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

Margueron, R., Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C. L., et al. (2008). Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol. Cell* 32, 503–518. doi:10.1016/j.molcel.2008.11.004

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

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

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci.U.S.A.* 105, 10513–10518. doi:10.1073/pnas.0804549105

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

Russo, F., Di Bella, S., Bonnici, V., Laganà, A., Rainaldi, G., Pellegrini, M., et al. (2014). A knowledge base for the discovery of function, diagnostic potential and drug effects on cellular and extracellular miRNAs. *BMC Genomics* 15:1–7. doi:10.1186/1471-2164-15-S3-S4

Russo, F., Di Bella, S., Nigita, G., Macca, V., Lagana, A., Giugno, R., et al. (2012). miRandola: extracellular circulating microRNAs database. *PLoS ONE* 7:e47786. doi:10.1371/journal.pone.0047786

Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S., and Romualdi, C. (2010). MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Res.* 38, W352–W359. doi:10.1093/nar/gkq423

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., 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

Song, Y., Xu, Y., Wang, Z., Chen, Y., Yue, Z., Gao, P., et al. (2012). MicroRNA-148b suppresses cell growth by targeting cholecystokinin-2 receptor in colorectal cancer. *Int. J. Cancer* 131, 1042–1051. doi:10.1002/ijc.26485

Stonebraker, M. (2010). SQL databases v. NoSQL databases. *Commun. ACM* 53, 10–11. doi:10.1145/1721654.1721659

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S. M., Ala, U., et al. (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous MRNAs. *Cell* 147, 344–357. doi:10.1016/j.cell.2011.09.029

Tesoriero, C. (2013). *Getting Started with OrientDB*. Birmingham: Packt Publishing Ltd.

UniProt Consortium. (2013). Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res.* 41, D43–D47. doi:10.1093/nar/gks1068

Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., et al. (2006). NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res.* 34(Suppl. 1), D150–D152. doi:10.1093/nar/gkj025

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.

Received: 10 October 2014; accepted: 24 November 2014; published online: 10 December 2014.

Citation: Bonnici V, Russo F, Bombieri N, Pulvirenti A and Giugno R (2014) Comprehensive reconstruction and visualization of non-coding regulatory networks in human. Front. Bioeng. Biotechnol. ***2****:69. doi: 10.3389/fbioe.2014.00069*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 12 December 2014
doi: 10.3389/fbioe.2014.00071

ncPred: ncRNA-disease association prediction through tripartite network-based inference

Salvatore Alaimo[1], Rosalba Giugno[2]*† and Alfredo Pulvirenti[2]*†

[1] *Department of Mathematics and Computer Science, University of Catania, Catania, Italy*
[2] *Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy*

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Helder I. Nakaya, Emory University, USA
Mikhail P. Ponomarenko, Russian Academy of Sciences, Russia

****Correspondence:***
Rosalba Giugno and Alfredo Pulvirenti, Department of Clinical and Experimental Medicine, c/o Department of Mathematics and Computer Science, University of Catania, Viale A. Doria 6, 95125 Catania, Italy
e-mail: giugno@dmi.unict.it; apulvirenti@dmi.unict.it

† Rosalba Giugno and Alfredo Pulvirenti have contributed equally to this work.

Motivation: Over the past few years, experimental evidence has highlighted the role of microRNAs to human diseases. miRNAs are critical for the regulation of cellular processes, and, therefore, their aberration can be among the triggering causes of pathological phenomena. They are just one member of the large class of non-coding RNAs, which include transcribed ultra-conserved regions (T-UCRs), small nucleolar RNAs (snoRNAs), PIWI-interacting RNAs (piRNAs), large intergenic non-coding RNAs (lincRNAs) and, the heterogeneous group of long non-coding RNAs (lncRNAs). Their associations with diseases are few in number, and their reliability is questionable. In literature, there is only one recent method proposed by Yang et al. (2014) to predict lncRNA-disease associations. This technique, however, lacks in prediction quality. All these elements entail the need to investigate new bioinformatics tools for the prediction of high quality ncRNA-disease associations. Here, we propose a method called *ncPred* for the inference of novel ncRNA-disease association based on recommendation technique. We represent our knowledge through a tripartite network, whose nodes are ncRNAs, targets, or diseases. Interactions in such a network associate each ncRNA with a disease through its targets. Our algorithm, starting from such a network, computes weights between each ncRNA-disease pair using a multi-level resource transfer technique that at each step takes into account the resource transferred in the previous one.

Results: The results of our experimental analysis show that our approach is able to predict more biologically significant associations with respect to those obtained by Yang et al. (2014), yielding an improvement in terms of the average area under the ROC curve (AUC). These results prove the ability of our approach to predict biologically significant associations, which could lead to a better understanding of the molecular processes involved in complex diseases.

Availability: All the *ncPred* predictions together with the datasets used for the analysis are available at the following url: http://alpha.dmi.unict.it/ncPred/

Keywords: ncRNAs-diseases association predictions, lncRNAs functional characterization, network-based inference, tripartite networks, resource transfer algorithm

1. INTRODUCTION

In recent years, great efforts have been employed in the study of non-coding RNAs (ncRNAs), a class of genes involved in a wide variety of biological functions. Small ncRNAs, such as siRNA, miRNA, and piRNA, are highly conserved in different species and have a key role in transcriptional and post-transcriptional silencing of genes. Long ncRNA (transcribed RNA molecules whose length is greater than 200 nucleotides) instead are poorly preserved and have the task of regulating gene expression through mechanisms still largely unknown (Mercer et al., 2009; Ponting et al., 2009; Wilusz et al., 2009). It has been shown that these molecules are involved in the regulation of gene expression by acting as controllers of processes such as RNA maturation or transportation, or altering chromatin structure. ncRNAs have great variety in structure and in gene regulation outcomes, however, several similarities can be identified in the way they act (Wang and Chang, 2011).

The connection between diseases and de-regulation of small ncRNAs has been established for years. However, recent studies show that mutations and de-regulations of lncRNAs are heavily involved in the development or progression of several diseases (Wapinski and Chang, 2011). Alterations in the structure (primary or secondary), or in the expression levels are the main underlying causes of diseases, from cancer to neurodegenerative disorders (Wapinski and Chang, 2011).

Pasmant et al. (2011) highlight how the expression of the lncRNA *ANRIL*, antisense transcript to *INK4b* gene, is correlated with the epigenetic silencing of *INK4a*, or *p16 protein*, which is involved in the regulation of cell cycle. High levels of *ANRIL* were found in prostate cancer tissues (Yap et al., 2010). Yap

et al. (2010), also, hypothesizes that this transcript is an initiating factor in tumor formation due to its silencing action on the *INK4b/ARF/INK4a* locus. Other experimental evidence link *ANRIL* de-regulation to a number of pathologies, including coronary disease, intracranial aneurysm, and type II diabetes (Pasmant et al., 2011).

Another example of correlation between lncRNAs and diseases is the *HOTAIR* transcript, which is involved in the progression of breast cancer by chromatin landscape remodeling (Burd et al., 2010). In particular, increased expression of tHOTAIR is an index of poor prognosis and tumor metastasis. Gupta et al. (2010) show that *HOTAIR* is also responsible for invasiveness and metastasis in epithelial cancer cells and its inhibition may lead to a reduction of invasiveness in cells where *PRC2 complex* is highly activated.

Further evidence of lncRNAs-diseases correlation is the transcript called *MALAT-1*, an RNA of more than 8000*nt* present in chromosome 11*q*13, whose over-expression is related to bad prognosis in patients with non-small cell lung cancer (Ji et al., 2003). In addition, the antisense transcript of β-*secretase-1* (*BACE1-AS*) has been identified in high concentrations in subjects with Alzheimer's disease and in amyloid precursor protein transgenic mice (Faghihi et al., 2008).

Therefore, despite the enormous importance that ncRNAs show in connection with several diseases, the number of entities, which somehow has been functionally characterized and associated to diseases, is extremely small (Wapinski and Chang, 2011). For this purpose, the developing a methodology that is able to predict ncRNA-disease interactions is crucial in order to formulate new hypotheses on the molecular mechanisms underlying complex diseases, and to identify potential new biomarkers for their diagnosis, treatment and prevention. Despite the use of such a methodology could be very helpful by making the search for new associations more focused and less costly, it must be emphasized that the task of determining, which are beneficial remains a responsibility of bio-physicians. They, indeed by identifying appropriate patient groups and properly documenting such cases, can establish the actual relationship, while also allowing a broader understanding of the underlying phenomena.

In this direction, Yang et al. (2014) developed a method, which exploits a bipartite network and a propagation algorithm to predict new associations that can be evaluated through appropriate *in vitro* experiments. Yang et al. (2014) based their method on the database assembled by Chen et al. (2013): a collection of approximately 1028 experimentally validated interactions among 322 lncRNAs and 221 diseases. The database has been further extended, through deep literature mining, to include additional interactions. The database includes also 478 experimentally validated interactions among 126 lncRNAs and 236 protein coding genes. For such genes a modulation in expression values is known to be carried out by such ncRNAs.

In this paper we present *ncPred*, a resource propagation methodology, which uses a tripartite network to guide the inference process of novel ncRNA-disease associations. The tripartite network allows the introduction of two levels of interaction: ncRNA-target and target-disease. Here, we call targets a group of biomolecules (i.e., genes, microRNAs, proteins) whose activity is modulated by a ncRNA (e.g., regulation of expression, binding to improve the efficiency of its activity, or binding to help the formation of complexes). In this way, we can exploit the greater quantity of known interactions between targets (i.e., proteins and miRNAs) and diseases to build a wider knowledge base and obtain a greater number of high quality predictions.

To perform a proper evaluation of our method, we applied a k-fold Cross-Validation procedure to the (Chen et al., 2013) database, remodeled to include information on targets. A further analysis uses a database of experimentally verified interactions between ncRNAs and miRNAs shown in Helwak et al. (2013).

2. MATERIALS AND METHODS

2.1. ALGORITHM

Let $O = \{o_1, o_2, \ldots, o_n\}$ be a set of non-coding RNAs (ncRNAs), let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of targets (i.e., genes, microRNA), and let $D = \{d_1, d_2, \ldots, d_p\}$ be a set of diseases. The ncRNA-target and target-disease interactions can be represented in a tripartite graph $G(O, T, D, E)$, where E is the set of interactions (edges) between nodes in O and T and nodes in T and D. Such a graph, can be represented by using a pair of adjacency matrices $A^{OT} = \left\{a_{ij}^{OT}\right\}_{n\times m}$ and $A^{TD} = \left\{a_{rs}^{TD}\right\}_{m\times p}$ where $a_{ij}^{OD} = 1$ if o_i is connected to t_j in G, and $a_{rs}^{TD} = 1$ if t_r is connected to d_s in G.

Our technique is based on the concept of resources transfer within the network. We refer to Alaimo et al. (2013) for details of resources transfer (drug-targeting) in bipartite networks. The bipartite network carries a prior knowledge which can be used to infer novel interactions. Starting from such a network, it computes weights between each pair of target. Those weights can be seen as the likelihood by which we can affirm that if a drug is associated with a target then it may be associated with another one. For each prediction, the algorithm also associates a score indicating the degree of certainty of the interaction.

In this paper, due to the tripartite network, we developed a multi-level transfer approach that at each step takes into account the resource transferred in the previous one (see **Figure 1** for an example). In the first level of the transfer, the resource is moved from the nodes in T (targets) to nodes in O (ncRNAs) and vice versa. In the second level, the resource is moved from D nodes to T nodes and it is combined with the resource of the previous step. Then, the resources are moved back to the D nodes. In this way, we define a methodology for the computation of a combined weight matrix $W^C = \left\{w_{ij}^c\right\}_{m\times p}$, where w_{ij}^c corresponds to the likelihood allowing us to claim that if a ncRNA interacts with a target t_i then it may be associated with the pathology d_j.

To compute such a matrix, we start by defining two partial weight matrices corresponding to the intermediate levels of transfer. These two matrices are then used to obtain the combined weight matrix and, therefore, compute the recommendations.

Let $k'(x)$ be the degree of node x in the ncRNA-target sub-network and $k''(y)$ the degree of node y in the target-disease sub-network.

The matrix $W^T = \left\{w_{ij}^T\right\}_{m\times m}$, associated with the first level of transfer, can be defined as:

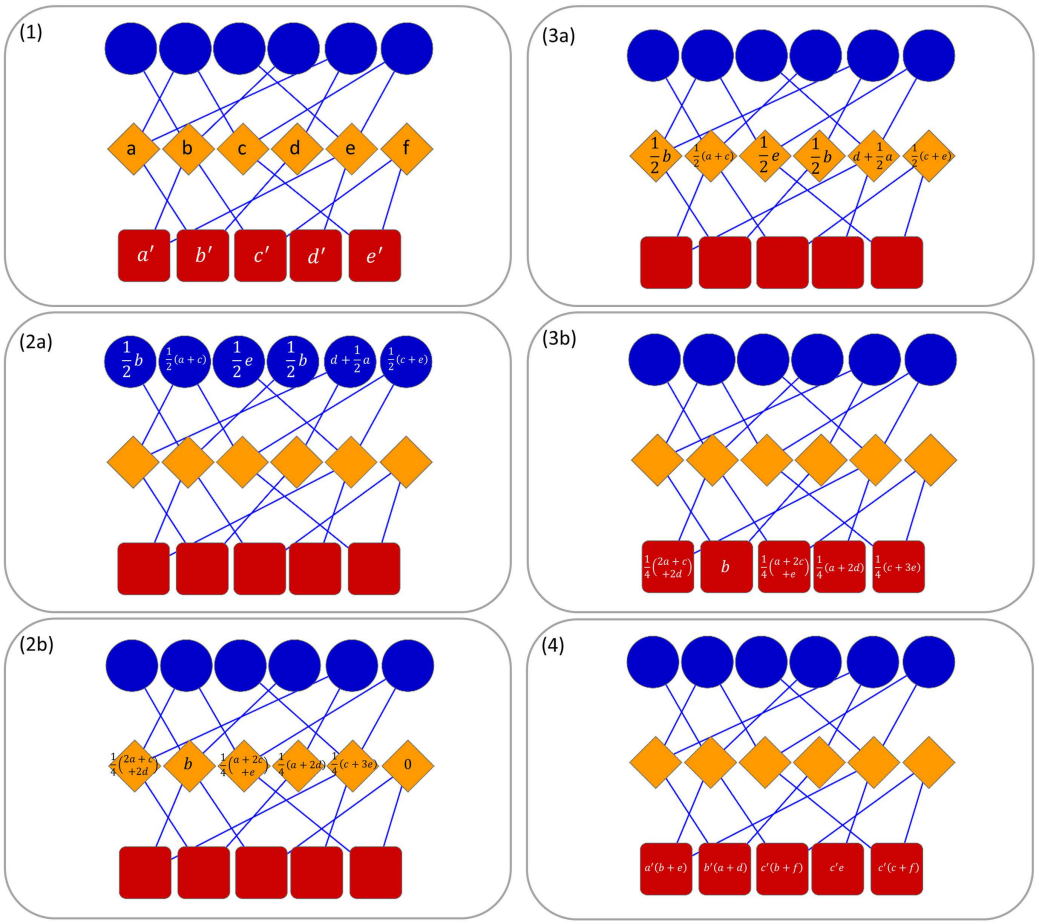


FIGURE 1 | Operating principle of ncPred in a tripartite network. Here, we represent ncRNAs in blue, targets in orange, and diseases in red. Without loss of generality, and in order to simplify the reading of the image, we decided to put λ_1 and λ_2 to 1, so as to obtain a uniform distribution of resources in the network. In the first step, a resource is assigned to each target and disease node (1). Thereafter, two separate transfer process are launched to compute the resource in target nodes (2a, 2b) and disease nodes (3a, 3b). Finally, resources are combined to obtain the total quantity in each disease node (4). In (4), the literals are used only for example purposes due to lack of space. They are to be replaced with the values computed in steps (2b) and (3b).

$$w_{ij}^{T} = \frac{1}{k'\,(t_i)^{(1-\lambda_1)}\;k'\,(t_j)^{\lambda_1}} \sum_{l=1}^{n} \frac{a_{li}^{OT}\,a_{lj}^{OT}}{k'\,(o_l)}, \qquad (1)$$

where w_{ij}^{T} corresponds to the likelihood that given a ncRNA interacting with target t_i, then it may also interact with target t_j. By using such an equation, we assign higher weights to the pairs of targets that share many ncRNAs, rather than those who share only a few.

The same applies to $W^D = \left\{w_{ij}^{D}\right\}_{p\times p}$, matrix associated with the second level of the transfer, where:

$$w_{ij}^{D} = \frac{1}{k''\,(d_i)^{(1-\lambda_2)}\;k''\,(d_j)^{\lambda_2}} \sum_{l=1}^{m} \frac{a_{li}^{TD}\,a_{lj}^{TD}}{k''\,(t_l)}. \qquad (2)$$

In equation 2, w_{ij}^{D} indicates whether we can assert that given a target associated with the disease d_i, it may also be linked to the disease d_j. w_{ij}^{D} is higher for the disease pairs, which are associated to many common targets with respect to those with fewer common targets.

In equations 1 and 2, the $\lambda_1 \in [0, 1]$ and $\lambda_2 \in [0, 1]$ parameters are used to tune the quality of the predictions. Parameter values close to zero indicate that the resource of a node is computed as the average of those in its neighborhood, while values close to one indicate that the resource is uniformly distributed among the nodes of its neighborhood. In terms of predictions, lambda values close to zero correspond to conservative predictions, while values close to one correspond to a larger number of predictions.

Therefore, the combined weight matrix $W^C = \left\{w_{ij}^{c}\right\}_{m\times p}$ can be obtained as:

$$w_{ij}^{C} = \sum_{t=1}^{m} \left[w_{it}^{T} \sum_{r=1}^{p} \left(a_{tr}^{TD} \cdot w_{rj}^{D} \right) \right]. \qquad (3)$$

In equation 3, the weight of a target-disease pair is computed by taking into account both the targets with a similar neighborhood

and the diseases with a similar neighborhood. In this way, a larger weight is assigned to those pairs for which more frequently there is a path, which passes through them.

Given the above weights, it is now possible to compute the recommendation matrix $R = \{r_{ij}\}_{n \times p}$ as:

$$R = A^{OT} \cdot W^{C}. \quad (4)$$

We call each r_{ij} prediction score for the pair (i, j). For each ncRNA o_i, its list of predictions R_i can be obtained by selecting those disease-prediction score pairs for which there is no path with o_i in the tripartite network. Such a list is sorted in descending order with respect to the value of r_{ij}, as the higher the score, the greater the belief that the ncRNA will have some connection with that particular disease.

Table 1 | Description of the datasets: number of ncRNAs, targets and diseases together with the count of interactions, average degree, density, modularity, number of connected components, and average path length.

Metrics	Chen et al. (2013)	Helwak et al. (2013)
ncRNAs	119	338
Targets	110	179
Diseases	514	134
ncRNAs–targets interactions	247	1699
Targets–diseases interactions	1005	1572
Average degree	1.572	5.025
Density	0.002	0.008
Modularity	0.609	0.274
Number of connected components	24	1
Average path length	1.572	1.734

2.2. DATASETS AND BENCHMARKS

We evaluated our method using two datasets containing experimentally verified interactions between ncRNAs, targets, and diseases. The first data set (Figure S1 in Supplementary Material) was built by collecting from (Chen et al., 2013) 478 interactions between lncRNAs and genes. These interactions were mapped by converting each target identifier to its Entrez Id. This allowed us to remove about 230 duplicates or superseded interactions. From the remaining targets, we then extracted 1005 experimentally validated gene-disease associations by searching in DisGeNET (Bauer-Mehren et al., 2010).

The second data set (Figure S2 in Supplementary Material) was obtained by collecting about 4000 lncRNA-miRNA interactions found by Helwak et al. (2013) by applying the CLASH methodology (Kudla et al., 2011). Each association indicates that a lncRNA contains one or more binding sites for miRNAs. From such a list, we removed all targets not present in miR2Disease database (Jiang et al., 2009), obtaining 1699 lncRNA-miRNA associations. Finally, using Jiang et al. (2009), we recovered 1572 miRNA-disease associations. **Table 1** provides a summary of the two datasets together with some metrics that can further elucidate their characteristics. Moreover, in **Figure 2**, we calculated the degree distribution of the two networks. These show that they can be considered scale-free networks.

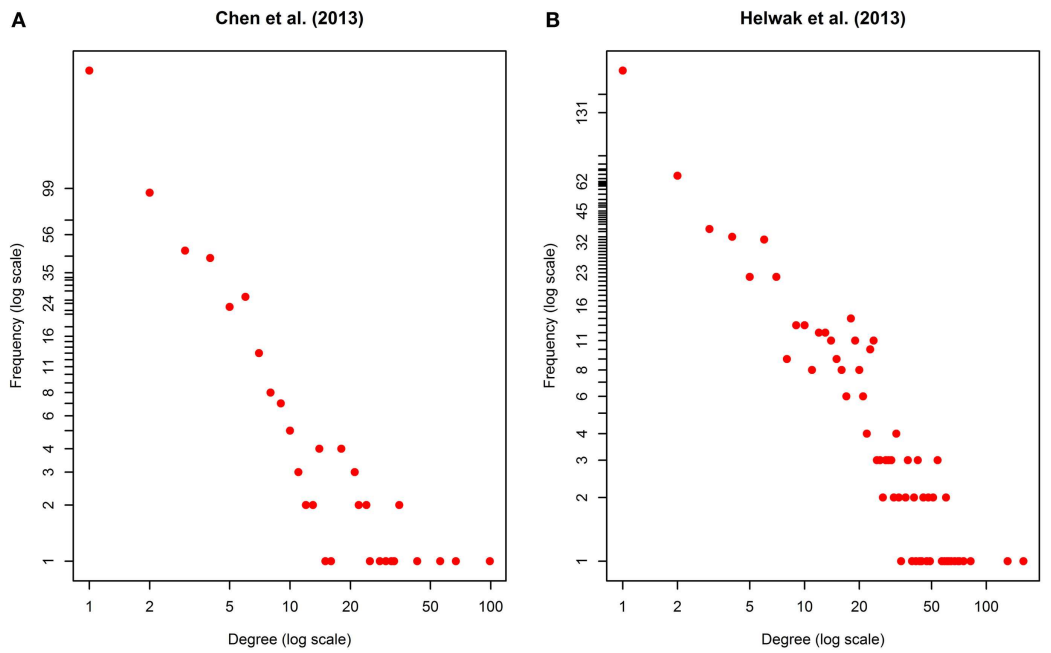


FIGURE 2 | Degree distribution of the two networks used as datasets: **(A)** Chen et al. (2013), **(B)** Helwak et al. (2013). The two plots are in log-log scale. As can be seen the degree distribution for the two networks can be approximated to an exponential one. We can therefore assume that the two networks are scale-free.

For the evaluation of our method, we applied a 10-fold cross-validation procedure repeated 30 times to obtain more reliable results. Each fold is built in the following way. Given the tripartite graph, we selected all possible pairs of ncRNA-disease interactions. Then, we randomly partitioned them into each fold. We make sure that the tripartite network generated from each fold is not disconnected. *ncPred* makes predictions only on connected networks. We considered the following four metrics (Alaimo et al., 2013) to assess the performance of our method: precision and recall enhancement, recovery, personalization, and Surprizal. The first two establish the ability of the method to recover the interactions of the test set, therefore, obtaining biologically relevant predictions. The other two measure the ability of the method to propose unexpected interactions, which may lead to novel insights onto ncRNA functions. Special care should be given to the precision and recall enhancement metrics. They measure the reliability of the prediction algorithm by comparing the standard precision and recall with a null model. Such a model is defined as a methodology that randomly assigns ncRNA-disease pairs. This implies that values greater than one are to be considered synonymous of higher quality and, therefore, reliability.

3. RESULTS

As stated earlier, to evaluate the power of our method, we applied a 10-fold cross-validation procedure repeated 30 times and averaged results to obtain more reliable estimates. In **Table 2**, we illustrate the behavior of *ncPred*, comparing it with Yang et al. (2014), in terms of precision and recall enhancement. The results demonstrate that *ncPred* clearly outperforms its competitor. In particular, we can see that while Yang et al. (2014) obtains a recall close to the null model, *ncPred* has much better results. This is crucial since the recall measures the ability of the algorithm to recover existing interactions in the network, and is therefore a sign of their reliability, namely their biological relevance.

In **Figure 3**, we report the receiver operating characteristic (ROC) curves computed on both datasets. The simulations were repeated 30 times and their results were averaged to obtain a more accurate evaluation. Both methods show a high true positive rate against low false positive rate, although *ncPred* is clearly able to achieve better results. This is also shown in **Table 2**, where we can see a significant increase in the average area under the ROC curve (AUC). Such a significance is further proved by the results shown in **Table 3**. By applying the Friedman rank sum test, we determined that the performance improvement achieved by our algorithm is

Table 2 | Comparison of ncPred and Yang et al. (2014) through the precision and recall enhancement metric, and the average area under ROC curve (AUC) calculated for each of the two datasets listed in Table 1.

Dataset	$e_P(20)$		$e_R(20)$		$AUC(20)$	
	Yang et al. (2014)	ncPred	Yang et al. (2014)	ncPred	Yang et al. (2014)	ncPred
Chen et al. (2013)	5.5113	12.3290	0.7297	1.6636	0.6217 ± 0.0178	0.7566 ± 0.0218
Helwak et al. (2013)	1.8654	5.8197	1.6509	5.6572	0.7069 ± 0.0084	0.7669 ± 0.0093

*The results were obtained using the optimal values for λ_1 and λ_2 parameters as shown in **Table 3**.*

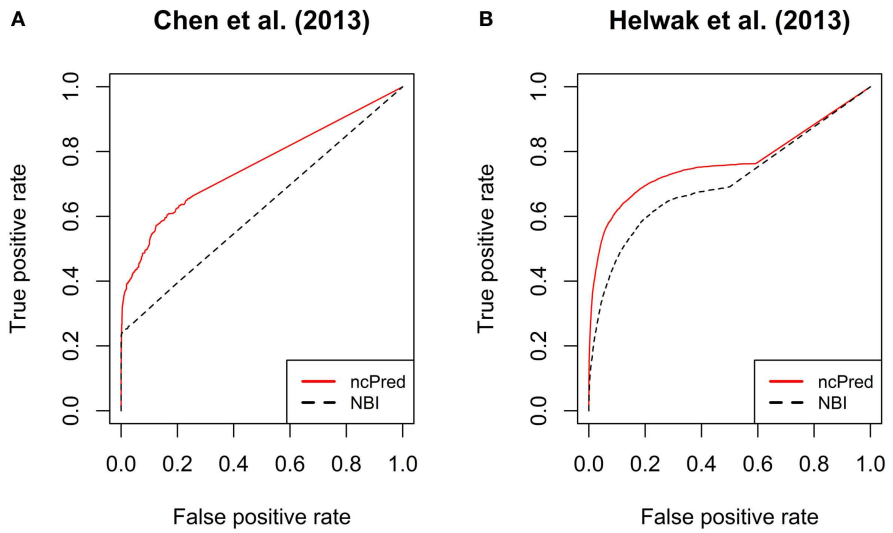


FIGURE 3 | Comparison between ncPred and Yang et al. (2014) by means of receiver operating characteristic (ROC) curves, computed for the recommendation lists built on our two datasets. Such curves measure the quality of the algorithms in terms of false positives rate against true positives rate. **(A,B)** are independent since computed on two separate datasets. The significance of the difference highlighted between ncPred and Yang et al. (2014) was measured by applying the Friedman rank sum test as assessed in **Table 4**.

statistically significant (i.e., the *p*-value is close to zero on both datasets).

Regarding the parameters λ_1 and λ_2, we performed a comprehensive analysis to establish the relationship between them and the prediction quality. In the supporting materials, we report the results of such analysis. The results indicate that there is no specific law, which governs their behavior. The peculiar characteristics of each dataset greatly affect the performances and, consequently, the parameters. It is, therefore, necessary to perform an *a priori* analysis in order to determine, which values give the best results. In our experiments, we used such an analysis to determine the best parameters in terms of precision and recall enhancement (see **Table 4** for details on their values). By looking at the characteristics of our data sets, the values obtained from such an analysis allowed us to suppose that the two parameters are close to zero in Helwak et al. (2013) dataset because of the greater density. This implies that to maintain high quality predictions it is necessary to reduce their number to avoid the introduction of noise. On the other hand, the Chen et al. (2013) dataset has a lower density. This allows us to produce a higher number of predictions before they start losing quality. Therefore, this explains the lambda values closer to one. It is important to point out that in order to determine the best parameters an analysis was performed considering only precision and recall enhancement, since they are closely related to the biological significance of the predictions. In this context, we report in **Table 2** only precision and recall enhancement and the AUC, ignoring the other metrics, which are available in the supporting materials.

Finally, assuming that the number of targets dominates the ncRNA one, we can state that the computational complexity of our method is $O(m^2 p)$. However, it is quite straightforward to implement parallelization and optimization techniques to make the computation faster.

3.1. CASE STUDIES

The analysis of the predictions for each non-coding showed that *ncPred* is able to find exactly the same predictions provided by Yang et al. (2014). The main difference between the two algorithms lies in the different scores given to each prediction. As highlighted in the previous section, ncPred is clearly able to provide more substantially accurate predictions.

Table 3 | Friedman rank sum test applied to establish the statistical significance in the performance improvement of ncPred compared to Yang et al. (2014).

Dataset	Friedman χ^2	*p*-Value
Chen et al. (2013)	1026.315	$<2.2 \times 10^{-16}$
Helwak et al. (2013)	6537.915	$<2.2 \times 10^{-16}$

Table 4 | Optimal values of λ_1 and λ_2 parameters for the datasets used in our experiments.

Dataset	λ_1	λ_2
Chen et al. (2013)	0.5	1
Helwak et al. (2013)	0.2	0.2

To further demonstrate the ability of our method, we reviewed in detail the results of five diseases (i.e., Alzheimer's Disease, Myocardial Infarction, Pancreatic Cancer, Parkinson's disease, and Gastric Cancer) as case studies. The top 10 predicted genes for each case are listed in **Table 5**. **Table 5** also shows the rank obtained by applying on our dataset, the Yang et al. (2014) method. In this context, the two datasets were taken together in order to start from a wider knowledge base.

3.1.1. Alzheimer's disease

Alzheimer's disease (AD) is one of the most common forms of dementia (Hebert et al., 2003). Recent studies indicate that it affects approximately 0.40% of the world population (Brookmeyer et al., 2007). The disease is, at present, untreatable, and it is characterized by a progressive loss of mnemonic, cognitive, and intellectual capacity, which ultimately leads to the death of the patient. Among the first 10 ncRNAs, we find *PVT1* a lncRNA,

Table 5 | List of top 10 predictions computed by ncPred and their rank obtained with Yang et al. (2014) for five case studies (Alzheimer's Disease, Myocardial Infarction, Pancreatic Cancer, Parkinson's Disease, and Gastric Cancer).

ncRNA	ncPred rank	Yang et al. (2014) rank	ncRNA	ncPred rank	Yang et al. (2014) rank
ALZHEIMER'S DISEASE					
PVT1	1	3	B2 SINE RNA	6	28
MEG3	2	19	TP53TG1	7	22
TUG1	3	32	WRAP53	8	23
lincRNA-p21	4	21	Kcnq1ot1	9	48
CDKN2B-AS1	5	20	Evf2	10	35
MYOCARDIAL INFARCTION					
H19	1	43	Kcnq1ot1	6	23
SRA1	2	24	PVT1	7	47
TUG1	3	26	CDKN2B-AS1	8	25
7SL	4	29	B2 SINE RNA	9	17
BDNF-AS1	5	34	Airn	10	18
PANCREATIC CANCER					
HOTAIR	1	16	PCAT1	6	40
LINC00312	2	15	ncRNACCND1	7	9
Kcnq1ot1	3	25	Six3OS	8	45
Xist	4	43	Airn	9	14
TERRA	5	10	RepA	10	47
PARKINSON'S DISEASE					
PVT1	1	11	LINC00312	6	24
MEG3	2	16	TP53TG1	7	20
TUG1	3	26	WRAP53	8	21
BACE1-AS	4	23	CDKN2B-AS1	9	27
lincRNA-p21	5	19	B2 SINE RNA	10	40
GASTRIC CANCER					
PTENP1	1	38	Evf2	6	60
LINC00312	2	15	Airn	7	13
Xist	3	1	TERRA	8	18
PCAT1	4	29	B2 SINE RNA	9	40
Six3OS	5	39	RepA	10	37

which regulates the transcription of *MYC* on the long distance (Carramusa et al., 2007). In Jiang et al. (2013), *MYC* has been characterized as the source of the main pathway substantially active in AD, thus having an important role in disease progression. Such a discovery confirms that *PVT1* could play a key role in the progress of AD. We have also identified the lncRNA *MEG3* that activates *TP53* and improves its binding affinity to target gene promoter (Liao et al., 2011). *TP53* was identified in Tan et al. (2012) as potential biomarker for AD. Therefore, further analysis to confirm *MEG3* role in AD are needed.

3.1.2. Myocardial Infarction

Myocardial infarction (MI) is a heart condition that occurs when the proper flow of blood to a part of the heart stops, and the heart muscle is damaged due to lack of sufficient oxygen. Genome-wide association studies have identified 27 epigenetic factors that are associated with an increased risk of MI (Feero et al., 2011). For example, the genomic locus 9p21 has one of the strongest associations with the pathology (Feero et al., 2011). The majority of such factors have been identified in regions implicated in other heart diseases (Feero et al., 2011). Among our predictions, we identified the lncRNA *SRA1* that Friedrichs et al. (2009) found crucial in cardiomyopathies. This leads us to assume a possible link with MI. In the top 10 predictions we also found the lncRNA *7SL*, which, by hybridizing to the *reverse-Alu-element-containing 3′UTR of MnSOD* gene, represses its expression (Lipovich et al., 2010). Overexpression of *MnSOD* has been identified as a possible protection against MI in transgenic mice (Chen et al., 1998). This could be a cue for further investigations to understand the role such a lncRNA.

3.1.3. Pancreatic cancer

Pancreatic cancer is an aggressive disease whose 5-year survival rate is extremely low (Amundadottir et al., 2009). The analysis of the predictions obtained by our algorithm has provided the association with lncRNA *HOTAIR*, whose overexpression has been associated with a poor prognosis in pancreatic cancer, as well as show a pro-oncogenic activity (Kim et al., 2012). A further lncRNA is *Airn*. The deletion of its promoter in paternal allele results in aberrant activation of *IGF2R* (Nagano and Fraser, 2009), whose polymorphisms are associated with an increased risk of pancreatic cancer (Dong et al., 2012).

3.1.4. Parkinson's disease

Parkinson's disease (PD) is a degenerative disorder of the central nervous system. The main cause of the disease is the death of dopamine-generating cells in the substantia nigra. The cause of this death is still unknown, nevertheless, the process of aging and metabolic stress are its common triggers (Parlato and Liss, 2014). It is interesting to note that the response to stress conditions and mechanisms for quality control are compromised in patients with PD. The reduction in the transcription of rRNA (ribosomal ribonucleic acid) is an important strategy to maintain cellular homeostasis under stress. An altered transcription is associated with neurodegenerative disorders. There are many triggers for nucleolar stress, but they seem to depend on the extitp53 protein (Parlato and Liss, 2014). Our algorithm is able to identify two probable lncRNA associated with this function: *PVT1*, also associated with AD, whose gene locus is a target of *p53* (Barsotti et al., 2012), and *MEG3* that promotes the expression of *Tp53* and increases the binding affinity to the promoters of its target (Liao et al., 2011).

3.1.5. Gastric cancer

Gastric cancer is a disease typically characterized by an overall 5 years survival rate lower than 10%, mainly due to the plurality of common symptoms that lead to treatments only in advanced disease stages (Orditura et al., 2014). Among our predictions, we find the lncRNA *Xist*. In Weakley et al. (2011), it was identified as differentially expressed in stomach preneoplastic cells, which could be a symptom of gastric cancer. Another factor could be the lncRNA *Evf2*, which is a direct putative positive regulator of transcription factor *Dlx-2* (Lipovich et al., 2010). Increased expression of *Dlx-2* was correlated with more advanced stages of the disease (Tang et al., 2013).

4. DISCUSSION

In this paper, we propose *ncPred* to predict novel associations between ncRNAs and diseases. The aim is to compute ncRNA-disease association's prediction starting from a tripartite network. Such a network integrates information on ncRNAs, targeting (i.e., those genes, microRNAs, proteins whose activity is affected by non-coding RNA), and their associations with diseases in order to improve prediction quality and accuracy.

Our experimental analysis shows that our approach predicts more biologically significant associations with respect to Yang et al. (2014). This assertion is confirmed by the results obtained in terms of recall, which as described above measures biological quality of results. The use of Friedman rank sum test also showed that the difference between our predictions and those of Yang et al. (2014) is not random but due to a better interpretation of available information. The results showed that our method could provide interesting suggestions in the study of the implications between ncRNA and pathologies. However, as stated in the introduction, the method can only help to make such a search more targeted and less expensive, offering a ranking of associations from more probable to less probable. Determine whether those associations are useful still remains within the competence area of bio-physicians that can provide conclusive evidence by identifying suitable patients and documenting such cases.

Despite what stated earlier, our method still has some limitations that should be taken into account. Firstly, ncRNA-target associations are still too small in number. It may be necessary to resort to additional targeting prediction techniques so as to expand original knowledge base. Secondly, the methodology does not use the biological information accompanying each association (e.g., type of ncRNA-target interaction, conditions in which the target-disease association was detected, tissues in which associations have significance). For this reason, it may be useful to further expand the methodology by using such additional information, which could make the methodology more reliable in terms of significant predictions.

SUPPLEMENTARY MATERIAL

In the Supplementary Material (Data Sheet 1.pdf) we report the ncPred parameter tuning further details concerning the comparison with Yang et al. (2014). The Supplementary Material for this

article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00071/abstract

REFERENCES

Alaimo, S., Pulvirenti, A., Giugno, R., and Ferro, A. (2013). Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 29, 2004–2008. doi:10.1093/bioinformatics/btt307

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R. Z., Fuchs, C. S., Petersen, G. M., Arslan, A. A., et al. (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* 41, 986–990. doi:10.1038/ng.429

Barsotti, A. M., Beckerman, R., Laptenko, O., Huppi, K., Caplen, N. J., and Prives, C. (2012). p53-Dependent induction of PVT1 and MIR-1204. *J. Biol. Chem.* 287, 2509–2519. doi:10.1074/jbc.M111.322875

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics* 26, 2924–2926. doi:10.1093/bioinformatics/btq538

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement.* 3, 186–191. doi:10.1016/j.jalz.2007.04.381

Burd, C. E., Jeck, W. R., Liu, Y., Sanoff, H. K., Wang, Z., and Sharpless, N. E. (2010). Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6:e1001233. doi:10.1371/journal.pgen.1001233

Carramusa, L., Contino, F., Ferro, A., Minafra, L., Perconti, G., Giallongo, A., et al. (2007). The PVT-1 oncogene is a Myc protein target that is overexpressed in transformed cells. *J. Cell. Physiol.* 213, 511–518. doi:10.1002/jcp.21133

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi:10.1093/nar/gks1099

Chen, Z., Siu, B., Ho, Y.-S., Vincent, R., Chua, C. C., Hamdy, R. C., et al. (1998). Overexpression of MnSOD protects against myocardial ischemia/reperfusion injury in transgenic mice. *J. Mol. Cell. Cardiol.* 30, 2281–2289. doi:10.1006/jmcc.1998.0789

Dong, X., Li, Y., Tang, H., Chang, P., Hess, K. R., Abbruzzese, J. L., et al. (2012). Insulin-like growth factor axis gene polymorphisms modify risk of pancreatic cancer. *Cancer Epidemiol.* 36, 206–211. doi:10.1016/j.canep.2011.05.013

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

Feero, W. G., Guttmacher, A. E., O'Donnell, C. J., and Nabel, E. G. (2011). Genomics of cardiovascular disease. *N. Engl. J. Med.* 365, 2098–2109. doi:10.1056/NEJMra1105239

Friedrichs, F., Zugck, C., Rauch, G.-J., Ivandic, B., Weichenhan, D., Müller-Bardorff, M., et al. (2009). HBEGF, SRA1, and IK: three cosegregating genes as determinants of cardiomyopathy. *Genome Res.* 19, 395–403. doi:10.1101/gr.076653.108

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076. doi:10.1038/nature08975

Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A., and Evans, D. A. (2003). Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.* 60, 1119–1122. doi:10.1001/archneur.60.8.1119

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi:10.1016/j.cell.2013.03.043

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P. M., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041. doi:10.1038/sj.onc.1206928

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA desregulation in human disease. *Nucleic Acids Res.* 37(Suppl. 1), D98–D104. doi:10.1093/nar/gkn714

Jiang, W., Zhang, Y., Meng, F., Lian, B., Chen, X., Yu, X., et al. (2013). Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease. *Bioinformatics* 29, 2596–2602. doi:10.1093/bioinformatics/btt423

Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., et al. (2012). HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* 32, 1616–1625. doi:10.1038/onc.2012.193

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10010–10015. doi:10.1073/pnas.1017386108

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., et al. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.* 39, 3864–3878. doi:10.1093/nar/gkq1348

Lipovich, L., Johnson, R., and Lin, C.-Y. (2010). MacroRNA underdogs in a microRNA world: evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA. *Biochim. Biophys. Acta* 1799, 597–615. doi:10.1016/j.bbagrm.2010.10.001

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

Nagano, T., and Fraser, P. (2009). Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mamm. Genome* 20, 557–562. doi:10.1007/s00335-009-9218-1

Orditura, M., Galizia, G., Sforza, V., Gambardella, V., Fabozzi, A., Laterza, M. M., et al. (2014). Treatment of gastric cancer. *World J. Gastroenterol.* 20, 1635. doi:10.3748/wjg.v20.i7.1635

Parlato, R., and Liss, B. (2014). How Parkinson's disease meets nucleolar stress. *Biochim. Biophys. Acta* 1842, 791–797. doi:10.1016/j.bbadis.2013.12.014

Pasmant, E., Sabbagh, A., Vidaud, M., and Bièche, I. (2011). ANRIL, a long, non-coding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444–448. doi:10.1096/fj.10-172452

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

Tan, M., Wang, S., Song, J., and Jia, J. (2012). Combination of p53 (ser15) and p21/p21 (thr145) in peripheral blood lymphocytes as potential Alzheimer's disease biomarkers. *Neurosci. Lett.* 516, 226–231. doi:10.1016/j.neulet.2012.03.093

Tang, P., Huang, H., Chang, J., Zhao, G.-F., Lu, M.-L., and Wang, Y. (2013). Increased expression of DLX2 correlates with advanced stage of gastric adenocarcinoma. *World J. Gastroenterol.* 19, 2697. doi:10.3748/wjg.v19.i17.2697

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi:10.1016/j.molcel.2011.08.018

Wapinski, O., and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends Cell Biol.* 21, 354–361. doi:10.1016/j.tcb.2011.04.001

Weakley, S. M., Wang, H., Yao, Q., and Chen, C. (2011). Expression and function of a large non-coding RNA gene XIST in human cancer. *World J. Surg.* 35, 1751–1756. doi:10.1007/s00268-010-0951-0

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

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS ONE* 9:e87797. doi:10.1371/journal.pone.0087797

Yap, K. L., Li, S., Muñoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., et al. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674. doi:10.1016/j.molcel.2010.03.021

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.

Received: 29 September 2014; accepted: 25 November 2014; published online: 12 December 2014.

Citation: Alaimo S, Giugno R and Pulvirenti A (2014) ncPred: ncRNA-disease association prediction through tripartite network-based inference. Front. Bioeng. Biotechnol. 2:71. doi: 10.3389/fbioe.2014.00071

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 January 2015
doi: 10.3389/fbioe.2014.00088

Discovery of protein–lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets

Jun-Hao Li†, Shun Liu†, Ling-Ling Zheng†, Jie Wu, Wen-Ju Sun, Ze-Lin Wang, Hui Zhou, Liang-Hu Qu* and Jian-Hua Yang*

RNA Information Center, Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol, Sun Yat-sen University, Guangzhou, China

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Igor B. Rogozin, National Institutes of Health, USA
Thiruvarangan Ramaraj, National Center for Genome Resources, USA

****Correspondence:***
Liang-Hu Qu and Jian-Hua Yang, Biotechnology Research Center, Sun Yat-sen University, Guangzhou 510275, China
e-mail: lssqlh@mail.sysu.edu.cn; yangjh7@mail.sysu.edu.cn

† Jun-Hao Li, Shun Liu and Ling-Ling Zheng have contributed equally to this work.

Long non-coding RNAs (lncRNAs) are emerging as important regulatory molecules in developmental, physiological, and pathological processes. However, the precise mechanism and functions of most of lncRNAs remain largely unknown. Recent advances in high-throughput sequencing of immunoprecipitated RNAs after cross-linking (CLIP-Seq) provide powerful ways to identify biologically relevant protein–lncRNA interactions. In this study, by analyzing millions of RNA-binding protein (RBP) binding sites from 117 CLIP-Seq datasets generated by 50 independent studies, we identified 22,735 RBP–lncRNA regulatory relationships. We found that one single lncRNA will generally be bound and regulated by one or multiple RBPs, the combination of which may coordinately regulate gene expression. We also revealed the expression correlation of these interaction networks by mining expression profiles of over 6000 normal and tumor samples from 14 cancer types. Our combined analysis of CLIP-Seq data and genome-wide association studies data discovered hundreds of disease-related single nucleotide polymorphisms resided in the RBP binding sites of lncRNAs. Finally, we developed interactive web implementations to provide visualization, analysis, and downloading of the aforementioned large-scale datasets. Our study represented an important step in identification and analysis of RBP–lncRNA interactions and showed that these interactions may play crucial roles in cancer and genetic diseases.

Keywords: long non-coding RNA, RNA-binding protein, GWAS, CLIP-Seq, RNA-Seq

INTRODUCTION

Mammalian genomes encode thousands of long non-coding RNAs (lncRNAs) (Wang and Chang, 2011; Guttman and Rinn, 2012). lncRNAs play important roles in a variety of biological processes that have been implicated in regulating tumorigenesis through interaction with RNA-binding proteins (RBPs) (Konig et al., 2011; Wang and Chang, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). However, for the majority of lncRNAs, the mechanism underlying their interaction with RBPs remains unknown (Konig et al., 2011; Wang and Chang, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013).

The control and function of lncRNA are governed by the specificity of RBPs (Wang and Chang, 2011; Guttman and Rinn, 2012). Increasing evidence suggests that many RBP–lncRNA interactions play important roles in correct transcriptional regulation (Konig et al., 2011; Wang and Chang, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). One emerging theme that many lncRNAs regulate gene expression by directing chromatin modificators to specific target regions (Ulitsky and Bartel, 2013). Significant fractions (20% in human) of lincRNAs are interacted with PRC2 and other chromatin-modifying complexes (Khalil et al., 2009; Guttman et al., 2011). The functional outcomes of some binding events have been revealed. For example, HOTAIR, which is transcribed from human HOX locus, guides repressor PRC2 to specific mammalian loci to silence gene expression and to promote cancer metastasis (Rinn et al., 2007; Wang et al., 2011). Besides, many lncRNAs have been shown to interact with other types of RBPs, including DNA methyltransferases (Schmitz et al., 2010; Di Ruscio et al., 2013), transcription factors (Wang et al., 2014), and splicing factors (Tripathi et al., 2010; Gong and Maquat, 2011; Yin et al., 2012). However, deciphering the interactions between hundreds of RBPs and thousands of lncRNAs remains a daunting challenge.

Genome-wide association studies (GWAS) have identified thousands of common genetic variants related to specific traits or disease phenotypes, and many of these variants (about 88%) lie in non-coding regions, which could potentially influence processing and expression of ncRNAs (Sethupathy and Collins, 2008; Hindorff et al., 2009; Ryan et al., 2010; Cabili et al., 2011; Kumar et al., 2013; Ning et al., 2014). For example, single nucleotide polymorphism (SNP) within miR-125a gene alters the processing of pri-miRNA by DGCR8 and causes recurrent pregnancy loss in a Han-Chinese population (Duan et al., 2007; Hu et al., 2011). Another study found that a papillary thyroid carcinoma-associated SNP, rs944289 affects the expression of lncRNA PTCSC3 by changing the binding activity of C/EBPα transcription factor (Cabili et al., 2011; Jendrzejewski et al., 2012). Although the genetic variants in interaction sites of RBP–lncRNA may interfere lncRNA functions and affected the susceptibility to human diseases, the relationships between genetic variants and interaction sites were yet unexplored.

Recent advances in high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation (HITS-CLIP, CLIP-Seq, PAR-CLIP, CLASH, iCLIP) have provided powerful ways to identify RBP-associated RNAs and map such interactions in the genome (Chi et al., 2009; Hafner et al., 2010; Konig et al., 2011; Helwak et al., 2013; Fu, 2014; Fu and Ares, 2014). The application of CLIP-Seq methods has reliably identified Argonaute (Ago) binding sites and miRNA-target interactome (Chi et al., 2009; Hafner et al., 2010; Helwak et al., 2013). In fact, many more studies to date have been focused on understanding the function of RBPs in RNA metabolism (Konig et al., 2011; Fu, 2014), such as pre-mRNA splicing (Fu and Ares, 2014). While an increasing number of RBPs have been explored using CLIP technologies, binding peaks mapped to non-protein-coding genes have been routinely discarded and not further analyzed. However, this data will be a rich trove well worthy of mining RBP–lncRNA relationships.

In this study, we performed a large-scale integration of public RBP binding sites generated by high-throughput CLIP-Seq technology and identified thousands of RBP–lncRNA interactions. Furthermore, by combining GWAS and RNA-Seq data, we explored clinically relevant RPB–lncRNA interactions that may facilitate the translation of genetic studies of complex diseases into therapeutics.

MATERIALS AND METHODS

INTEGRATION OF RBP BINDING SITES FROM PUBLISHED CLIP DATA

HITS-CLIP, PAR-CLIP, and iCLIP binding clusters/peaks data were retrieved from the gene expression omnibus and sequence read archive (SRA) (Barrett et al., 2013), the supplementary data of original references or directly from authors upon request. All binding sites coordinates were converted to hg19 and mm10 assemblies using the UCSC LiftOver Tool (Meyer et al., 2013).

RBP TARGET SITES SCANNING IN ANNOTATED lncRNA TRANSCRIPTS

Human gene annotations were acquired from GENCODE Version 17 (Harrow et al., 2012). Mouse gene annotations were extracted from Ensembl Gene Release 72 (Hubbard et al., 2009) and LiftOver to mm10 assembly. lncRNAs were further filtered to remove the transcripts overlapping with protein-coding genes. The aforementioned RBP CLIP clusters were used to intersect with the coordinates of all annotated transcripts to find their RBP binding sites, which were fed to Circos (Krzywinski et al., 2009) for visualization.

TCGA TUMOR EXPRESSION DATA AND EXPRESSION CORRELATION OF RBPs AND lncRNAs

The Cancer Genome Atlas (TCGA) RNA-Seq expression datasets (level 3, IlluminaHiSeq_RNASeqV2) for 14 cancer types and gene annotation file (TCGA.hg19.June2011.gaf) were downloaded from TCGA Data Portal (Cancer Genome Atlas Research Network, 2008). Expression of 397 known lncRNAs can be measured in TCGA level 3 RNA-Seq data. Expression correlation (Pearson correlation coefficient) between lncRNAs and RBPs was estimated using co-expression program (the program is available from the authors upon request), which was written in C language and ALGLIB library, and p-value was adjusted with the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995).

IDENTIFICATION OF DISEASE-RELATED SNPs IN RBP BINDING SITES ASSOCIATED WITH lncRNAs

Disease/phenotype associated SNPs were curated from published GWAS data provided by the NHGRI GWAS Catalog (Welter et al., 2014), Johnson and O'Donnell (2009), dbGAP (Mailman et al., 2007), and GAD (Becker et al., 2004). Additional SNPs in linkage disequilibrium (LD) with reported disease-related loci were selected with the criteria requiring an r^2 value over 0.5 in at least one of the four populations (CEU, CHB, JPT, and YRI) genotype data of the HapMap project (release 28) (International HapMap 3 Consortium et al., 2010). For each SNP, rs ID were lifted to dbSNP build 141 based on the "RsMergeArch.bcp" and "SNPHistory.bcp" table from dbSNP, and genomic coordinates were lifted to the hg19 assembly using the UCSC LiftOver tool. All these disease-related SNPs or LD SNPs were mapped to exons and splicing sites (2 nt in the intron that is close to an exon) of the annotated lncRNA transcripts and further examined whether they were located in any RBP binding clusters.

DATA VISUALIZATIONS

RNA-binding protein–lncRNA interactions were deposited in our starBase V2.0 (Li et al., 2014) under the "Protein–RNA" section[1]. For each interaction, we provided links to our enhanced deepView genome browser[2], which was written using a GD graphics library for PHP, to visualize RBP binding sites, lncRNAs, and other annotation tracks in an integrated display style similar to that of UCSC genome browser.

RESULTS

THE GENOME-WIDE BINDING MAP OF RNA-BINDING PROTEINS AND THE ANNOTATION OF RBP–lncRNA INTERACTIONS

We curated 117 published CLIP-Seq datasets to profile the genome-wide binding maps of 65 RBP. Unique binding sites of distinct RBPs varied from thousands to millions, and the genomic context distributions of binding sites for different RBPs distinguished from each other (**Figure 1**; Table S1 in Supplementary Material). For example, PUM2, a translational repressor during embryonic development and cell differentiation (Huang et al., 2011), predominately bound to 3′UTR regions of protein genes, while another translation inhibitor FMRP (Napoli et al., 2008) tended to interact with CDS. The discrepancy in binding context preferences for RBPs could root from different amounts of available datasets, usages of various variants of CLIP-Seq, varying sequencing depth, and/or genuine distinctions in the underlying recognition mechanism of RBPs.

Despite that the majority of RBP binding sites were mapped to protein-coding genes, on average 1.1% of RBP binding sites lay within exons of human lncRNAs. In total, 21,073 and 1,662 RBP–lncRNA interactions were identified in human and mouse, respectively (**Table 1**). It is noteworthy that most well-studied lncRNAs interacted with chromatin modificators, acting as tethers or scaffolds (Khalil et al., 2009; Kung et al., 2013). Thus, we considered the binding features of Ezh2, a subunit of PRC2 complexes, by analyzing the CLIP-Seq data from mouse embryonic stem cells

[1]http://starbase.sysu.edu.cn/rbpLncRNA.php
[2]http://starbase.sysu.edu.cn/browser.php

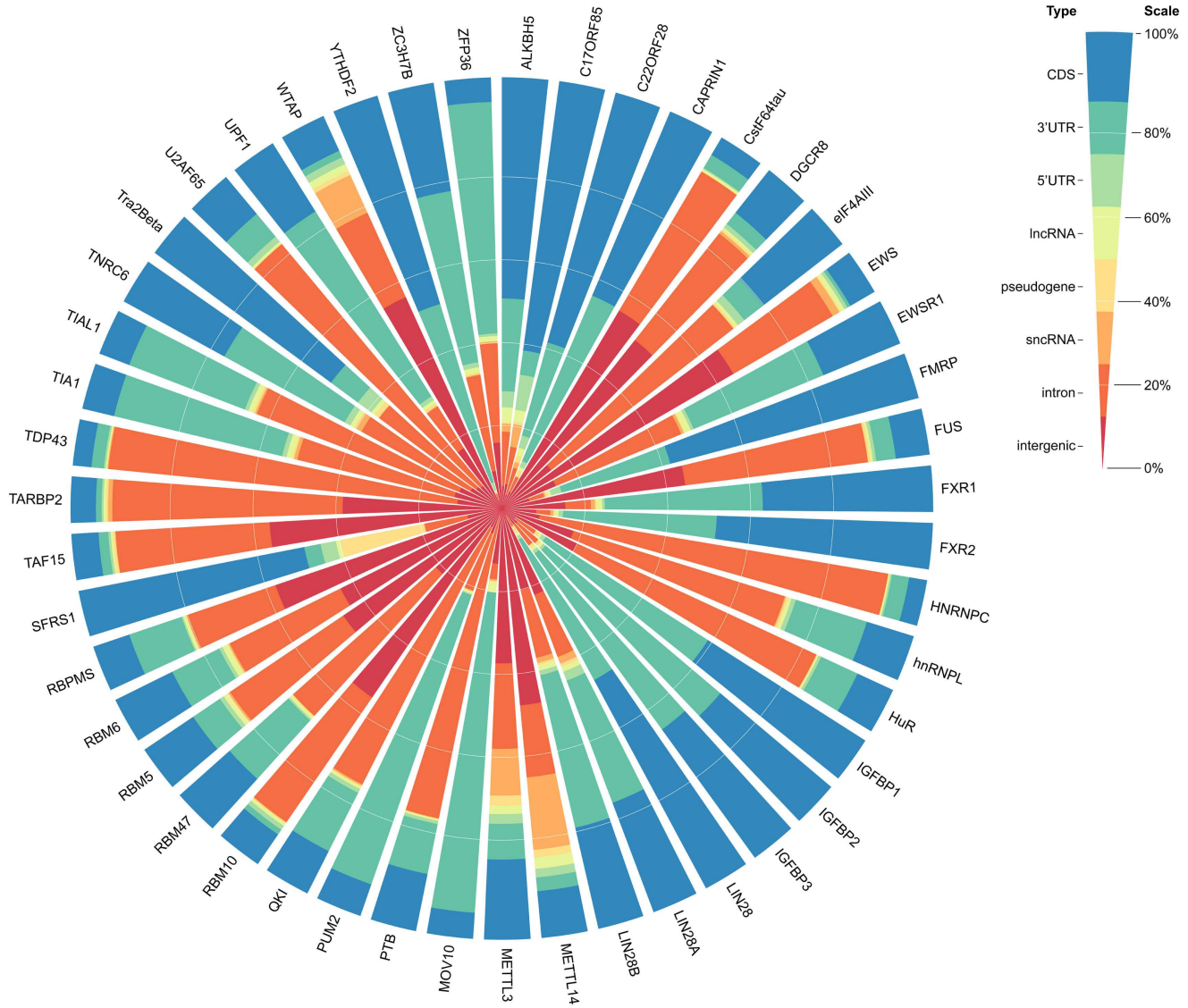


FIGURE 1 | The genomic context distributions of binding sites for 47 human RBPs. Binding sites are mapped to genomic features in the following priority order: CDS, 3′UTR, 5′UTR, lncRNA, pseudogene, sncRNA, intron, intergenic.

Table 1 | The summary of CLIP-Seq datasets used in this study and the resulting RBP–lncRNA interactions.

Species	Experiments	RBPs	Cell lines/ tissues	RBP binding sites mapped to lncRNAs	RBP–lncRNA interactions
Human	90	47	13	84,356	21,073
Mouse	27	18	20	5,330	1,662

(Kaneko et al., 2013). Our results demonstrated that Ezh2 interacted with 35 lncRNAs including many imprinted RNAs, such as Tsix, Meg3, Rian, and Pvt1 (**Figure 2**), which was consistent with the epigenetic features of PRC2 (Zhao et al., 2010).

EXPLORING COMBINATORIAL EFFECTS AMONG RBPs

For the 12,255 human lncRNAs, 56.8% were found bound to at least 1 RBP. Surprisingly, 16 lncRNAs, including GAS5 and NEAT1, harbored binding sites of over 30 RBPs (**Figure 3**; Table S2 in Supplementary Material), indicating their diverse roles in biological processes when accompanied with different RBPs. Since one lncRNA could interact with multiple RBPs, it could be expected that some RBP binding sites were overlapped with each other. Therefore, we explored combinatorial effects among RBPs by employing integrated CLIP-Seq datasets. For example, we utilized PAR-CLIP data generated in HEK293 and intersect binding sites of three RNA destabilizer HuR, Ago2, and MOV10. The results showed that tens of lncRNAs, including cancer-related lncRNAs TUG1, DLEU2, and GAS5, were bound by at least two of the

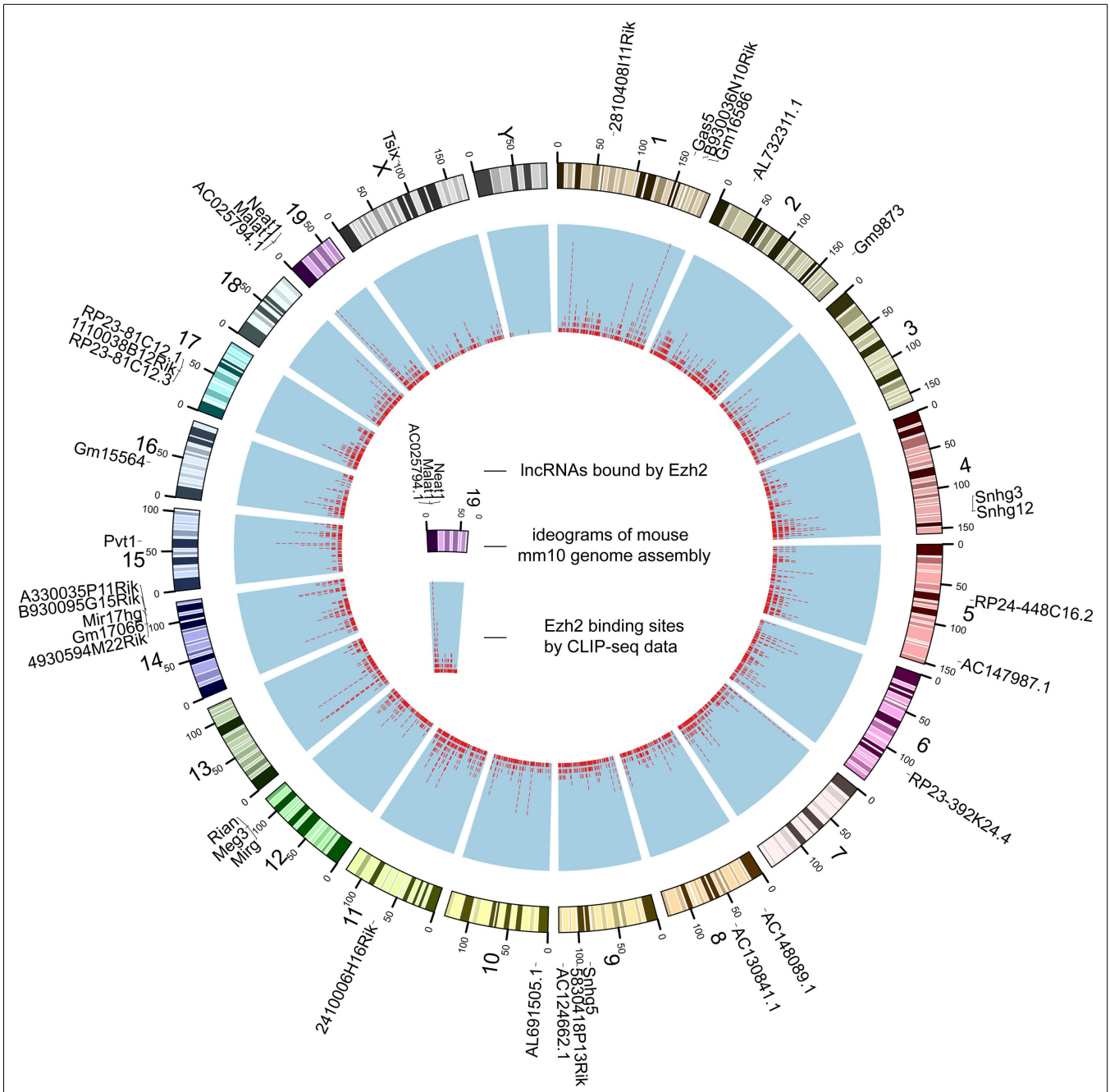


FIGURE 2 | The genome-wide binding map of Ezh2 in mouse. The outer track is mouse chromosomes labeled with lncRNAs bound by Ezh2. The red tiles of the inner track represent the genomic coordinates of corresponding binding sites.

three RBPs at identical binding sites (**Figure 4**). This phenomenon suggested that the stabilities of these lncRNAs were likely under joint control of these three RBPs, which could be explained by their confirmed interplays in HEK293 (Chendrimada et al., 2007) and Hela cells (Kim et al., 2009).

EXPRESSION ASSOCIATION OF RBP–lncRNA INTERACTIONS

To realize the roles of RBP–lncRNA interactions in cancer, we preformed co-expression analysis between RBPs and lncRNAs by virtue of 90 human CLIP-Seq datasets and expression data from more than 6,000 tumor samples in 14 types of cancer. Up to 583 pairs concerning 47 RBPs and 49 lncRNAs showed strong correlation at expression levels in at least 1 cancer type (**Figure 5A**). Marvelously, PUM2 and TUG1 involved with cell cycle regulation (Khalil et al., 2009; Huang et al., 2011) showed significant positive expression correlation ($p < 0.05$) in all 14 cancer types (**Figure 5B**). Two potential PUM2 binding sites on TUG1 have the consensus recognition

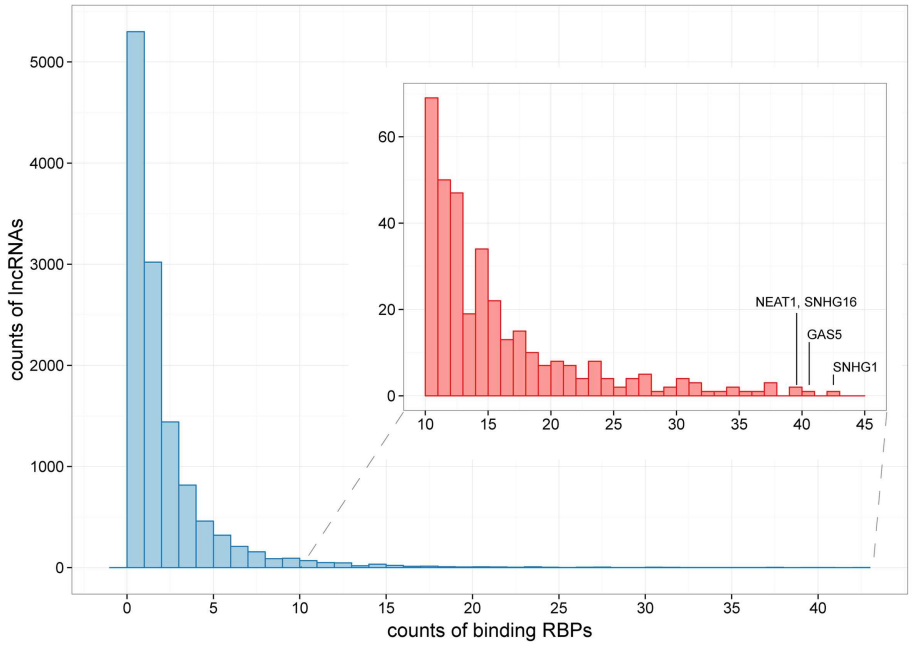


FIGURE 3 | The distribution of lncRNAs bound by different numbers of RBPs. Histograms showing counts of lncRNAs bound by over 10 RBPs are zoomed in at the subpanel. SNHG1, GAS5, NEAT1, and SHNG16 are marked, which are bound by 42, 40, 39, and 39 RBPs, respectively.

motif UGURUAUA, which was highly conserved in mammals (**Figure 5C**).

PREDICTING GWAS-ASSOCIATED RBP BINDING SITES IN lncRNAs

Although GWAS over the years have revealed a significant number of genetic variants related to diseases or phenotypes, a considerable portion of these identified loci are not within protein-coding genes and therefore not functionally explained to date (Hindorff et al., 2009). Here, we tried to fill this gap by connecting RBP binding sites in lncRNAs and potential disease-related SNPs.

Altogether, 87,677 unique disease-related SNPs were collected from four public GWAS data source (Table S3 in Supplementary Material, detailed in Section "Materials and Methods"). Considering that additional SNPs in LD with reported disease-related loci may also map to RBP binding sites in lncRNAs, we perform LD analysis to extracted SNPs that had high LD relationship with disease-related SNPs using a threshold of $r^2 > 0.5$ in at least one population from the HapMap CEU, CHB, JPT, and YRI genotype data, which yielded a total of 895,968 disease-related or LD SNPs.

We found that 2431 of these SNPs were mapped to the exons of 2089 transcripts of 1489 lncRNA genes, among which 162 SNPs were also located in at least 1 binding sites of 29 RBPs (Table S4 in Supplementary Material). For example, three disease-related SNPs, namely, rs16902485, rs10283090, and rs2720659, resided in the exons of lncRNA PVT1. According to the GWAS annotations of Johnson and O'Donnell (2009), the latter two of the three SNPs were associated with "type II diabetes mellitus," which was in good accordance with the recent reports showing that PVT1 may contribute to diabetic nephropathy (Hanson et al., 2007; Alvarez and DiStefano, 2011; Alwohhaib et al., 2014). These SNPs were also overlapped with binding sites of U2AF65, HuR, and eIF4AIII, respectively (**Figure 6**), suggesting variants in these sites might result in impaired binding of these RBPs to PVT1, which thereby might lead to the development of corresponding diseases.

Next, we checked whether disease-related SNPs might be located in the splicing sites of lncRNAs and affect the alternative splicing of lncRNAs. We defined a splicing site as the 2 nt within an intron close to the exon–intron junction. As a result, we found that only 24 SNPs lay within lncRNA splicing sites (Table S4 in Supplementary Material), among which only 1 SNP, rs17207481, was overlapped with binding sites of FUS and HuR. These results suggested that SNPs exerted limited effects on disease occurrence through the mechanism of disturbing alternative splicing of lncRNAs.

COMPARATIVE ANALYSIS OF RBP TARGETS USING THE deepView GENOME BROWSER

To facilitate comparative analysis of the CLIP-Seq datasets and exploration of RBP–lncRNA interactions, we developed the improved deepView Genome Browser[2] in starBase V2.0 (Li et al., 2014). In the query page of the browser, users can input one interested genomic region in the "search term" and select corresponding genome assembly to gain an integrated view of various genomic features. Information on binding sites of RBPs, predicted miRNA-target sites overlapped with CLIP-Seq data, as well as gene annotations from RefSeq and Ensembl were provided in toggleable tracks. The image of the browser will be updated immediately by clicking the "refresh tracks" button when users change track options. **Figure 7** illustrated the visualization of FUS–MEG3 interactions with deepView. Users can click the "zoom in" or "zoom out"

FIGURE 4 | The genome-wide binding map of HuR, Ago2, and MOV10 in human. The outermost track represents ideograms of chr1, chr13, and chr22 in human genome. lncRNAs bound by these RBPs are labeled on the periphery, and those bound by at least two of the three RBPs at identical binding sites are colored red. The blue, green, and purple tracks indicate the binding positions of HuR, Ago2, and MOV10, respectively.

button at the top to shrink or extend on the center of the annotation tracks window by 1.5-, 3-, or 10-folds. Clicking the "View region at UCSC" button will redirect users to the UCSC page and exhibit the current region on the UCSC genome browser. To explore RBP binding sites on a particular gene, users can type its gene symbol in the position textbox and then click the "GO" button to update the display image to determine, which RBPs might participate in regulating the gene. Our visualization method allows a direct comparison of binding patterns of different RBPs, binding preferences of one particular RBP in different cell lines and tissues, and genomic contexts of RBP binding sites.

DISCUSSION

Although a few dozen lncRNAs have been characterized to some extent and reported to function in important cellular processes, the functions of most annotated lncRNAs are

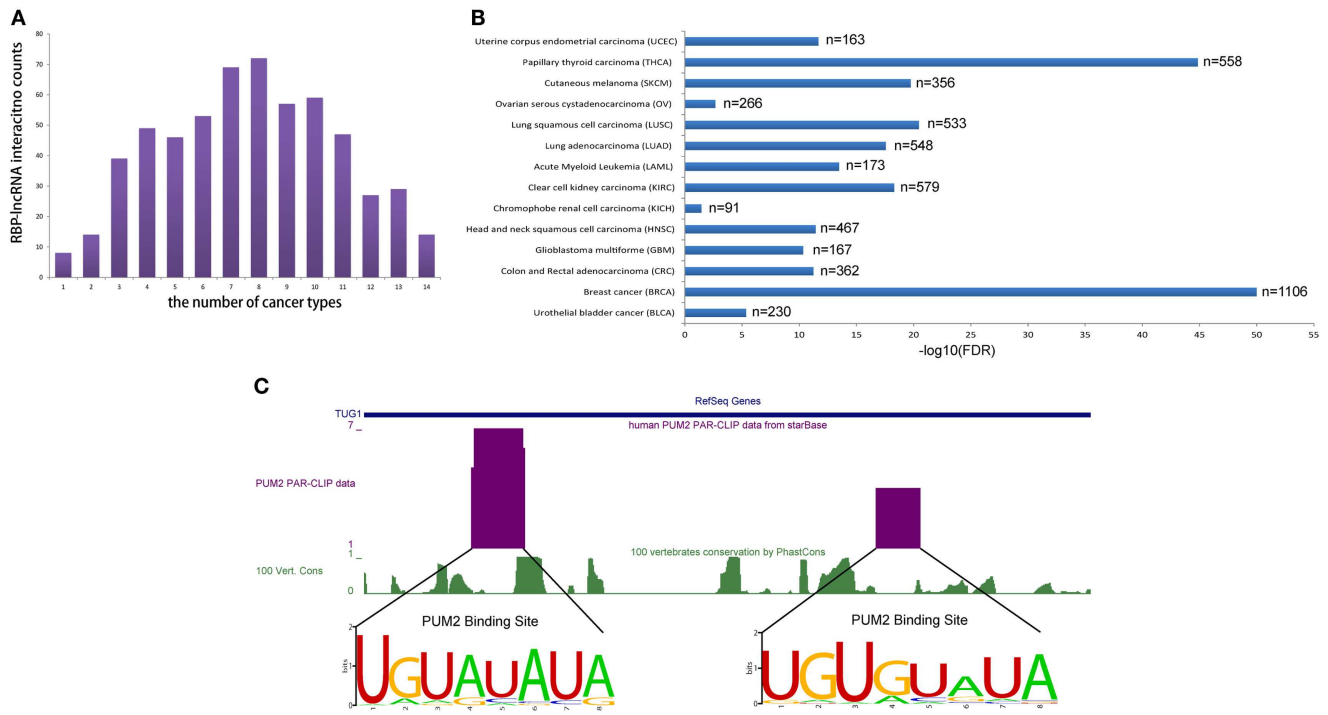


FIGURE 5 | RBP–lncRNA interactions are supported by co-expression analysis in 14 types of cancers. **(A)** Histograms show RBP–lncRNA interactions with expression association (Pearson correlation, $p < 0.05$) in at least one cancer type. **(B)** The expression levels of PUM2 and TUG1 are positively correlated ($p < 0.05$) in all 14 cancers. **(C)** The PUM2 binding sites on TUG1 are inferred from PAR-CLIP data, and the consensus recognition motif UGURUAUA are conserved in mammals.

unknown (Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). Several bioinformatics resources and tools have made efforts to functionally annotate lncRNAs (Da Sacco et al., 2012), such as fRNAdb (Kin et al., 2007) and ncFANs (Liao et al., 2011). These tools mainly inferred lncRNA function by their differential expression in distinct biological states or their co-expression patterns with protein-coding genes, but little attention was paid to the relationship of lncRNAs and their bounded proteins. In this study, by analyzing a large set of RBP binding sites derived from all available CLIP-Seq experimental techniques (PAR-CLIP, HITS-CLIP, iCLIP, CLASH), we have shown extensive and complex RBP–lncRNA interaction networks (**Figure 1**).

Recent studies have revealed that many lncRNAs function through specific interactions with RBPs, but whether these interactions are direct and specific remains controversial. RBP–lncRNA interactions identified by low stringent immunoprecipitation of non-cross-linked RNA–protein complexes, such as RIP-Chip and RIP-Seq, may contain indirect binding relationships (Konig et al., 2011). In comparison to previously reported significant fractions (10% in mouse) of PRC2-associated lncRNAs (Zhao et al., 2008), we found that a relatively small fraction (~1%) of lncRNAs were bound by Ezh2 in mouse (**Figure 2**). Therefore, we provide enhanced resolution to determine lncRNA functional networks based on RBP–lncRNA interactions supported by high-throughput CLIP-Seq data. More than 80,000 binding clusters identified from 65 different RBPs represent a valuable resource for resolving some obstacles that have arisen in efforts to understand lncRNA action. Nevertheless, although CLIP-Seq is designed to detect direct binding events of proteins and RNAs, the resulting data might still contain false positives and false negatives, which may root from every cumbersome step of this technique. To minimize the impact of such false discoveries, we filtered the origin results by the reported FDR and provided evidences such as number of CLIP reads and number of supporting experiments, which may help users to gain RBP–lncRNA interactions of high-confidence.

By cross analysis of binding maps for multiple RBPs, this study offers a new resource to understanding joint control of target lncRNA expression. While only 65 RBPs were analyzed, we found that many of the RBPs bound to the same lncRNA (**Figure 3**). This is consistent with the compelling idea that lncRNAs can serve as scaffolds that assemble many relevant RBPs to regulate gene expression (Wang and Chang, 2011; Ulitsky and Bartel, 2013). At the same time, we also identified hundreds of identical binding sites that bound by multiple different RBPs in lncRNAs (**Figure 4**), probably reflecting competition among RBPs that binding on a given lncRNA.

Our combined analysis of CLIP-Seq data and GWAS data revealed hundreds of disease-related SNPs resided in the RBP binding sites of lncRNAs (Table S4 in Supplementary Material). Unlike the sporadic attempts on simply finding genetic variants associated with disease susceptibility within lncRNA genes

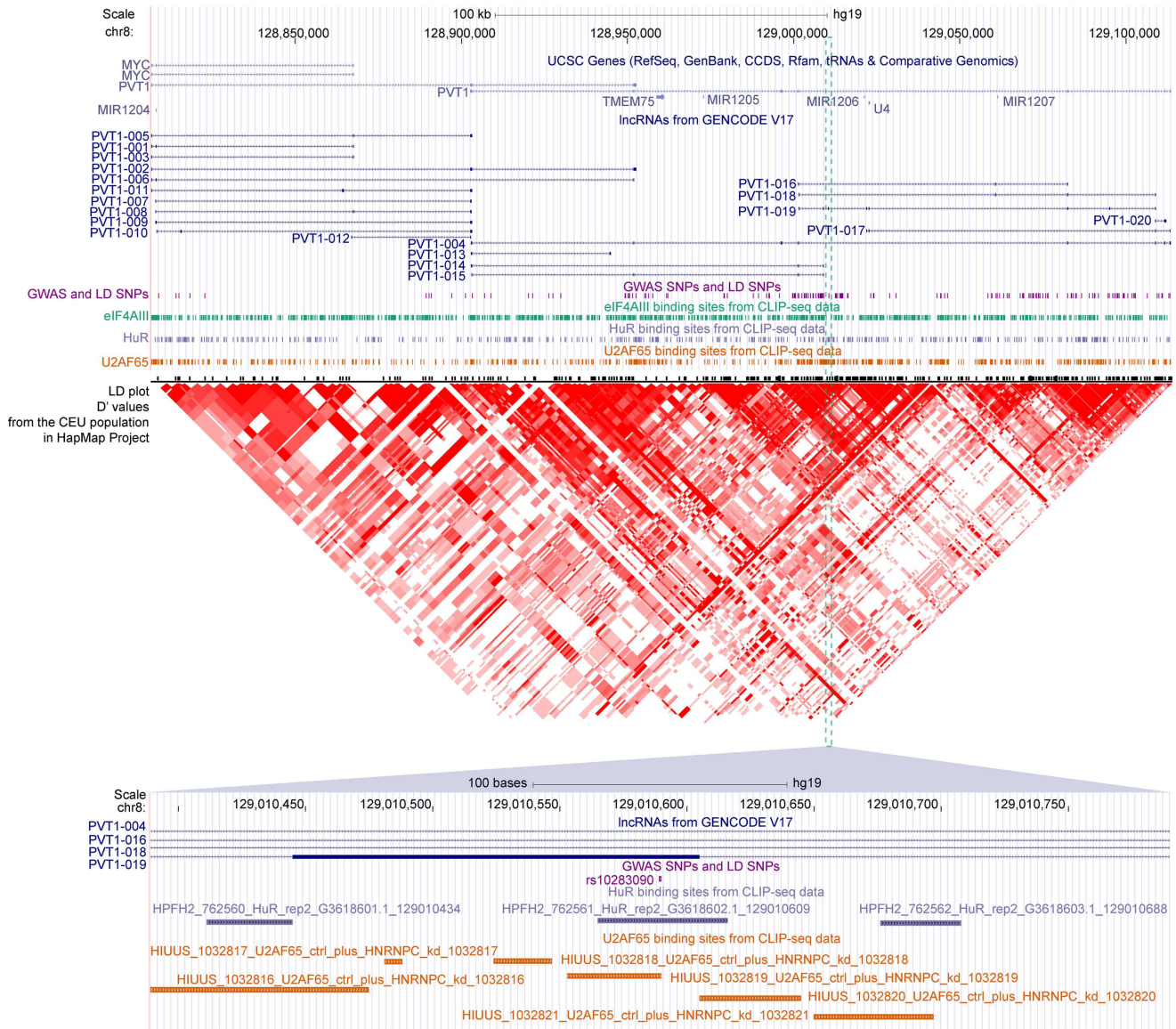


FIGURE 6 | The GWAS-associated SNPs and binding sites of three RBPs in the locus of PVT1. Gene annotations from UCSC, lncRNAs from GENCODE, GWAS, and LD SNPs, binding sites of eIF4AIII/HuR/U2AF65 and LD plot from HapMap are shown accordingly. The SNP rs10283090 overlapped with binding sites of HuR and U2AF65 are zoomed in at the bottom panel.

(Bochenek et al., 2013; Mirza et al., 2014), our approaches focused on SNPs that might impact on the binding events between RBPs and lncRNAs. Since most lncRNAs fulfill their roles through by forming complex with their protein partners, our results provide insights on the functions of lncRNAs from the perspective of RBP binding malfunction in diseases, which in turn may contribute to disease etiology.

Overall, our studies and the accompanying datasets demonstrated that one single lncRNA will generally be bound and regulated by one or multiple RBPs, the combination of which may coordinately determine the final regulatory outcome. We have also shown that an exhaustive and high-resolution RBP–lncRNA interaction map will help to discover genetic variations that contribute to complex genetic diseases by affecting post-transcriptional gene regulation.

AUTHOR CONTRIBUTIONS

Jian-Hua Yang, Liang-Hu Qu, and Jun-Hao Li conceived the project. Jun-Hao Li, Shun Liu, Ling-Ling Zheng, and Jian-Hua Yang performed the computational and statistical analysis. Jun-Hao Li, Shun Liu, Ling-Ling Zheng, Jian-Hua Yang, Liang-Hu Qu, Jie Wu, Wen-Ju Sun, Ze-Lin Wang, and Hui Zhou wrote the manuscript. Liang-Hu Qu and Jian-Hua Yang supervised the project. All authors read and approved the final manuscript.

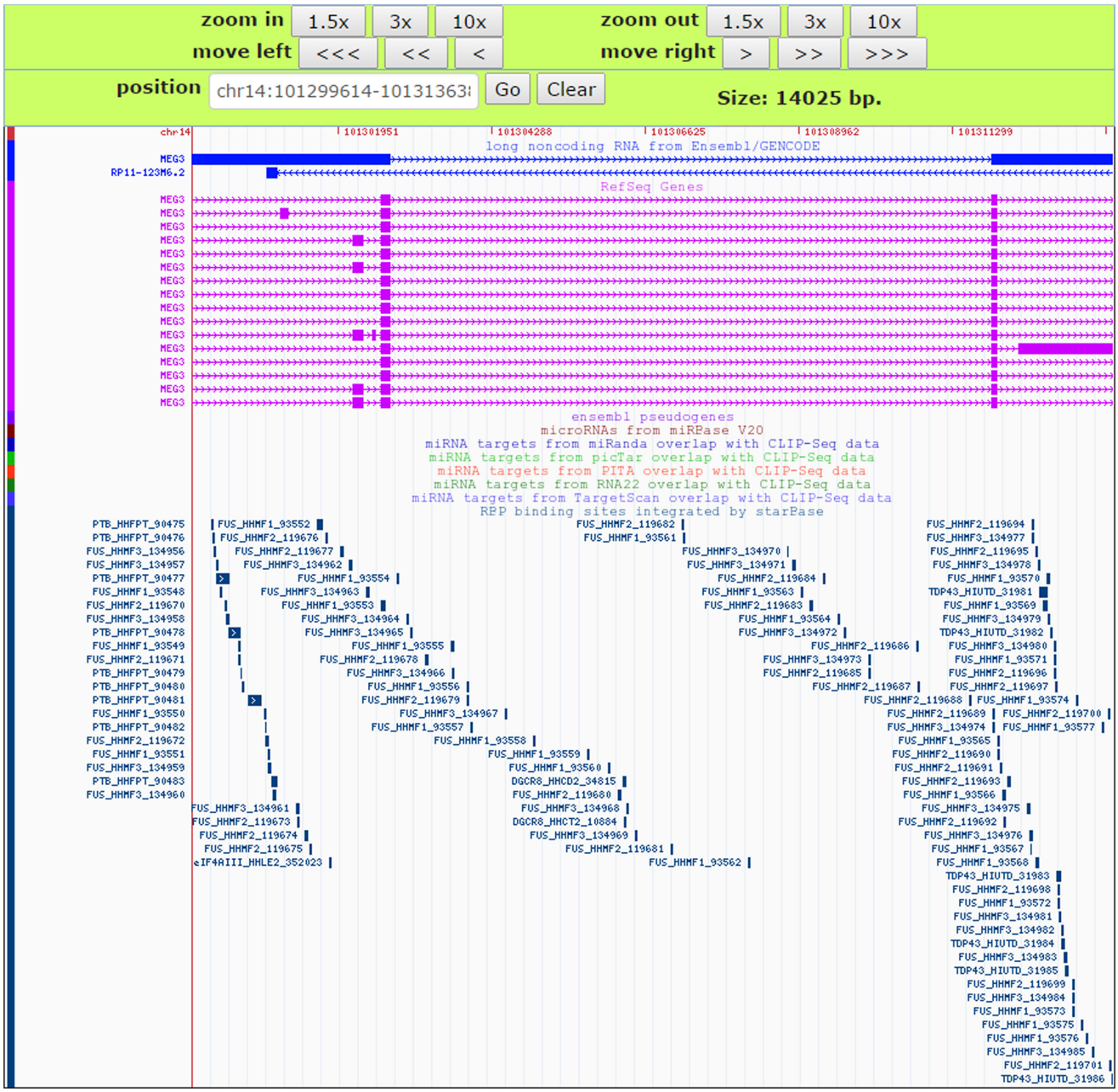


FIGURE 7 | An instance for displaying RBPs target sites in the deepView Browser of starBase V2.0. The predictive FUS binding sites on MEG3 are visible in the RBP binding sites track. In this track, the binding sites of other RBPs such as TDP-43 and PTB on MEG3 are also showed, which facilitates comparative analysis of binding events of multiple RBPs.

ACKNOWLEDGMENTS

This research is supported by the Ministry of Science and Technology of China, National Basic Research Program (No. 2011CB811300); the National Natural Science Foundation of China (No. 91440110, 31230042, 31370791, 31471223, 31401975); the funds from Guangdong Province (No. S2012010010510, S2013010012457); the project of Science and Technology New Star in ZhuJiang Guangzhou city (No. 2012J2200025); Fundamental Research Funds for the Central Universities (No. 2011330003161070, 14lgjc18); China Postdoctoral Science Foundation (No. 200902348). This research is supported in part by the Guangdong Province Key Laboratory of Computational Science and the Guangdong Province Computational Science Innovative Research Team.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00088/abstract

Table S1 | The summary of binding sites distribution across genomic features for 47 human RBPs.

Table S2 | Counts of binding RBPs in the 12,255 human lncRNAs.

Table S3 | Disease/trait related SNPs collected from four public GWAS databases.

Table S4 | Disease/trait related SNPs overlapped with RBP binding sites in lncRNAs.

REFERENCES

Alvarez, M. L., and DiStefano, J. K. (2011). Functional characterization of the plasmacytoma variant translocation 1 gene (PVT1) in diabetic nephropathy. *PLoS ONE* 6:e18671. doi:10.1371/journal.pone.0018671

Alwohhaib, M., Alwaheeb, S., Alyatama, N., Dashti, A. A., Abdelghani, A., and Hussain, N. (2014). Single nucleotide polymorphisms at erythropoietin, superoxide dismutase 1, splicing factor, arginine/serin-rich 15 and plasmacytoma variant translocation genes association with diabetic nephropathy. *Saudi J. Kidney Dis. Transpl.* 25, 577–581.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36, 431–432. doi:10.1038/ng0504-431

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300.

Bochenek, G., Hasler, R., El Mokhtari, N. E., Konig, I. R., Loos, B. G., Jepsen, S., et al. (2013). The large non-coding RNA ANRIL, which is associated with atherosclerosis, periodontitis and several forms of cancer, regulates ADIPOR1, VAMP3 and C11ORF10. *Hum. Mol. Genet.* 22, 4516–4527. doi:10.1093/hmg/ddt299

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi:10.1101/gad.17446611

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi:10.1038/nature07385

Chendrimada, T. P., Finn, K. J., Ji, X. J., Baillat, D., Gregory, R. I., Liebhaber, S. A., et al. (2007). MicroRNA silencing through RISC recruitment of eIF6. *Nature* 447, 823–U821. doi:10.1038/nature05841

Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460, 479–486. doi:10.1038/nature08170

Da Sacco, L., Baldassarre, A., and Masotti, A. (2012). Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. *Int. J. Mol. Sci.* 13, 97–114. doi:10.3390/ijms13010097

Di Ruscio, A., Ebralidze, A. K., Benoukraf, T., Amabile, G., Goff, L. A., Terragni, J., et al. (2013). DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* 503, 371–376. doi:10.1038/nature12598

Duan, R., Pak, C., and Jin, P. (2007). Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum. Mol. Genet.* 16, 1124–1131. doi:10.1093/hmg/ddm062

Fu, X.-D. (2014). Non-coding RNA: a new frontier in regulatory biology. *Natl. Sci. Rev.* 1, 190–204. doi:10.1016/j.pt.2013.04.003

Fu, X. D., and Ares, M. Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* 15, 689–701. doi:10.1038/nrg3778

Gong, C., and Maquat, L. E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3'UTRs via Alu elements. *Nature* 470, 284–288. doi:10.1038/nature09701

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477, 295–300. doi:10.1038/nature10398

Guttman, M., and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* 482, 339–346. doi:10.1038/nature10887

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141. doi:10.1016/j.cell.2010.03.009

Hanson, R. L., Craig, D. W., Millis, M. P., Yeatts, K. A., Kobes, S., Pearson, J. V., et al. (2007). Identification of PVT1 as a candidate gene for end-stage renal disease in type 2 diabetes using a pooling-based genome-wide single nucleotide polymorphism association study. *Diabetes* 56, 975–983. doi:10.2337/db06-1072

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi:10.1101/gr.135350.111

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi:10.1016/j.cell.2013.03.043

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi:10.1073/pnas.0903103106

Hu, Y., Liu, C. M., Qi, L., He, T. Z., Shi-Guo, L., Hao, C. J., et al. (2011). Two common SNPs in pri-miR-125a alter the mature miRNA expression and associate with recurrent pregnancy loss in a Han-Chinese population. *RNA Biol.* 8, 861–872. doi:10.4161/rna.8.5.16034

Huang, Y. H., Wu, C. C., Chou, C. K., and Huang, C. Y. (2011). A translational regulator, PUM2, promotes both protein stability and kinase activity of Aurora-A. *PLoS ONE* 6:e19718. doi:10.1371/journal.pone.0019718

Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., et al. (2009). Ensembl 2009. *Nucleic Acids Res.* 37, D690–D697. doi:10.1093/nar/gkn828

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi:10.1038/nature09298

Jendrzejewski, J., He, H., Radomska, H. S., Li, W., Tomsic, J., Liyanarachchi, S., et al. (2012). The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8646–8651. doi:10.1073/pnas.1205654109

Johnson, A. D., and O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC Med. Genet.* 10:6. doi:10.1186/1471-2350-10-6

Kaneko, S., Son, J., Shen, S. S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1258–1264. doi:10.1038/nsmb.2700

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672. doi:10.1073/pnas.0904715106

Kim, H. H., Kuwano, Y., Srikantan, S., Lee, E. K., Martindale, J. L., and Gorospe, M. (2009). HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev.* 23, 1743–1748. doi:10.1101/gad.1812509

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., et al. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.* 35, D145–D148. doi:10.1093/nar/gkl837

Konig, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2011). Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* 13, 77–83. doi:10.1038/nrg3141

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi:10.1101/gr.092759.109

Kumar, V., Westra, H. J., Karjalainen, J., Zhernakova, D. V., Esko, T., Hrdlickova, B., et al. (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.* 9:e1003201. doi:10.1371/journal.pgen.1003201

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

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Liao, Q., Xiao, H., Bu, D., Xie, C., Miao, R., Luo, H., et al. (2011). ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res.* 39, W118–W124. doi:10.1093/nar/gkr432

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186. doi:10.1038/ng1007-1181

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* 41, D64–D69. doi:10.1093/nar/gks1048

Mirza, A. H., Kaur, S., Brorsson, C. A., and Pociot, F. (2014). Effects of GWAS-associated genetic variants on lncRNAs within IBD and T1D candidate loci. *PLoS ONE* 9:e105723. doi:10.1371/journal.pone.0105723

Napoli, I., Mercaldo, V., Boyl, P. P., Eleuteri, B., Zalfa, F., De Rubeis, S., et al. (2008). The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. *Cell* 134, 1042–1054. doi:10.1016/j.cell.2008.07.031

Ning, S., Zhao, Z., Ye, J., Wang, P., Zhi, H., Li, R., et al. (2014). LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. *BMC Bioinformatics* 15:152. doi:10.1186/1471-2105-15-152

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

Ryan, B. M., Robles, A. I., and Harris, C. C. (2010). Genetic variation in microRNA networks: the implications for cancer research. *Nat. Rev. Cancer* 10, 389–402. doi:10.1038/nrc2867

Schmitz, K. M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* 24, 2264–2269. doi:10.1101/gad.590910

Sethupathy, P., and Collins, F. S. (2008). MicroRNA target site polymorphisms and human disease. *Trends Genet.* 24, 489–497. doi:10.1016/j.tig.2008.07.004

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* 39, 925–938. doi:10.1016/j.molcel.2010.08.011

Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46. doi:10.1016/j.cell.2013.06.020

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi:10.1016/j.molcel.2011.08.018

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120–124. doi:10.1038/nature09819

Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., et al. (2014). The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science* 344, 310–313. doi:10.1126/science.1251456

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi:10.1093/nar/gkt1229

Yin, Q. F., Yang, L., Zhang, Y., Xiang, J. F., Wu, Y. W., Carmichael, G. G., et al. (2012). Long noncoding RNAs with snoRNA ends. *Mol. Cell* 48, 219–230. doi:10.1016/j.molcel.2012.07.033

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., et al. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* 40, 939–953. doi:10.1016/j.molcel.2010.12.011

Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J., and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322, 750–756. doi:10.1126/science.1163045

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.

Received: 19 October 2014; accepted: 22 December 2014; published online: 14 January 2015.

Citation: Li J-H, Liu S, Zheng L-L, Wu J, Sun W-J, Wang Z-L, Zhou H, Qu L-H and Yang J-H (2015) Discovery of protein–lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. Front. Bioeng. Biotechnol. ***2****:88. doi: 10.3389/fbioe.2014.00088*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 14 July 2016
doi: 10.3389/fbioe.2016.00060

Discovering miRNA Regulatory Networks in Holt–Oram Syndrome Using a Zebrafish Model

Romina D'Aurizio[1], Francesco Russo[1,2], Elena Chiavacci[3], Mario Baumgart[4], Marco Groth[4], Mara D'Onofrio[5], Ivan Arisi[5], Giuseppe Rainaldi[1], Letizia Pitto[3] and Marco Pellegrini[1]**

[1] Laboratory of Integrative Systems Medicine (LISM), Institute of Informatics and Telematics (IIT), Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy, [2] Department of Computer Science, University of Pisa, Pisa, Italy, [3] Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy, [4] Leibniz Institute on Ageing, Fritz Lipmann Institute (FLI), Jena, Germany, [5] Genomics Facility, Fondazione EBRI Rita Levi-Montalcini, Roma, Italy

OPEN ACCESS

Edited by:
Christian M. Zmasek,
J. Craig Venter Institute, USA

Reviewed by:
Eirini Trompouki,
Max Planck Institute for Immunobiology and Epigenetics, Germany
Claus Jürgen Scholz,
University of Würzburg, Germany

***Correspondence:**
Letizia Pitto
l.pitto@ifc.cnr.it;
Marco Pellegrini
marco.pellegrini@iit.cnr.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 18 January 2016
Accepted: 24 June 2016
Published: 14 July 2016

Citation:
D'Aurizio R, Russo F, Chiavacci E, Baumgart M, Groth M, D'Onofrio M, Arisi I, Rainaldi G, Pitto L and Pellegrini M (2016) Discovering miRNA Regulatory Networks in Holt–Oram Syndrome Using a Zebrafish Model. Front. Bioeng. Biotechnol. 4:60. doi: 10.3389/fbioe.2016.00060

MicroRNAs (miRNAs) are small non-coding RNAs that play an important role in the post-transcriptional regulation of gene expression. miRNAs are involved in the regulation of many biological processes such as differentiation, apoptosis, and cell proliferation. miRNAs are expressed in embryonic, postnatal, and adult hearts, and they have a key role in the regulation of gene expression during cardiovascular development and disease. Aberrant expression of miRNAs is associated with abnormal cardiac cell differentiation and dysfunction. Tbx5 is a member of the T-box gene family, which acts as transcription factor involved in the vertebrate heart development. Alteration of Tbx5 level affects the expression of hundreds of genes. Haploinsufficiency and gene duplication of Tbx5 are at the basis of the cardiac abnormalities associated with Holt–Oram syndrome (HOS). Recent data indicate that miRNAs might be an important part of the regulatory circuit through which Tbx5 controls heart development. Using high-throughput technologies, we characterized genome-widely the miRNA and mRNA expression profiles in WT- and Tbx5-depleted zebrafish embryos at two crucial developmental time points, 24 and 48 h post fertilization (hpf). We found that several miRNAs, which are potential effectors of Tbx5, are differentially expressed; some of them are already known to be involved in cardiac development and functions, such as miR-30, miR-34, miR-190, and miR-21. We performed an integrated analysis of miRNA expression data with gene expression profiles to refine computational target prediction approaches by means of the inversely correlation of miRNA–mRNA expressions, and we highlighted targets, which have roles in cardiac contractility, cardiomyocyte proliferation/apoptosis, and morphogenesis, crucial functions regulated by Tbx5. This approach allowed to discover complex regulatory circuits involving novel miRNAs and protein coding genes not considered before in the HOS such as miR-34a and miR-30 and their targets.

Keywords: zebrafish, heart, microRNA, NGS, microarray, Holt–Oram, data integration

1. INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNAs of about 20–23 nucleotides that play an essential role in a variety of biological important pathways from development and physiology to diseases such as cancer (Chen and Rajewsky, 2007; Small and Olson, 2011). miRNAs are mostly known to function by targeting complementary sequences in mRNA transcripts, usually in the 3′ untranslated region (3′ UTR) and so inhibiting the translation and altering the stability of mRNA (Bartel, 2004; Yates et al., 2013). The identification and validation of miRNA–mRNA interactions is fundamental for discerning the role of miRNAs in the complex context of regulatory networks. However, since the miRNA binding is mostly not a perfect one-to-one match with the complementary target sites, it is difficult to predict miRNA targets. Consequently, several computational methods and tools have been developed in the last years (Yue et al., 2009; Peterson et al., 2014). They encompass a range of different computational approaches, from the modeling of physical interactions, exploiting common features like seed match, conservation, free energy, and site accessibility to the incorporation of less common features extracted through machine learning techniques. Computational methods predict hundreds of thousands target mRNAs per miRNA, instead the number of experimentally validated targets is very low. One possibility to reduce the false positive rate is to combine high-throughput experimental data with sequence-based predictions (Huang et al., 2007; Muniategui et al., 2013). Although, this approach does not allow to identify miRNA targets that are repressed exclusively at the translational level. Since many miRNAs cause degradation of their targets (Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010; Subtelny et al., 2014), the integration of expression profiles has been proposed to be an effective strategy to discover true miRNA–target interactions (Gennarino et al., 2009; Nazarov et al., 2013; Albert et al., 2014).

In this work, we used expression values of miRNAs and mRNAs obtained with high-throughput technologies to study complex regulatory networks altered in the Holt–Oram syndrome (HOS). HOS is a rare autosomal congenital disease characterized by cardiac and upper limb malformations (Basson et al., 1997). Mutations in the T-box gene Tbx5, which encodes a key transcription factor for vertebrate heart development, are responsible for HOS (Horb and Thomsen, 1999; Goetz et al., 2006). Family members with identical Tbx5 mutations can display large variations in malformation severity and HOS penetrance. This peculiar characteristic of HOS can be explained by the observation that Tbx5 is part of an extremely complex regulatory network. Due to the high number of messenger RNAs that are targeted by one miRNA, miRNAs are the best candidates to orchestrate the downstream regulation of Tbx5 gene expression in embryonic heart development. We have recently shown that miRNAs are crucial components of this network (Chiavacci et al., 2012, 2015). In fact, we proved that in mouse cardiac cells and zebrafish embryos, Tbx5 is able to regulate several miRNAs and, in particular, miR-218 and miR-19 (Chiavacci et al., 2012, 2015). The dysregulation of both miRNAs has a severe impact on heart development, affecting early heart morphogenesis.

As a model system, zebrafish has been extensively used for studying early vertebrate development (Kimmel et al., 1995; Yao et al., 2014) over the last 20 years. In particular, the HOS model called heartstring (hts) mutant has been well established in zebrafish, and it recapitulates almost completely the HOS characteristics. Furthermore, the zebrafish hts mutant can be easily replicated with the injection in zebrafish eggs of a specific Tbx5 morpholino (small antisense ribo-oligonucleotides, which blocks target translation) (Garrity et al., 2002).

Here, we propose an integrative approach, which uses experimental data from zebrafish HOS model system and computational methods for investigating *in vivo* complex regulatory networks perturbed in this pathology across two different stages of zebrafish development, 24 and 48 hpf. Those two stages were chosen since they mark fundamental steps in heart development. By 24 hpf, the migration phase is concluded, and the heart tube lies along the anteroposterior axis of the embryo with the atrial end to the left of the midline. By 48 hpf, the heart development is substantially completed: the heart terminated the looping phase and functional valves are formed (Kimmel et al., 1995; Yao et al., 2014). We show that it is possible to use data integration methods for studying rare diseases, providing significant insight into biological processes, and identifying new potential markers and drug targets of clinical interest.

2. MATERIALS AND METHODS

2.1. Embryos Injection

The zebrafish line used in this study is the wild-type AB strain, the animals were raised and maintained under standard laboratory conditions (Westerfield, 1993). To silence the zebrafish gene, Tbx5a we used the antisense morpholino oligonucleotide MO-Tbx5a against the translational start site of the gene, the sequence of MO-Tbx5a was 5′-GAA AGG TGT CTT CAC TGT CCG CCA T-3′ (Garrity et al., 2002). The sequence of the control morpholino, MO-Ct, was 5′-CCT CTT ACC TCA GTT ACA ATT TAT A-3′. All morpholinos were supplied by Gene Tools LLC. Zebrafish morpholinos were injected into the yolk of 1-cell stage embryos with a constant injection volume, ~1 nl, using a microinjector (Tritech Research, Los Angeles, CA, USA). Zebrafish eggs were injected with 1.5 ng of MO-Tbx5a or 1.5 ng of MO-Ct, and embryos were collected at 24 and 48 hpf.

2.2. RNA Extraction, Library Preparation, Sequencing, and Microarray

For high-throughput DNA sequencing, total RNA was extracted from batch of $n = 50$ zebrafish embryos. The library preparation was done as described in (Baumgart et al., 2012). In detail, 500 ng of total RNA was used as input material. Library preparation was done using the TruSeq Small RNA Sample Prep (Illumina). The purified libraries were quantified on the Agilent DNA 1000 chip, diluted to 10 nM and subjected to sequencing-by-synthesis on Illumina HiSeq 2000 producing single-end 51 bp read length. Two independent batches of embryos were used for MO-Tbx5a and MO-Ct at 24 hpf, one for both condition at 48 hpf.

To measure mRNA expression, the Agilent Low Input Quick Amp labeling kit was used to retrotranscribe into the cDNA (from 200 ng total RNA), amplify, and incorporate the cyanine 3-labeled CTP (cRNA). The method uses the T7 RNA polymerase, which simultaneously amplifies and incorporates cyanine 3-labeled CTP. The fluorescent cRNA was purified and hybridized to the Agilent Zebrafish V3 Gene Expression Microarray 4 × 44, according to the manufacturer protocol. Three independent batches of embryos were assessed for MO-Tbx5a, while two for MO-Ct both at 24 and 48 hpf stages. Resulting images were quantified and text files containing raw values were analyzed.

2.3. Analysis of Sequencing and Microarray Data

Raw sequences were obtained and de-multiplexed using the Illumina pipeline CASAVA v1.8.2 FastQC v0.10.1[1], which was used for quality check, and primary reads were initially trimmed off to remove adapters sequence using Cutadapt v.1.2.1 (Martin, 2011). Employing FASTX_Toolkit (0.0.13.1), the reads with N calls were discarded. Remaining high quality reads, with a minimum length of 17 bp and a maximum 38 bp after clipping, were clustered for unique hits and mapped to zebrafish pre-miRNA sequences present into the mirBase (release 20) employing miRExpress (v2.1.3; Wang et al., 2009). We allowed 95% of sequence identity between read and reference sequence and a length tolerance range of 4 bp for mapping. miRNAs expression profiles were built by calculating the sum of read counts for each miRNA according to the alignment criteria. Differential expression analysis of miRNAs identified by miRExpress was performed using Bioconductor's package DESeq (Anders and Huber, 2010). The reads count, used as measure of miRNAs quantification, was first normalized by library size factors to a common scale. The analysis was then performed and *p*-values were estimated using a negative binomial distribution model and local regression to estimate the relationship between the dispersion and the mean of each miRNAs. Raw *p*-values were finally adjusted for multiple testing using the Benjamini and Hochberg (1995) procedure controlling the false discovery rate (FDR). miRNAs with an adjusted *p*-value <0.05 were considered to be differentially expressed.

For microarrays, pre-processing of the data included background correction using a normal-exponential convolution model (offset = 16) (Ritchie et al., 2007) and cyclic loess normalization (Ballman et al., 2004) implemented in Limma package v.3.14.4 (Smyth, 2004). Low-expressed probes were filtered out keeping probes that are at least 10% brighter than the 95th percentile of the negative controls on at least 2 arrays. The Agilent Single channel Expression Microarray 4 × 44K for Zebrafish contains 39344 probes, 39162 are unique. For 35073 of them, we were able to retrieve gene accession ids corresponding to 21956 unique gene IDs. The linear modeling approach and empirical Bayes statistics implemented by Limma were used for assessing differentially expression. Finally, *p*-values were adjusted for multiple testing by means of the Benjamini and Hochberg method to control the false discovery rate. Genes with FDR less than 0.05 and fold change (FC) higher than 1.3 were selected for downstream analysis.

All statistical analyses were conducted using *R* and available Bioconductor packages.[2]

2.4. Integrated Analyses of Zebrafish miRNA and mRNA Expression Profiles

In order to discover miRNA-target pairs involved in HOS, we combined inverse correlations between miRNA and mRNA expression for improving *in silico* microRNA target predictions (see **Figure 1**). We selected the significant differential expressed miRNAs and mRNAs and performed target prediction analysis. Since miRNAs act at the post-transcriptional level downregulating their targets binding on the 3′-UTR of mRNAs, in this study, we focused our attention on these sequences that we retrieved from the UCSC Table Browser.[3] We predicted miRNA target sites in the 3′-UTR using TargetScan Fish 6.2 (Lewis et al., 2005) and Pita (Kertesz et al., 2007) algorithms and then selected the consensus. Finally, we extracted the inversely correlated interactions (to reflect the typical miRNA–mRNA relationship) obtaining the final miRNA-target list.

3. RESULTS

In the following sections, we detail the expression profiles of both miRNAs and annotated genes, which resulted altered by Tbx5a depletion during early zebrafish developmental stages (24 and 48 hpf). Small RNAseq and microarray analysis were performed to generate, respectively, miRNA and mRNA profiles. Moreover, we describe the main results obtained by integrating experimental data with computational methods to investigate *in vivo* regulatory networks modified by Tbx5 dosage alteration.

[1]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2]http://www.bioconductor.com

[3]http://genome.ucsc.edu/

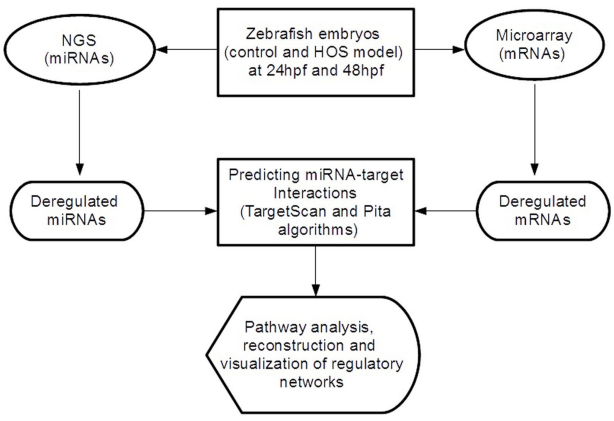


FIGURE 1 | Overview of the analytical workflow used in the study to identify inversely correlated putative target genes and to build altered regulatory networks in HOS.

3.1. Sequencing and Annotation of miRNAs Modulated by Tbx5a at 24 and 48 hpf

In order to assess miRNAs expression modulation in zebrafish embryos after Tbx5a depletion, we conducted massive parallel sequencing experiments producing between 12.7 and 25.1 million total sequencing reads were obtained for each given library (16.8 mean) and this ranged from 10.5 to 20.2 million reads of 17–38 length after adapter trimming. On average, around 760 thousands of reads mapped to zebrafish miRNAs, annotated in miRBase v20 identifying 367 mature miRNAs on average per sample (see Table S2 in Supplementary Material). Among them, 19 miRNAs resulted to be significantly modulated at 24 hpf and 33 at 48 hpf (Table S3 and S4 in Supplementary Material respectively). We selected the most variable miRNAs, in terms of expression fold-change between Tbx5 and Ct morphants for downstream analysis: miR-34a, miR-10d-5p, miR-30a, miR-210-3p, and miR-5p at 24 hpf, miR-34a, miR-462, miR-146a, miR-21, miR-7b, and miR-190b at 48 hpf (**Table 1**). Differently from experiments reported in our previous work (Chiavacci et al., 2015), the downregulation of miR-19a at 48 hpf was not significant. However, this miRNA was included in the list of miRNAs modulated by Tbx5 because: (1) Q-RT PCR analysis performed in four different sets of experiments confirmed miR-19a downregulation (**Figure 2B**), (2) this downregulation was clearly supported by physiological data and by *in situ* hybridization experiments already presented (Chiavacci et al., 2015). Besides miR-19a-3p, other seven differentially modulated miRNAs were measured by quantitative RT-PCR and fold-changes were compared in **Figures 2A,B** for 24 and 48 hpf, respectively. All modulation was confirmed except for miR-210-5p, which resulted not significative.

3.2. Tbx5 Sensitive Genes in Early Developmental Stages of Zebrafish

To characterize the gene expression profiles at 24 and 48 hpf of zebrafish development and evaluate how altered Tbx5 dosage influences the genome-wide transcription, we measured mRNAs using expression microarray technology (see Materials and Methods for details). mRNAs were extracted from zebrafish embryos injected with MO-Tbx5a or MO-Ct and collected at 24 and 48 hpf. Using an absolute FC cut-off of 1.3 and an adjusted p-value of 0.05, we identified 7100 differentially modulated genes after TBx5a silencing at 24hpf, while 2276 genes at 48 hpf. The magnitude of differential expression was formally tested to be biologically significant using the t-test relative to a threshold (TREAT) method (McCarthy and Smyth, 2009) implemented in Limma. The complete lists are available as Supplementary Material. Validation by relative Q-RT PCR was performed for some of the genes taking into consideration the relevance for the cardiac context. Q-RT PCR analysis confirmed the microarray data (**Figure 2A** for 24 hpf and **Figure 2B** for 48 hpf).

TABLE 1 | Selected differentially expressed miRNAs at 24 and 48 hpf.

Devel. stage	miRNA	FC	p-val	adj p-val
24 hpf	dre-miR-34a	2.82	1.03e−12	2.99e−10
	dre-miR-10d-5p	0.55	11.24e−07	6.77e−06
	dre-miR-30a	0.41	9.40e−12	1.02e−09
	dre-miR-210-3p	0.33	8.29e−12	1.02e−09
	dre-miR-210-5p	0.26	1.68e−10	1.22e−08
48 hpf	dre-miR-34a	6.62	7.43e−16	2.70e−14
	dre-miR-462	5.6	5.95e−10	7.63e−09
	dre-miR-146a	4.51	1.05e−09	1e27e−08
	dre-miR-21	2.84	1.65e−10	2.25e−09
	dre-miR-19a-3p[a]	0.68	8.52e−02	4.91e−02
	dre-miR-7b	0.10	1.83e−07	1.66e−06
	dre-miR-190b	0.01	1.21e−18	5.29e−17

[a]*Data for miR-19a-3p comes from our previous published data in (Chiavacci et al., 2015).*

To highlight most relevant functional categories among identified modulated genes, we performed a Functional Annotation Clustering using The Database for Annotation, Visualization and Integrated Discovery (DAVID) tool (Da Wei Huang and Lempicki, 2008). The Functional Annotation Clustering integrates the Kappa statistics to measure common genes between two annotations (e.g., ontological terms), and the fuzzy heuristic clustering to classify the groups of similar annotations according to kappa values. The resulting groups have similar biological meaning due to share similar gene members. We considered KEGG pathways and Gene Ontology terms performing the enrichment analysis of downregulated and upregulated genes (separately) at 24 and 48 hpf. Clusters that resulted significant ($p < 0.05$) are reported in **Table 2**. Interestingly, at 24 hpf, the top scoring functional categories contained genes that are involved in cell adhesion and ion binding. It is well known that morphogenesis requires specific cell adhesion molecules that are expressed in a precise developmental time, and the altered gene expression leads to heart defects (Buck et al., 1993; Kwee et al., 1995). In accordance with this observation and with our results, genes annotated with the term homophilic cell adhesion were also identified as significantly upregulated following inhibition of Tbx5a in a microarray-based expression profile performed in 56 hpf Tbx5 morphant zebrafish embryos (Mosimann et al., 2015). Furthermore, we found that the majority of modulated genes consisted of the cation and ion binding categories. In this context, the cation calcium has an important role in heart development, functions, and diseases (Arnolds et al., 2012; Crocini et al., 2014). The Calcium Binding Proteins (CaBPs) share a very similar domain organization with Calmodulin (CaM) and have been shown to have coevolved in vertebrate animals (McCue et al., 2010). The CaBPs have an important role during the development and in several diseases, such as Diastolic dysfunction that is characterized by slow or incomplete relaxation of the ventricles during diastole and is an important player in heart failure pathophysiology (Asp et al., 2013).

3.3. Changes in miRNA Expression and Integration with mRNA Profile Identify Potential miRNA–mRNA Target Pairs Involved in HOS

In this study, we integrated two target prediction algorithms, TargetScan and Pita, with miRNA and gene expression data to refine *in silico* predictions and reduce the number of false positive interactions. The resulted miRNA-target pairs consisted of 122 potential targets at 24 hpf for upregulated miRNAs and 372

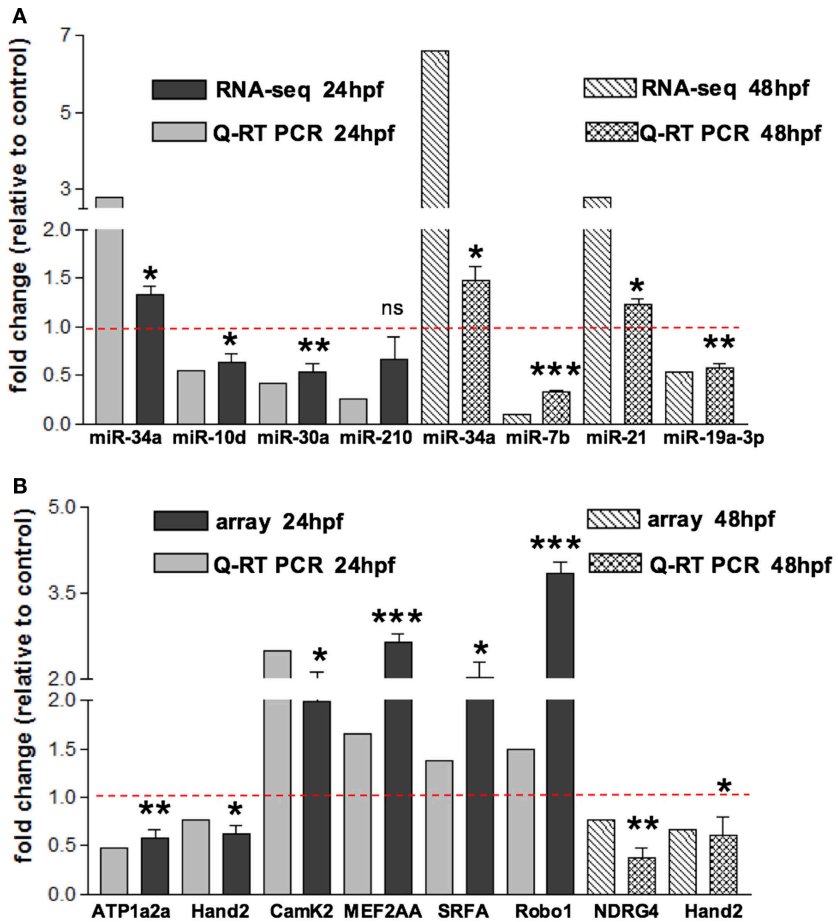


FIGURE 2 | Validation of small RNA seq profiling and array-based gene expression profiles by Quantitative RT-PCR. **(A)** Sequencing and corresponding Q-RT PCR expressions of eight of the miRNAs reported in **Table 1** and identified as differentially modulated in MO-Tbx5 vs. MO-Ct embryos at 24 hpf and at 48 hpf are reported. Values are expressed as fold change of MO-Tbx5 relative to MO-Ct. For Q-RT PCR, values are normalized on U6 expression. As pointed out in the results, miR-19a RNA-seq value is from Chiavacci et al. (2015). miR-210 is the 5p isoform. **(B)** Microarray and corresponding qRT-PCR expressions of eight genes showing differential expression in MO-Tbx5 compared to MO-Ct embryos at 24 and 48 hpf are reported. Values are expressed as fold change of MO-Tbx5 relative to MO-Ct. For Q-RT PCR, values are normalized on EF1, beta actin, and 18S expression. The values reported in the Q-RT PCR analysis are the mean of at least three independent microinjection experiments, *t*-test was used for statistical analysis: $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$.

potential targets for downregulated miRNAs (complete lists are in Table S5 and S6 in Supplementary Material). At 48 hpf, we discovered 142 potential targets for upregulated miRNAs and 162 for downregulated miRNAs (see Table S7 and S8 in Supplementary Material). Among them, several miRNA–mRNA interactions involved genes that are known to be connected to heart development or cardiac functions published in previous works (Table S9 and S10 in Supplementary Material). We summarized these finding in **Figures 3** and **4** and explore most interesting functional relations in the next section.

4. DISCUSSION

Tbx5 is a crucial transcription factor in heart development. In HOS murine model, it has been shown that even small alterations of this gene cause modulation of hundreds of genes (Mori et al., 2006). It has been suggested that the strong impact that Tbx5 has on gene expression is mainly the result of its ability to modulate other regulators, such as different transcription factors, in a very complex regulatory network. Our previous studies suggest that Tbx5 affects the embryo development by modulating also miRNAs (Chiavacci et al., 2012). Moreover, the fact that miR-19a replacement is able to partially rescue fins and cardiac defects in a zebrafish model of HOS, strongly supports the importance of miRNAs in Tbx5 regulatory circuits (Chiavacci et al., 2015).

In this study, we analyzed miRNA and mRNA expression profiles at two fundamental time points (24 and 48 hpf) of zebrafish embryos development after depletion of Tbx5 and compared them to the wild-type ones. We employed expression data to improve miRNA-target predictions of computational sequence-based methods by means of anticorrelation analysis of miRNA–mRNA expression levels. Repression by animal miRNAs, differently from plant miRNAs, leads to decreased translational efficiency and/or decreased mRNA levels. Although, the relative contributions of these two outcomes is still unknown and increasing experimental evidences show that changes in mRNA levels closely reflects the

TABLE 2 | Most significant categories from functional annotation clustering analysis of the deregulated transcripts were reported.

Time-regulation	Cluster	Term	Benjamini *p* value	Fold-enrich.
24 hpf, up genes	c1	GO:0007155 cell adhesion	6.97e–03	2.11
	c1	GO:0022610 biological adhesion	6.97e–03	2.11
	c1	GO:0007156 homophilic cell adhesion	8.87e–03	3.27
	c1	GO:0016337 cell–cell adhesion	1.28e–02	2.89
	c2	GO:0008270 zinc ion binding	5.71e–03	1.33
	c2	GO:0046914 transition metal ion binding	4.12e–03	1.28
	c3	GO:0006468 protein amino acid phosphorylation	1.24e–02	1.67
	c3	GO:0016310 phosphorylation	2.51e–02	1.57
24 hpf, down genes	c1	GO:0044429 mitochondrial part	2.56e–03	1.81
	c1	GO:0005739 mitochondrion	3.97e–03	1.62
	c1	GO:0031975 envelope	2.07e–02	1.59
	c1	GO:0031967 organelle envelope	2.10e–02	1.60
	c1	GO:0019866 organelle inner membrane	2.33e–02	1.88
	c1	GO:0005743 mitochondrial inner membrane	2.40e–02	1.88
	c1	GO:0005740 mitochondrial envelope	2.55e–02	1.74
	c1	GO:0031966 mitochondrial membrane	2.97e–02	1.71
	c2	GO:0046872 metal ion binding	4.10e–02	1.19
	c3	GO:0004672 protein kinase activity	4.90e–02	1.56
48 hpf, down genes	c1	GO:0043565 sequence-specific DNA binding	1.02e–08	2.25
	c1	GO:0003700 transcription factor activity	1.24e–06	1.94
	c1	GO:0030528 transcription regulator activity	2.57e–06	1.76
	c1	GO:0051252 regulation of RNA metabolic process	7.00e–05	1.71
	c1	GO:0006355 regulation of transcription. DNA-dependent	1.07e–04	1.72
	c1	GO:0003677 DNA binding	3.12e–03	1.44
	c1	GO:0045449 regulation of transcription	7.05e–03	1.46
	c2	GO:0019825 oxygen binding	1.07e–04	9.26
	c2	GO:0005344 oxygen transporter activity	1.07e–04	9.26
	c2	GO:0005833 hemoglobin complex	2.57e–04	10.68
	c2	GO:0015669 gas transport	4.79e–04	8.92
	c2	GO:0015671 oxygen transport	4.79e–04	8.92
	c3	dre00010: glycolysis/gluconeogenesis	2.85e–03	3.99

impact of miRNAs on gene expression suggesting that destabilization of target mRNAs by exonucleolytic activity is the main mechanism to decrease protein output (Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010; Subtelny et al., 2014). Therefore, the anticorrelation analysis of miRNA–mRNA expression levels may contribute to elucidate large portion of miRNA–mRNA regulatory networks affected by pathological conditions. This approach allowed us to identify putative miRNA–target interactions, and cardiac transcription factors that are particularly interesting in the context of HOS.

One of the most interesting miRNA identified as upregulated both at 24 and 48 hpf was miR-34a. MiR-34 family members (miR-34a, -34b, and -34c) are upregulated in the heart in response to stress and the silencing of the entire miR-34 family could protect the heart against pathological cardiac remodeling and improve cardiac functions (Bernardo et al., 2012). Moreover, miR-34a is induced in the aging heart and *in vivo* silencing of miR-34a reduces age-associated cardiomyocyte cell death. The inhibition of miR-34a reduces cell death and fibrosis following acute myocardial infarction and improves recovery of myocardial function (Boon et al., 2013). These recent studies show an emerging role of miR-34a (and the miR-34 family) as potential regulator of heart remodeling. Therapies that inhibit miR-34a could be useful for cardiac pathologies and HOS.

We discovered several potential miR-34a targets with a possible connection with heart development and HOS (**Figure 2**). The *ATPase Na^+/K^+ transporting, alpha 2a polypeptide (ATP1a2a)* transcript is downregulated in 24 hpf Tbx5a morphants. Since the ATP1a2a contributes to the Ca homeostasis by pumping sodium ions (Na^+) out of cells and potassium ions (K^+) into cells, a downregulation of this enzyme might have a negative impact in cardiac contractility and control of arrhythmias. This observation is consistent with the crucial role of Tbx5 in the regulation of cardiac contraction in embryos and in adults. Interestingly, HOS patients show diastolic filling abnormalities and downregulation of ATP2a2, which regulates Ca fluxes in the SR (Zhu et al., 2008).

Furthermore, our data suggest that miR-34a might impact cardiac contraction by regulating a member of the *ras homolog gene family*, rhoad. RhoA, controlling the Rho-kinase pathway, plays an important role in various fundamental cellular functions, including contraction and motility (Satoh et al., 2011). Moreover, in line with the pro apoptotic role exerted by miR-34a (Raver-Shapira et al., 2007), we observed upregulation of *dual-specificity tyrosine-(Y)-phospho-regulated kinase 2 (dyrk2)*, putative miR-30a target, which negatively regulates the cardiomyocyte growth by mediating repressor function of GSK-3 beta on eIF2B (Weiss et al., 2013) and upregulation of *caspase 8*, putative target of miR-219-5p downregulated at 48 hpf (**Figure 3**).

As expected, several genes affecting cardiac morphogenesis were identified in our analysis (**Figures 2** and **3**). Specifically *Roundabout Guidance Receptor 1 (Robo1)*, which is involved in heart tube formation in zebrafish (Fish et al., 2011) and tmem2,

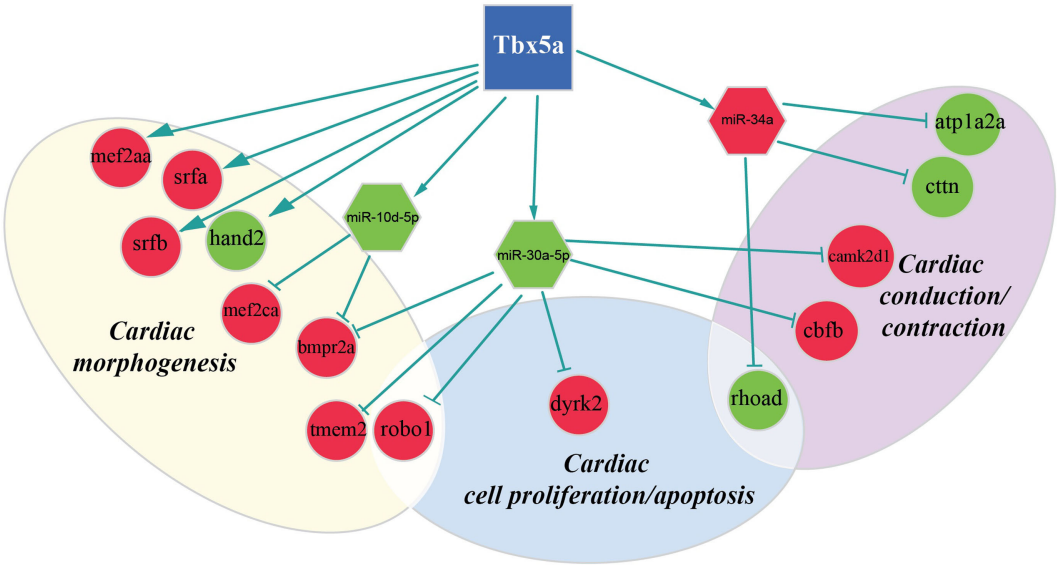


FIGURE 3 | Regulatory network altered in zebrafish HOS model at 24 hpf. Potential interactions involving Tbx5, transcriptional factors, miRNAs, and their targets are shown together with the functional impact in heart development.

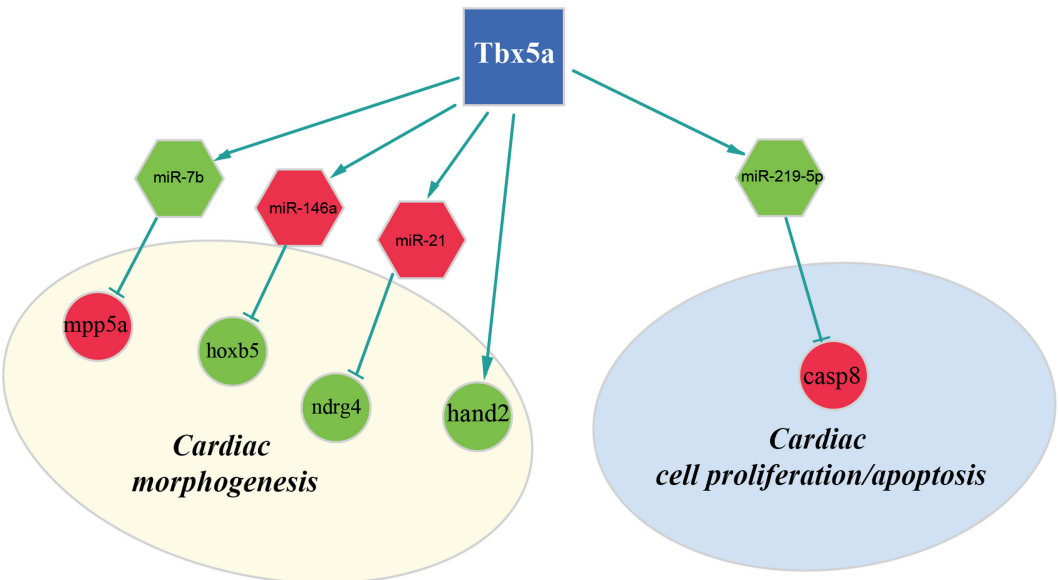


FIGURE 4 | Regulatory network altered in zebrafish HOS model at 48 hpf. Potential interactions involving Tbx5, miRNAs, and their targets are shown together with the functional impact in heart development.

whose expression in myocardial and endocardial tissues in zebrafish and mouse is required for regionally restrict atrioventricular canal boundary and endocardial cushion development. Both genes are putative miR-30a targets at 24 hpf.

Recently, a role for Tbx5 in the establishment of correct heart asymmetry in zebrafish embryos has been highlighted (Pi-Roig et al., 2014). Our data suggest that miR-30a and miR-10d might be contribute to this specific Tbx5 function by controlling respectively bmpr2a (Monteiro et al., 2008) and camk2d1 (Francescatto et al., 2010).

Another interesting miRNA that we found upregulated at 48 hpf is miR-21 whose deregulation in heart has been reported to contribute to cardiovascular disease (Jazbutyte and Thum, 2010). More recently, a crucial role of this miRNA in heart valve formation has been also shown in zebrafish (Banjo et al., 2013), and the alteration of cardiac valve morphology is one

of the hallmark of zebrafish HOS phenotype (Camarata et al., 2010; Chiavacci et al., 2012). Two predicted targets of miR-21 are NDRG1B and NDRG4, members of the *N-myc downstream regulated gene (NDRG) family*, which are downregulated in Tbx5 morphants at 48 hpf. Alterations of NDRG4 cause several of the cardiac defects that characterize the heartstring mutants and are significantly decreased in hearts with reduced Tbx5 activities (Qu et al., 2008). Therefore, we hypothesized that Tbx5 might affect NDRG4 expression through miR-21 modulation.

Although in this study, we used whole embryos for our analysis, we discovered important alterations on transcription factors with crucial roles in heart development. In particular, we observed downregulation both at 24and 48 hpf of the *bHLH transcription factor Hand2* (Yelon et al., 2000). Mutations in the hands off locus, which encodes for this transcription factor, cause defects in myocardial development in an early stage, produce a reduced number of myocardial precursors, and show delayed differentiation of the pectoral fin mesenchyme (Schindler et al., 2014). All these phenotypic characteristics are in line with the observed Tbx5 morphant phenotype. In HOS mouse hearts, a strong downregulation of Hand1 was observed (Mori et al., 2006). In mouse, Hand1 and Hand2 are members of the *Hand subfamily* and have partially redundant functions (Yelon et al., 2000; Tamura et al., 2014). However, in zebrafish, only one member of the hand family has been identified, the Hand2 transcription factor, which is able to perform several of the functions that in mammals are Hand1 specific (Togi et al., 2006; Reichenbach et al., 2008). Therefore, modulation of Hand1 in mouse or Hand2 in zebrafish might have similar functional consequences. Indeed, it has been shown that Hand2 is able to downregulate Nppa, a direct target of Tbx5 and Irx4 an other important cardiac transcription factor strongly downregulated in HOS mouse heart (Bruneau et al., 2001; Mori et al., 2006). In our analysis, we were not able to detect a significant modulation of Nppa gene whose expression is restricted to the cardiac tissue. On the contrary, at 48 hpf, we detected a downregulation of Irx4. This gene is not only expressed in the heart tissues but also present in the eye, a district which is relatively large at this time of development and where Tbx5 is functionally active.

Differently from Hand2, MEF2AA, MEF2CA, SRFB, and SRFA resulted upregulated in 24 hpf embryos depleted for Tbx5a. Among them, MEF2CA is a putative target of miR-10d, both miRNAs already mentioned as downregulated at 24 hpf. All of them codify for transcription factors largely expressed in mesodermal tissues and involved in cardiac developmental patterns highly active at 24 hpf. Consequently, alterations in the expression of these factors have important effects on cardiac development. However, it is hard to predict whether dysregulation of these genes might have positive or negative regulatory effects on their targets. For Tbx5 direct interactors, such as MEF2CA, the ratio between interactors seems more important than the absolute level of the specific factor (Takeuchi et al., 2011). For SRFs, it has been shown that a mild increase may pose either positive and/or negative modulatory effects on their target activation depending on the cofactor recruited (Miano, 2003; Zhang et al., 2003). Interestingly, a negative functional cooperator of SRF is SRFBP1 that we identified as upregulated in 24 hpf Tbx5a morphants. SRFBP1 is highly expressed in fetal and adult mouse heart and functionally interacts with SRF and myocardin in repressing the atrial natriuretic factor promoter activity (Zhang et al., 2004). The data suggest that the observed mild increase of SRF and SRFBP1 in zebrafish Tbx5a morphants might contribute in the downregulation of Nppa, which characterizes the HOS disease.

In conclusion, in this study, we proposed an integrative analysis of miRNA and mRNA expression profiles in a zebrafish model to study the impact of the downregulation of Tbx5 responsible of the HOS. We found several deregulated transcripts including important transcription factors for heart development and diseases, and several deregulated miRNAs with a potential role in the pathology. This model uncovered novel miRNAs and protein coding genes not considered before in the HOS such as miR-34a and miR-30 and their targets. Further dissection of these regulatory circuits will shed light on fundamental pathways in heart development that can contribute to the pathogenesis of human heart diseases. Identification of new TBX5 targets might not only help understand the complexity of HOS phenotype but also contribute in finding novel therapeutic strategies to treat congenital disease. Future experiments are needed to test the role of the identified miRNAs regulated by Tbx5 and the effects on their downstream targets.

4.1. Data Accession Codes

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Super Series accession number GSE64466.[4]

AUTHOR CONTRIBUTIONS

RD and LP conceived the work and interpreted the data; RD analyzed sequencing and microarray data; FR performed the integrative analysis; EC performed the *in vivo* experiments; MB and MG carried out sequencing experiments, MD and IA carried out microarray experiments; RD, FR, EC, and LP wrote the manuscript; GR, LP, and MP supervised the work; all authors read and approved the final manuscript.

ACKNOWLEDGMENTS

FR has been supported by a fellowship sponsored by Progetto Istituto Toscano Tumori Grant 2012 Prot.A00GRT.

FUNDING

The present work is partially supported by the Flagship project InterOmics (PB.P05, CUP B91J12000270001), funded by the Italian MIUR and CNR organizations, and by the joint IIT-IFC Laboratory of Integrative Systems Medicine (LISM).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fbioe.2016.00060

[4]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64466

REFERENCES

Albert, M. H., Mannert, J., Fleischmann, K. K., Schiemann, M., Pagel, P., Schmid, I., et al. (2014). Mirnome and transcriptome aided pathway analysis in human regulatory t cells. *Genes Immun.* 15, 303–312. doi:10.1038/gene.2014.20

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106

Arnolds, D. E., Liu, F., Fahrenbach, J. P., Kim, G. H., Schillinger, K. J., Smemo, S., et al. (2012). Tbx5 drives scn5a expression to regulate cardiac conduction system function. *J. Clin. Invest.* 122, 2509–2518. doi:10.1172/JCI62617

Asp, M. L., Martindale, J. J., Heinis, F. I., Wang, W., and Metzger, J. M. (2013). Calcium mishandling in diastolic dysfunction: mechanisms and potential therapies. *Biochim. Biophys. Acta* 1833, 895–900. doi:10.1016/j.bbamcr.2012.09.007

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature* 455, 64–71. doi:10.1038/nature07242

Ballman, K. V., Grill, D. E., Oberg, A. L., and Therneau, T. M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics* 20, 2778–2786. doi:10.1093/bioinformatics/bth327

Banjo, T., Grajcarek, J., Yoshino, D., Osada, H., Miyasaka, K. Y., Kida, Y. S., et al. (2013). Haemodynamically dependent valvulogenesis of zebrafish heart is mediated by flow-dependent expression of mir-21. *Nat. Commun.* 4, 1978. doi:10.1038/ncomms2978

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi:10.1016/S0092-8674(04)00045-5

Basson, C. T., Bachinsky, D. R., Lin, R. C., Levi, T., Elkins, J. A., Soults, J., et al. (1997). Mutations in human cause limb and cardiac malformation in holt-oram syndrome. *Nat. Genet.* 15, 30–35. doi:10.1038/ng0197-30

Baumgart, M., Groth, M., Priebe, S., Appelt, J., Guthke, R., Platzer, M., et al. (2012). Age-dependent regulation of tumor-related microRNAs in the brain of the annual fish *Nothobranchius furzeri*. *Mech. Ageing Dev.* 133, 226–233. doi:10.1016/j.mad.2012.03.015

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300.

Bernardo, B. C., Gao, X.-M., Winbanks, C. E., Boey, E. J., Tham, Y. K., Kiriazis, H., et al. (2012). Therapeutic inhibition of the mir-34 family attenuates pathological cardiac remodeling and improves heart function. *Proc. Natl. Acad. Sci. U.S.A.* 109, 17615–17620. doi:10.1073/pnas.1206432109

Boon, R. A., Iekushi, K., Lechner, S., Seeger, T., Fischer, A., Heydt, S., et al. (2013). MicroRNA-34a regulates cardiac ageing and function. *Nature* 495, 107–110. doi:10.1038/nature11919

Bruneau, B. G., Nemer, G., Schmitt, J. P., Charron, F., Robitaille, L., Caron, S., et al. (2001). A murine model of holt-oram syndrome defines roles of the t-box transcription factor tbx5 in cardiogenesis and disease. *Cell* 106, 709–721. doi:10.1016/S0092-8674(01)00493-7

Buck, C. A., Baldwin, H., DeLisser, H., Mickanin, C., Shen, H., Kennedy, G., et al. (1993). Cell adhesion receptors and early mammalian heart development: an overview. *C. R. Acad. Sci. III* 316, 838–859.

Camarata, T., Krcmery, J., Snyder, D., Park, S., Topczewski, J., and Simon, H.-G. (2010). Pdlim7 (LMP4) regulation of tbx5 specifies zebrafish heart atrioventricular boundary and valve formation. *Dev. Biol.* 337, 233–245. doi:10.1016/j.ydbio.2009.10.039

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* 8, 93–103. doi:10.1038/nrg1990

Chiavacci, E., D'Aurizio, R., Guzzolino, E., Russo, F., Baumgart, M., Groth, M., et al. (2015). MicroRNA 19a replacement partially rescues fin and cardiac defects in zebrafish model of holt oram syndrome. *Sci. Rep.* 5, 18240. doi:10.1038/srep18240

Chiavacci, E., Dolfi, L., Verduci, L., Meghini, F., Gestri, G., Evangelista, A. M. M., et al. (2012). MicroRNA 218 mediates the effects of tbx5a over-expression on zebrafish heart development. *PLoS ONE* 7:e50536. doi:10.1371/journal.pone.0050536

Crocini, C., Coppini, R., Ferrantini, C., Yan, P., Loew, L. M., Tesi, C., et al. (2014). Defects in t-tubular electrical activity underlie local alterations of calcium release in heart failure. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15196–15201. doi:10.1073/pnas.1411557111

Da Wei Huang, B. T. S., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Fish, J. E., Wythe, J. D., Xiao, T., Bruneau, B. G., Stainier, D. Y. R., Srivastava, D., et al. (2011). A slit/miR-218/robo regulatory loop is required during heart tube formation in zebrafish. *Development* 138, 1409–1419. doi:10.1242/dev.060046

Francescatto, L., Rothschild, S. C., Myers, A. L., and Tombes, R. M. (2010). The activation of membrane targeted camk-ii in the zebrafish kupffer's vesicle is required for left-right asymmetry. *Development* 137, 2753–2762. doi:10.1242/dev.049627

Garrity, D. M., Childs, S., and Fishman, M. C. (2002). The heartstrings mutation in zebrafish causes heart/fin Tbx5 deficiency syndrome. *Development* 129, 4635–4645.

Gennarino, V. A., Sardiello, M., Avellino, R., Meola, N., Maselli, V., Anand, S., et al. (2009). MicroRNA target prediction by expression analysis of host genes. *Genome Res.* 19, 481–490. doi:10.1101/gr.084129.108

Goetz, S. C., Brown, D. D., and Conlon, F. L. (2006). Tbx5 is required for embryonic cardiac cell cycle progression. *Development* 133, 2575–2584. doi:10.1242/dev.02420

Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835–840. doi:10.1038/nature09267

Hendrickson, D. G., Hogan, D. J., McCullough, H. L., Myers, J. W., Herschlag, D., Ferrell, J. E., et al. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol.* 7:e1000238. doi:10.1371/journal.pbio.1000238

Horb, M. E., and Thomsen, G. H. (1999). Tbx5 is essential for heart development. *Development* 126, 1739–1751.

Huang, J. C., Babak, T., Corson, T. W., Chua, G., Khan, S., Gallie, B. L., et al. (2007). Using expression profiling data to identify human microRNA targets. *Nat. Methods* 4, 1045–1049. doi:10.1038/nmeth1130

Jazbutyte, V., and Thum, T. (2010). MicroRNA-21: from cancer to cardiovascular disease. *Curr. Drug Targets* 11, 926–935. doi:10.2174/138945010791591403

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39, 1278–1284. doi:10.1038/ng2135

Kimmel, C. B., Ballard, W. W., Kimmal, S. R., Ullmann, B., and Schilling, T. F. (1995). Stages of embryonic-development of the zebrafish. *Dev. Dyn.* 203, 253–310. doi:10.1002/aja.1002030302

Kwee, L., Baldwin, H. S., Shen, H. M., Stewart, C. L., Buck, C., Buck, C. A., et al. (1995). Defective development of the embryonic and extraembryonic circulatory systems in vascular cell adhesion molecule (VCAM-1) deficient mice. *Development* 121, 489–503.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi:10.1016/j.cell.2004.12.035

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* 17, 10–12. doi:10.14806/ej.17.1.200

McCarthy, D. J., and Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a treat. *Bioinformatics* 25, 765–771. doi:10.1093/bioinformatics/btp053

McCue, H. V., Haynes, L. P., and Burgoyne, R. D. (2010). Bioinformatic analysis of CaBP/calneuron proteins reveals a family of highly conserved vertebrate Ca2+-binding proteins. *BMC Res. Notes* 3:118. doi:10.1186/1756-0500-3-118

Miano, J. M. (2003). Serum response factor: toggling between disparate programs of gene expression. *J. Mol. Cell. Cardiol.* 35, 577–593. doi:10.1016/S0022-2828(03)00110-X

Monteiro, R., van Dinther, M., Bakkers, J., Wilkinson, R., Patient, R., ten Dijke, P., et al. (2008). Two novel type II receptors mediate bmp signalling and are required to establish left-right asymmetry in zebrafish. *Dev. Biol.* 315, 55–71. doi:10.1016/j.ydbio.2007.11.038

Mori, A. D., Zhu, Y., Vahora, I., Nieman, B., Koshiba-Takeuchi, K., Davidson, L., et al. (2006). Tbx5-dependent rheostatic control of cardiac gene expression and morphogenesis. *Dev. Biol.* 297, 566–586. doi:10.1016/j.ydbio.2006.05.023

Mosimann, C., Panáková, D., Werdich, A. A., Musso, G., Burger, A., Lawson, K. L., et al. (2015). Chamber identity programs drive early functional partitioning of the heart. *Nat. Commun.* 6, 8146. doi:10.1038/ncomms9146

Muniategui, A., Pey, J., Planes, F. J., and Rubio, A. (2013). Joint analysis of miRNA and mRNA expression data. *Brief. Bioinformatics* 14, 263–278. doi:10.1093/bib/bbs028

Nazarov, P. V., Reinsbach, S. E., Muller, A., Nicot, N., Philippidou, D., Vallar, L., et al. (2013). Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function. *Nucleic Acids Res.* 41, 2817–2831. doi:10.1093/nar/gks1471

Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L., and Congdon, C. B. (2014). Common features of microRNA target prediction tools. *Front. Genet.* 5:23. doi:10.3389/fgene.2014.00023

Pi-Roig, A., Martin-Blanco, E., and Minguillon, C. (2014). Distinct tissue-specific requirements for the zebrafish tbx5 genes during heart, retina and pectoral fin development. *Open Biol.* 4, 140014. doi:10.1098/rsob.140014

Qu, X., Jia, H., Garrity, D. M., Tompkins, K., Batts, L., Appel, B., et al. (2008). Ndrg4 is required for normal myocyte proliferation during early cardiac development in zebrafish. *Dev. Biol.* 317, 486–496. doi:10.1016/j.ydbio.2008.02.044

Raver-Shapira, N., Marciano, E., Meiri, E., Spector, Y., Rosenfeld, N., Moskovits, N., et al. (2007). Transcriptional activation of miR-34a contributes to p53-mediated apoptosis. *Mol. Cell* 26, 731–743. doi:10.1016/j.molcel.2007.05.017

Reichenbach, B., Delalande, J.-M., Kolmogorova, E., Prier, A., Nguyen, T., Smith, C. M., et al. (2008). Endoderm-derived sonic hedgehog and mesoderm Hand2 expression are required for enteric nervous system development in zebrafish. *Dev. Biol.* 318, 52–64. doi:10.1016/j.ydbio.2008.02.061

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., et al. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 23, 2700–2707. doi:10.1093/bioinformatics/btm412

Satoh, K., Fukumoto, Y., and Shimokawa, H. (2011). Rho-kinase: important new therapeutic target in cardiovascular diseases. *Am. J. Physiol. Heart Circ. Physiol.* 301, H287–H296. doi:10.1152/ajpheart.00327.2011

Schindler, Y. L., Garske, K. M., Wang, J., Firulli, B. A., Firulli, A. B., Poss, K. D., et al. (2014). Hand2 elevates cardiomyocyte production during zebrafish heart development and regeneration. *Development* 141, 3112–3122. doi:10.1242/dev.106336

Small, E. M., and Olson, E. N. (2011). Pervasive roles of microRNAs in cardiovascular biology. *Nature* 469, 336–342. doi:10.1038/nature09783

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, Article3. doi:10.2202/1544-6115.1027

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., and Bartel, D. P. (2014). Poly(a)-tail profiling reveals an embryonic switch in translational control. *Nature* 508, 66–71. doi:10.1038/nature13007

Takeuchi, J. K., Lou, X., Alexander, J. M., Sugizaki, H., Delgado-Olguín, P., Holloway, A. K., et al. (2011). Chromatin remodelling complex dosage modulates transcription factor function in heart development. *Nat. Commun.* 2, 187. doi:10.1038/ncomms1187

Tamura, M., Amano, T., and Shiroishi, T. (2014). The Hand2 gene dosage effect in developmental defects and human congenital disorders. *Curr. Top. Dev. Biol.* 110, 129–152. doi:10.1016/B978-0-12-405943-6.00003-8

Togi, K., Yoshida, Y., Matsumae, H., Nakashima, Y., Kita, T., and Tanaka, M. (2006). Essential role of Hand2 in interventricular septum formation and trabeculation during cardiac development. *Biochem. Biophys. Res. Commun.* 343, 144–151. doi:10.1016/j.bbrc.2006.02.122

Wang, W.-C., Lin, F.-M., Chang, W.-C., Lin, K.-Y., Huang, H.-D., and Lin, N.-S. (2009). miRExpress: analyzing high-throughput sequencing data for profiling microrna expression. *BMC Bioinformatics* 10:328. doi:10.1186/1471-2105-10-328

Weiss, C. S., Ochs, M. M., Hagenmueller, M., Streit, M. R., Malekar, P., Riffel, J. H., et al. (2013). Dyrk2 negatively regulates cardiomyocyte growth by mediating repressor function of GSK-3 on eIF2B. *PLoS ONE* 8:e70848. doi:10.1371/journal.pone.0070848

Westerfield, M. (1993). *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Brachydanio rerio)*. Oregon: University of Oregon Press Eugene.

Yao, Y., Ma, L., Jia, Q., Deng, W., Liu, Z., Zhang, Y., et al. (2014). Systematic characterization of small RNAome during zebrafish early developmental stages. *BMC Genomics* 15:117. doi:10.1186/1471-2164-15-117

Yates, L. A., Norbury, C. J., and Gilbert, R. J. C. (2013). The long and short of microRNA. *Cell* 153, 516–519. doi:10.1016/j.cell.2013.04.003

Yelon, D., Ticho, B., Halpern, M. E., Ruvinsky, I., Ho, R. K., Silver, L. M., et al. (2000). The bHLH transcription factor hand2 plays parallel roles in zebrafish heart and pectoral fin development. *Development* 127, 2573–2582.

Yue, D., Liu, H., and Huang, Y. (2009). Survey of computational algorithms for microRNA target prediction. *Curr. Genomics* 10, 478–492. doi:10.2174/138920209789208219

Zhang, X., Azhar, G., Furr, M. C., Zhong, Y., and Wei, J. Y. (2003). Model of functional cardiac aging: young adult mice with mild overexpression of serum response factor. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 285, R552–R560. doi:10.1152/ajpregu.00631.2002

Zhang, X., Azhar, G., Zhong, Y., and Wei, J. Y. (2004). Identification of a novel serum response factor cofactor in cardiac gene regulation. *J. Biol. Chem.* 279, 55626–55632. doi:10.1074/jbc.M405945200

Zhu, Y., Gramolini, A. O., Walsh, M. A., Zhou, Y.-Q., Slorach, C., Friedberg, M. K., et al. (2008). Tbx5-dependent pathway regulating diastolic function in congenital heart disease. *Proc. Natl. Acad. Sci. U.S.A.* 105, 5519–5524. doi:10.1073/pnas.0801779105

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