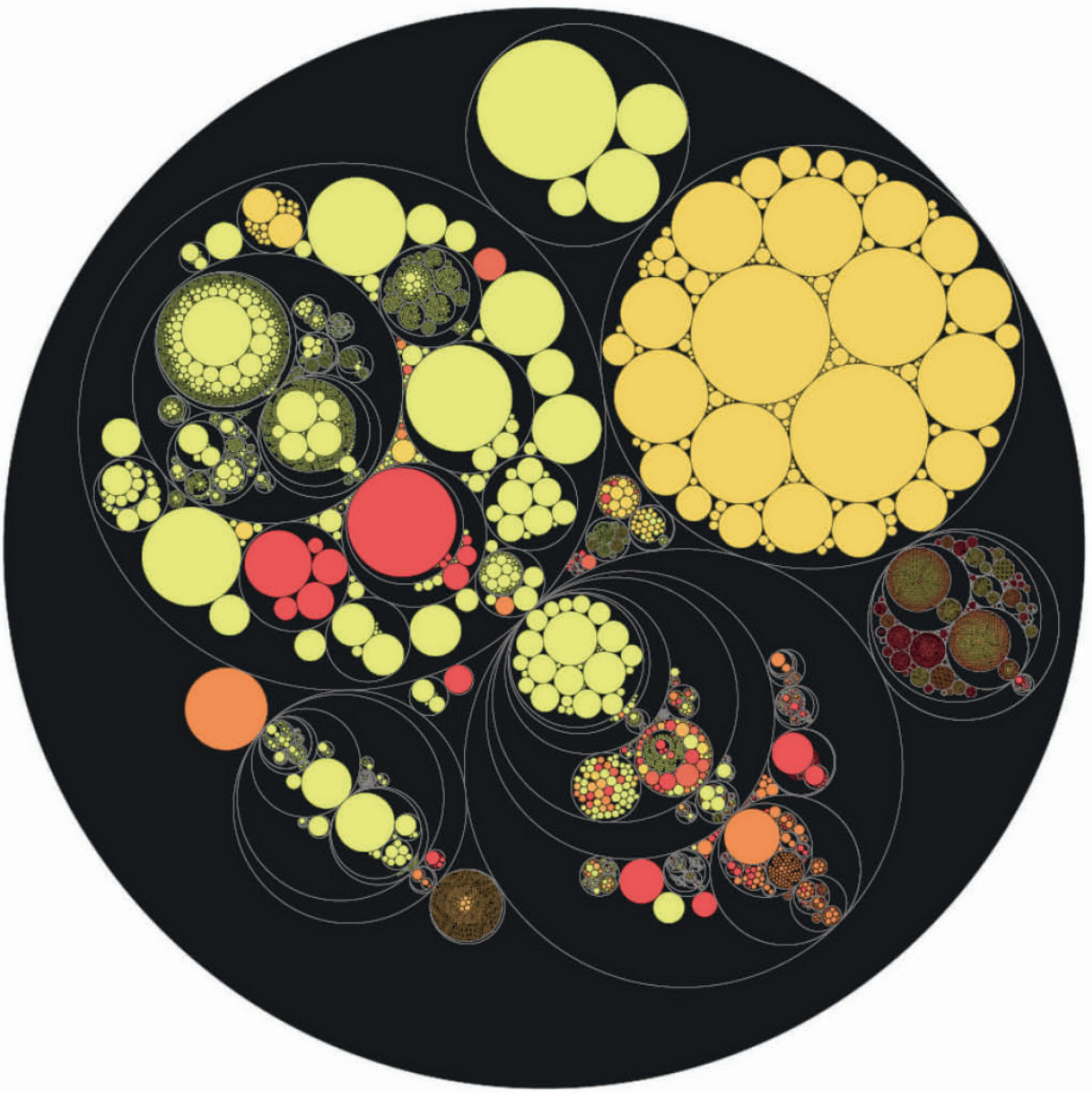
NEW TRENDS ON GENOME AND TRANSCRIPTOME CHARACTERIZATIONS

EDITED BY: Rosalba Giugno and Vincenzo Manca
PUBLISHED IN: Frontiers in Genetics



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-610-9
DOI 10.3389/978-2-88945-610-9

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

NEW TRENDS ON GENOME AND TRANSCRIPTOME CHARACTERIZATIONS

Topic Editors:
Rosalba Giugno, University of Verona, Italy
Vincenzo Manca, University of Verona, Italy

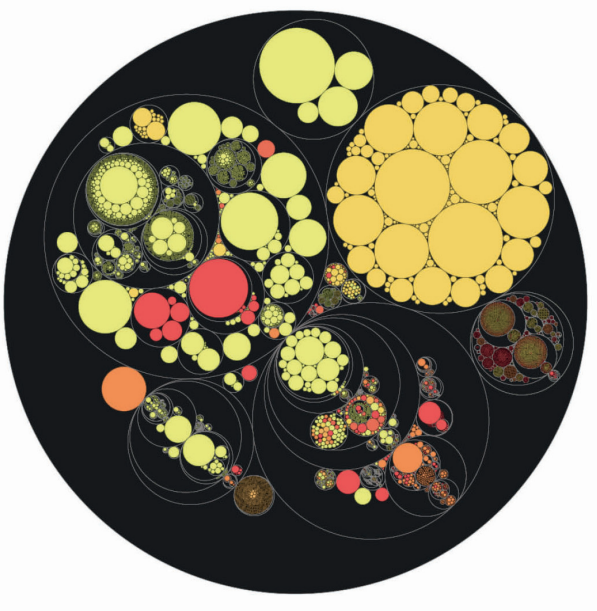


Image: Rosario Lombardo, COSBI UNITN-Microsoft, PhD Thesis Unconventional Computations and Genome Representations, 2013.

Citation: Giugno, R., Manca, V., eds (2018). New Trends on Genome and Transcriptome Characterizations. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-610-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 20 August 2018
doi: 10.3389/fgene.2018.00322

Editorial: New Trends on Genome and Transcriptome Characterizations

Rosalba Giugno and Vincenzo Manca*

Department of Computer Science, University of Verona, Verona, Italy

Keywords: algorithms, data analysis, computational biology, bioinformatics, analysis methods

Editorial on the Research Topic

New Trends on Genome and Transcriptome Characterizations

Worldwide, National Health Systems are investing to fit the requirements of "precision medicine." This term refers to the prevention and treatment of diseases that take into account the individual characteristics of patients, from their genetic variability to their different life style. However, the aims of precision medicine can only be fully realized when the internal mechanisms of diseases are understood and a deep knowledge of their individual variability is reached. Such a high level of comprehension can not only allow physicians to maximize the benefit against dangerous side effects, but can also promote the discovery of new treatments and prevention procedures. The same attitude, which in medicine can be synthetized as a shift from pathologies to patients, can be applied in the field of agriculture, when general principles are adapted and regulated according to the specific environments and local situations where cultures are realized.

These objectives require financial and intellectual resources to collect large amounts of genome and transcriptome data that are specifically related to the phenomena of interest in the different fields of applications (diseases in many contexts of related pathologies, or agricultural settings at the production level). However, a second methodological aspect is the powerful combination of information theoretical concepts with specific algorithmic and computational tools. This informational perspective is intended to extract deep biological meaning that often escape from simple statistical analyses of macroscopic phenomena. In fact, long range correlations or deep mathematical regularities, surprisingly enough, seem to relate with biological structures and functions that are encoded in a multilevel organization of genomes and of their expression. The search for new information-based categories will provide novel interpretation of classical biological concepts using this new informational approach.

The results of this methodological innovation are going to have a wide range of applications in the public health and in many economic sectors that contribute positively to human lifestyle and to progress of countries. In particular, in the agri-food sector, the study of the variability of genomes and transcriptomes implies the improvement of productivity and quality of products. In this field the combination of plant biotechnology with bioinformatics, in comparison with the traditional techniques of phenotypic analysis of plants, will provide a remarkable increase of speed and efficiency in the selection of progenies with superior characteristics.

This Research Topic collected contributions from computer scientists, bioinformaticians, and geneticists who designed or applied unconventional methods to understand pathology in medical and agricultural contexts. The issue comprises 12 articles, with 9 original research articles and 3 reviews.

Mosca et al. present a network smoothing method to predict gene modules involved in Alzheimer disease. The approach prioritizes the role of genes in the disease by the grade of their smoothing, which reflect the interaction topology of those genes with known genes in the AD. It is interesting to notice that, while the paper was under review, 3 of the several genes predicted in

OPEN ACCESS

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

****Correspondence:***
Rosalba Giugno
rosalba.giugno@univr.it

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

Received: *07 June 2018*
Accepted: *30 July 2018*
Published: *20 August 2018*

Citation:
Giugno R and Manca V (2018) Editorial: New Trends on Genome and Transcriptome Characterizations. Front. Genet. 9:322. doi: 10.3389/fgene.2018.00322

this work were, independently, added to the genes-association database (SFARI) of The Simons Foundation Autism Research Initiative.

Pedersen et al. present an integrative approach to combine information from tissue specific protein interaction networks with genome-wide gene association in Type 2 diabetes (T2D). They first constructed pancreatic and beta-cell protein complexes to be used as reference model for scaffold procedures, and then, by means of them, identified a set of 24 islet protein complexes probably disregulated or disfunctioning in T2D. Human specific tissues are investigated in Mayne et al. to assess the variegation of diseases depending on sex. By making use of meta-analysis approach, authors demonstrate that, among other results, even cortical regions may influence sexually dimorphic traits. Moreover, a large percentage of genes whose expression was sex-biased had androgen or estrogen hormone response elements.

Establishing the exceptional method for calculating deregulated genes has attracted many research efforts and we expect to see other attempts, in particular in the emerging scenario of single-cell RNA sequencing. Dal Molin et al. compare four different tools dedicated to differential expression of single cell RNA sequencing and extended two methods, commonly used for single cell data, for the analysis of bulk RNA sequencing data. The results on real and synthetic datasets (which imitate unimodal and bimodal distributions) reveal the limitations of each tool, by showing that no tool outperforms the others. Bulk RNA sequencing data related to ovarian follicle are deeply analyzed by Battaglia et al., a team devoted to this subject, aimed at characterizing non-coding RNAs to improve medical practice in infertility disorders, concerning with diagnosis, treatment, and discovery of biomarkers, for oocyte quality, in Assisted Reproductive Treatment. Computational methods in agriculture, combining bulk RNA sequencing and advanced bioinformatics methods, are presented by Wang et al. to detect new small RNAs and other interfering RNAs having a role in the tomato ethylene signaling pathway and fruit ripening.

Chen et al. investigate the genome duplication in rosids, whereas Cao et al. provide to assess the evolution of WUSCHEL-related homeobox transcription factors in rosaceae. Alagarasan et al. characterize the ZIP gene family in Setaria italic. Three reviews complete the issue. Chiara and Pavesi present the best practices for read pre-processing in the identification of human variome, i.e., casual mutations of haplotypes, in order to overcome the lost quality due to the high variability of data (gene panels, exomes, or whole genomes).

Vandin presents computational methods, mostly involving network analysis, for characterizing inter-tumor heterogeneity coming from pathways commonly mutated across different patients, and intra-tumor heterogeneity coming from bulk or single cell sequencing data. Finally, Alaimo et al. review computational techniques in agriculture area for grapevine meta-omics analysis. Authors discussed also the current knowledge of microbiome in plants, its differences varying according to the parts of plants, and its role to cause or protect from diseases.

AUTHOR CONTRIBUTIONS

RG and VM handled manuscripts and edited the Research Topic.

FUNDING

This work has been partially supported by the following projects: GNCS-INDAM, Fondo Sociale Europeo, and National Research Council Flagship Projects Interomics; JOINT PROJECTS 2016-JPVR16FNCL; JOINT PROJECTS 2017-B33C17000440003; project of the Italian Ministry of Education, Universities and Research (MIUR) "Dipartimenti di Eccellenza 2018–2022."

ACKNOWLEDGMENTS

We would like to thank the reviewers for their exceptional support in ensuring a rigorous selection of high quality contributions for this research topic and the members Frontiers editorial board for all their professionalism and assistance.

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 Genetics

ORIGINAL RESEARCH
published: 25 September 2017
doi: 10.3389/fgene.2017.00129

Network Diffusion-Based Prioritization of Autism Risk Genes Identifies Significantly Connected Gene Modules

Ettore Mosca[1], Matteo Bersanelli[2], Matteo Gnocchi[1], Marco Moscatelli[1], Gastone Castellani[2], Luciano Milanesi[1] and Alessandra Mezzelani[1]*

[1] Bioinformatics Group, Institute of Biomedical Technologies, National Research Council of Italy, Segrate, Italy, [2] Applied Physics Group, Department of Physics and Astronomy, University of Bologna, Bologna, Italy

Autism spectrum disorder (ASD) is marked by a strong genetic heterogeneity, which is underlined by the low overlap between ASD risk gene lists proposed in different studies. In this context, molecular networks can be used to analyze the results of several genome-wide studies in order to underline those network regions harboring genetic variations associated with ASD, the so-called "disease modules." In this work, we used a recent network diffusion-based approach to jointly analyze multiple ASD risk gene lists. We defined genome-scale prioritizations of human genes in relation to ASD genes from multiple studies, found significantly connected gene modules associated with ASD and predicted genes functionally related to ASD risk genes. Most of them play a role in synapsis and neuronal development and function; many are related to syndromes that can be in comorbidity with ASD and the remaining are involved in epigenetics, cell cycle, cell adhesion and cancer.

Keywords: autism spectrum disorder, biological networks, network diffusion, data integration, gene module

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Nicola Ancona,
Institute of Intelligent Systems for Automation, Consiglio Nazionale Delle Ricerche (CNR), Italy
Giovanni Ciriello,
University of Lausanne, Switzerland

****Correspondence:***
Ettore Mosca
ettore.mosca@itb.cnr.it

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

Received: *08 March 2017*
Accepted: *04 September 2017*
Published: *25 September 2017*

Citation:
Mosca E, Bersanelli M, Gnocchi M, Moscatelli M, Castellani G, Milanesi L and Mezzelani A (2017) Network Diffusion-Based Prioritization of Autism Risk Genes Identifies Significantly Connected Gene Modules. Front. Genet. 8:129. doi: 10.3389/fgene.2017.00129

INTRODUCTION

"Autism spectrum disorder" (ASD) includes clinically and etiologically wide range of neurodevelopmental disorders such as the less severe disorders Asperger's syndrome and pervasive developmental disorder, not otherwise specified, as well as the most severe childhood disintegrative disorder. ASD symptoms are recognized mainly by the complex behavioral phenotype that manifests within the first 3 years of life: difficult in communication and social interaction, limited interests and repetitive behaviors (National Institute of Mental Health, 2013).

Genetics play a crucial role in autism pathogenesis (Devlin and Scherer, 2012). Indeed, ASD has a high-heritability index (0.85–0.92) (Monaco and Bailey, 2001), a significant sib recurrence risks (8.6%) and 64% concordance among monozygotic twins (Smalley, 1998). Thousands of causative or predisposing genetic variations have been found in ∼30% of autistic patients (O'Roak et al., 2012), thus making autism a complex multifactorial disorder involving many genes and loci contributing to the phenotype. Genetic variations involved in ASD are chromosomal abnormalities (∼5%), copy number variations (CNVs) (10–20%) and single-gene mutations (∼5%) (Miles, 2011). Although the role of genetics in ASD etiology is recognized for ∼70% of cases, the causative factor is still unknown.

The approaches currently used to disentangle the genetic complexity of ASDs include large genome-wide association studies (GWAS), CNV testing and genome sequencing. Interestingly, the

application of these different approaches yielded many non-overlapping genes, which may suggest different molecular mechanisms within connected pathways (Pinto et al., 2014). The analysis of molecular interactions and pathways is therefore crucial for the interpretation of the results emerging from genome-scale studies on a pathology marked by a significant genetic heterogeneity. Indeed biological pathways associated with a specific pathology are likely to be more conserved than individual genetic variations, because multiple combinations of variations might perturb each pathway (Barabási et al., 2011). Network-based and pathway-based analyses can therefore provide a functional explanation to non-overlapping genes and narrow the targets for therapeutic intervention (Devlin and Scherer, 2012).

One of the problems that network-based analyses can solve is indeed the identification of the so-called disease modules, i.e., network regions associated with a disease (Barabási et al., 2011). Recently, molecular interaction networks have been used in the analysis of ASD genetic data to define gene networks associated with ASD. The identification of a subnetwork with desired properties from a large biological network (like the one formed by all protein-protein interactions) poses many challenges and therefore several approaches have been proposed (Mitra et al., 2013).

Regarding the integration of networks and ASD genetic data, Cristino et al. (2014) studied the interacting partners of genes known to be associated with ASDs and other related disorders; Noh et al. (2013) identified a significantly interconnected network of genes affected by CNVs; Li et al. (2014) studied the association between ASDs and genes forming topological communities (clusters of genes with a high density of connection between genes of the community and less connections with genes outside the community); Gilman et al. (2011) found functionally connected clusters of genes affected by CNVs.

Recently, network smoothing index (NSI) was proposed as a network-based quantity that allows to define a network region enriched with *a priori* information (Bersanelli et al., 2016). The NSI is based on network diffusion, a method that simulates the flow of a fluid throughout a network. NSI quantifies the network relevance of each gene in relation to a set of input genes (e.g., ASD genes), considering the whole network and mitigating the importance of hubs.

In this work, we use network diffusion and the NSI to propose a possible disease module for ASD, encompassing the network regions most frequently hit by molecular variations reported in several studies and collected in curated public databases. Moreover, our study introduces a network-based genome-wide prioritization of genes in relation to their known and predicted relevance for ASDs.

MATERIALS AND METHODS

Molecular Interactions

STRING interactions were collected from STRING (version 10), a database of direct and indirect PPIs (Szklarczyk et al., 2015). Native identifiers were mapped to Entrez Gene (Brown et al., 2015) identifiers. In case multiple proteins mapped to the same gene identifier, only the pair of gene identifiers with the highest STRING confidence score was considered. A total of 11,535 genes and 207,157 links with confidence score $\geq$700 was retained.

ASD Risk Genes

ASD risk genes were collected from The Simons Foundation Autism Research Initiative SFARI Gene database (Abrahams et al., 2013, version available in July 2015) and from Li et al. (2014).

SFARI Gene provides a publicly available database where genes are scored according to the strength of the evidence of gene's association with autism. In particular, genes are assigned to 7 categories (Supplementary Table 1): "syndromic" (S), "high confidence" (1), "strong candidate" (2), "suggestive evidence" (3), "Minimal evidence" (4), "Hypothesized" (5), and "Not supported" (6). SFARI genes were divided into to broad classes of high and low strength of association with ASD. Genes belonging to categories S, 1, 2, 3, 1S, 2S, 3S, 4S were included in SFARIh list, while genes of categories 4 and 5 were grouped into SFARIl. Native gene identifiers were converted to Entrez Gene (Brown et al., 2015) identifiers.

In addition to SFARIh genes, we considered 5 sources (namely dnCNVn, dnCNV3s, rCNV, dMUT, mMUT) of genes harboring CNVs and mutations associated with ASDs, proposed by large recent studies (Li et al., 2014). dnCNVn contains genes from an ASD-associated network composed of genes with *de novo* CNVs identified in 181 individuals and genes previously implicated in ASDs (Noh et al., 2013). dnCNV3s contains genes with *de novo* CNVs found in all three independent studies on more than 1,000 families (Levy et al., 2011; Sanders et al., 2011; Pinto et al., 2014). rCNV contains genes with rare CNVs found in a study involving approximately 1000 ASD individuals of European ancestry and matched controls (Pinto et al., 2010). dMUT and mMUT include genes with, respectively, disruptive and missense mutations (Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012; Li et al., 2014).

Only genes occurring in STRING network were considered in network-based analyses (**Table 1**). Therefore, whenever appropriate, we will specifically refer to the original gene lists and the corresponding derived lists with only genes occurring in STRING network using suffixes "i" and "n0" respectively (e.g., SFARIh_{i}, $\text{SFARIh}_{\text{n0}}$).

Network-Based Analysis

Given an input gene list L and a gene network encoded as the n-by-n symmetrically normalized adjacency matrix W (Bersanelli et al., 2016), the n-sized vector X_0 was defined in order to have positive quantities only in its elements representing the genes in L, and null values for all the other genes. Network diffusion finds the vector X_*, in which the quantities initially available in X_0 are subject to smoothing according to the pattern of interactions W. The vector X_* was calculated using an iterative procedure (Zhou et al., 2004), as described in Bersanelli et al. (2016):

$$X_{t+1} = \alpha W X_t + (1-\alpha) X_0,\ X_* = \lim_{t\to\infty} X_t,$$

TABLE 1 | Overlap between lists of genes harboring variations associated with ASDs.

	dnCNVn	dnCNV3s	rCNV	dMUT	mMUT	SFARIh
dnCNVn	154 203	39	9	2	6	11
dnCNV3s	50	263 530	11	3	12	12
rCNV	9	14	221 396	2	9	3
dMUT	2	4	14	45 67	1	16
mMUT	6	16	4	3	251 365	19
SFARIh	16	18	16	22	24	154 206

Gene list size (diagonal) and number of genes co-occurring between pairs of lists (off-diagonal elements); lower triangle and diagonal (bottom): original gene lists; upper triangle and diagonal (top): original gene lists with only genes occurring in STRING network.

where α [here set to 0.7 as in previous works (Bersanelli et al., 2016)] is a parameter that weights to which extent the initial information is retained or spread throughout the network. In the independent smoothing of each of the six ASD gene lists described above, genes belonging to the list were set to 1. In the joint analysis of all gene lists, genes belonging to SFARIh_{n0} were set to 1, while genes belonging to other lists were set to 0.5: this setting was chosen so that genes strongly associated with ASD had a higher priority.

For each gene g, the network smoothing index S quantifies the network proximity of g to genes marked by a positive value in X_0, i.e., associated with ASD, as ratio between gene values after and before network diffusion:

$$S(g) = \frac{X_*(g)}{X_0(g) + \varepsilon}$$

where ε is a small positive quantity that weights the importance of the initial values X_0. In order to mitigate the tendency of hub genes to gather excessive amounts of information only because of their central position, the permutation-adjusted network smoothing index S_p was introduced as,

$$S_p(g) = -log_{10}(p_S(g)) \cdot S(g)$$

where $p_S(g)$ is an empirical p-value, computed using K permutations of X_0, each one denoted as X_0^k, and the corresponding $S^k(g)$ (calculated using X_0^k):

$$p_S(g) = \frac{1 + \#\{S^k(g) \geq S(g)\}}{K + 1}.$$

In the analysis of the six ASD risk gene lists, ε values were defined in order to predict genes in network proximity to the input genes. Given a gene set of size N, ε was set in order to obtain, among the first N top ranking genes by S_p, a ratio of 1:1 between (i) the number of input genes and (ii) the number of genes in network proximity to input genes. The resulting values were 0.21 for dnCNVn_{n0} and 0.19 for dnCNV3s_{n0}, RcnV_{n0}, dMUT_{n0}, mMUT_{n0} and SFARIh_{n0}. In the joint analysis of all gene lists ε was set equal to 1, because the analysis was mainly aimed at defining a network-based prioritization of the 956 input genes, rather than at predicting other genes in network proximity. In all these analyses we used $K = 999$.

Network resampling (NR) shows to which extent a network score, resulting from the combination of gene scores, is expected if links among genes are shuffled. Also in this case permutations are used to define the null model. Given a number m of genes at the top of a ranked gene list, NR consists of two steps. First, a non-decreasing quadratic objective function $\Omega(m)$ is defined:

$$\Omega(m) = S_p^T(m) \cdot A_m \cdot S_p(m),$$

where $S_p(m)$ is the vector referring to the first m-scoring genes and A_m is the adjacency matrix between such genes. In the second step, q permutations of A_m are defined keeping the same degree distribution. Lastly, an empirical p-value (p_N) is calculated to quantify the fraction of times the objective function calculated on a permuted network, $\Omega^k(m)$, is greater than or equal to $\Omega(m)$. The procedure is repeated for different m, providing an overview on whether gene links and gene scores determine significant network scores when moving down in the ranked gene list (see figures below).

Network resampling (NR) was applied to genes ranked by S_p in descending order and using a total of 200 permutations, which was enough to underline the presence of significantly connected components (gene modules).

Pathway Analysis

Pathway analysis was carried out using over-representation analysis (ORA). ORA estimates the significance of a pathway in relation to an input gene list, calculating the hypergeometric probability of finding the observed number of input genes that are also members of the considered pathway, in the context of a background set of genes. As a background we considered all the genes occurring in original lists and all genes occurring in the gene network. Gene-pathway associations were downloaded from NCBI Biosystems (version: February 2017) (Geer et al., 2010); in particular, only pathways (gene sets) with a number of genes between 10 and 200 were considered. Hypergeometric probabilities were calculated using "phyper" and "dhyper" R functions, and were corrected for multiple hypotheses testing using the Benjamini-Hochberg method, implemented in "p.adjust" R routine. The similarity between two gene sets (A, B) was calculated using the overlap coefficient: $o = |A \cap B| / \min(|A|, |B|)$.

RESULTS

Network Location of Genes Associated with ASDs in SFARI Database

With the aim of characterizing the functional relations among SFARI genes and predict relevant risk genes for ASDs, we

considered direct and indirect protein-protein interactions (PPI) and quantified, *via* the permutation-adjusted NSI (*Sp*) (Bersanelli et al., 2016), the network proximity of each human gene in relation to the network location of 154 genes reported as strongly associated with ASD ($SFARIh_{n0}$ list, **Table 1**).

We found several genes in significant network proximity to $SFARIh_{n0}$ genes, with high *S* and low p_S (**Figure 1A** and Supplementary Table 2). Interestingly, among these genes, we found a significant number of genes having minimal/hypothetical evidences of association with autism in SFARI ($SFARIl_{n0}$) (**Table 2**).

In order to assess whether genes ranked by S_p formed a significantly connected gene module, we applied the NR approach (Bersanelli et al., 2016). We observed a significantly connected gene module (M_{SFARI}) resulting from the top 244 genes (**Figures 1B–D**). This module includes 142 (out of 154) $SFARIh_{n0}$ genes, 9 $SFARIl_{n0}$ genes and 93 genes not in SFARI.

These 93 genes include regulators of synaptic development and plasticity, are involved in syndromic conditions in comorbidity with ASD, regulate epigenetic mechanisms and a few are associated to cancer (Supplementary Table 2). For example, among the 93 genes that have a relevant position within the module (**Figure 1C**) we found cancer genes that control cell proliferation, [e.g., Tumor Protein P53 (TP53), AKT Serine/Threonine Kinase 1 (AKT1), Mechanistic Target Of Rapamycin (MTOR), C-Terminal Binding Protein 1 (CTBP1)], a process that was recently proposed as a common denominator of cancer and ASDs (Crawley et al., 2016), and genes with relevant role for brain function e.g., histone deacetylase-1 (HDAC1), histone deacetylase-3 HDAC3 (Volmar and Wahlestedt, 2015) and contactin-2 (CNTN2) (Anderson et al., 2012). These 93

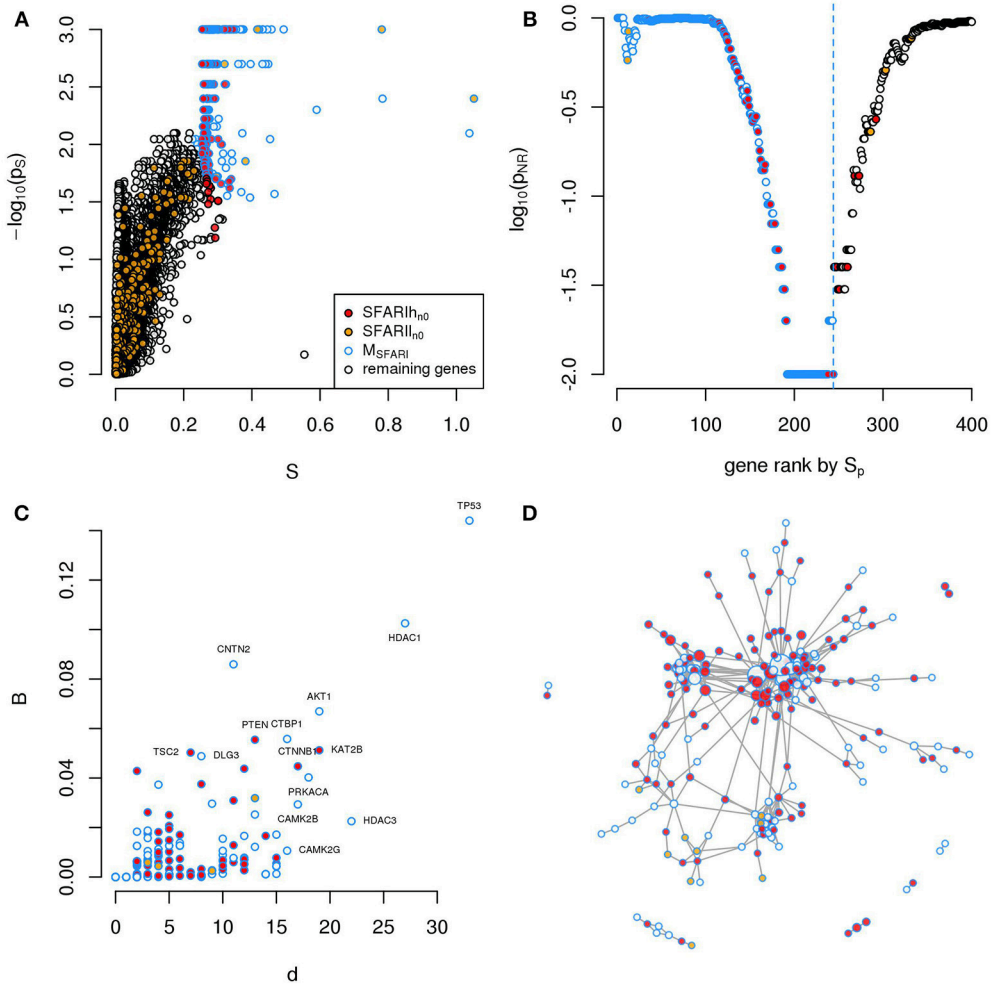


FIGURE 1 | A significantly connected gene module based on SFARIh genes. The network-diffusion based analysis of SFARIh genes (red) leads to the definition of a significantly connected gene module of 244 genes (blue border). **(A)** Network smoothing index and corresponding estimated *p*-value (p_S). **(B)** For each rank *n* (horizontal axis), the estimated *p*-value (p_{NR}) that quantifies the significance of the gene network defined by the genes ranked up to the *n*-th rank. **(C)** Number of interactions (*d*) and normalized betweenness (*B*) of genes in M_{SFARI}. **(D)** Visualization of M_{SFARI} gene module, in which genes are circles and PPI are links; circle size is proportional to gene degree. **(A–D)** *G*: all genes included in the STRING network.

predicted genes act as a bridge between SFARIh genes that were not directly linked in STRING (**Figure 1D**).

Interestingly, while this manuscript was under review, 3 out of the 93 predicted genes were added to the SFARI database independently from our study. Namely, HADAC3 was added among "hypothesized" genes, while TANC2 and PPP2R1B among "minimal evidence" genes.

Genes in Network Proximity to Genes Harboring Variations Associated with ASD

In addition to the SFARIh gene list, we considered five other lists of genes found altered in ASD subjects by previous studies. These lists vary in size from 67 to 530 and from 45 to 263 genes after integration with STRING gene network. The percent overlap between lists is low (**Table 1**) and most of the genes occur only in one list (**Figure 2**). In this situation, as introduced earlier, the information on how gene products interact to regulate biological functions can be used to explain the heterogeneity of ASD risk gene lists. Firstly, we analyzed each list separately to underline the specificities and commonalities of each list. Subsequently, we used network information to define a prioritization among all ASD risk genes proposed in the considered studies (union of all gene lists).

We calculated the *Sp* of all genes in STRING network considering as input each of the six ASD gene lists ($SFARIh_{n0}$, $dnCNVn_{n0}$, $dnCNV3s_{n0}$, $rCNV_{n0}$, $dMUT_{n0}$, and $mMUT_{n0}$) and selected the top $2n$ genes ranked by decreasing values of *Sp*, where n is list size (Supplementary Table 3). Note that almost all these genes are in significant network proximity ($p_S < 0.05$) to the corresponding input genes (**Figure 3** and Supplementary Table 3), which are also ranked among the top $2n$ genes. For convenience we will refer to these network-based gene lists—which contains input and predicted genes—as $SFARIh_{n^*}$, $dnCNVn_{n^*}$, $dnCNV3s_{n^*}$, $rCNV_{n^*}$, $dMUT_{n^*}$, and $mMUT_{n^*}$.

Only 14 genes occur in three or more input gene lists (**Figure 2**). Among those genes, at least DLGAP2 (Discs Large Homolog Associated Protein 2) and SYNGAP1 (Synaptic Ras GTPase Activating Protein 1) are worth mentioning. In fact, DLGAP2, a post-synaptic density protein with probable implication in ASD pathogenesis (Chien et al., 2013) scored as "minimal evidence" in SFARI (SFARIl), is part of all three CNV lists and was predicted as functionally related to SFARIh genes and dMUT genes. Furthermore, DLGAP1 (Discs Large Homolog Associated Protein-1) was predicted as functionally related to genes of 3 input lists, including SFARIh (**Figure 2B**). Similarly, SYNGAP1, which codes an autism related brain-specific synaptic Ras GTP-activating protein (Berryer et al., 2013), occur in three input lists (dnCNVhc, dnCNV3s and SFARIh) and was predicted as functionally related to genes harboring rCNVs.

Globally, a total of 913 genes were predicted as functionally related to at least one ASD risk gene list (**Table 3** and Supplementary Table 4). Interestingly, 106 of these genes were already proposed as ASD risk genes in 1 or more

TABLE 2 | Number of SFARIl genes in network proximity to SFARIh genes.

|M ∩ SFARIl|	|M|	|G-M|	|SFARIl|	E(|M ∩ SFARIl|)	*p*
2	10	11,371	216	0.190	$1.46{\cdot}10^{-2}$
9	100	11,281	216	1.90	$1.15{\cdot}10^{-4}$
9*	102*	11,279*	216*	1.94*	$1.34{\cdot}10^{-4}$*
20	300	11,081	216	5.69	$1.07{\cdot}10^{-6}$
22	400	10,981	216	7.59	$7.27{\cdot}10^{-6}$
26	500	10,881	216	9.49	$2.87{\cdot}10^{-6}$

*M, genes in network proximity to SFARIh genes; G, all genes; |M ∩ SFARIl| size of the intersection between M and SFARIl; $E(|M \cap SFARIl|) = \frac{|M|}{|G|} \cdot |SFARIl|$, expected |M ∩ SFARIl|; p: probability that |M ∩ SFARIl| is greater than or equal to the observed value in a hypergeometric experiment; the asterisks underline the overlap obtained using the top ranking 244 genes ranked by S_p (102 after removing SFARIh genes) composing a significantly connected gene module (M_{SFARIl}) (**Figure 1**).*

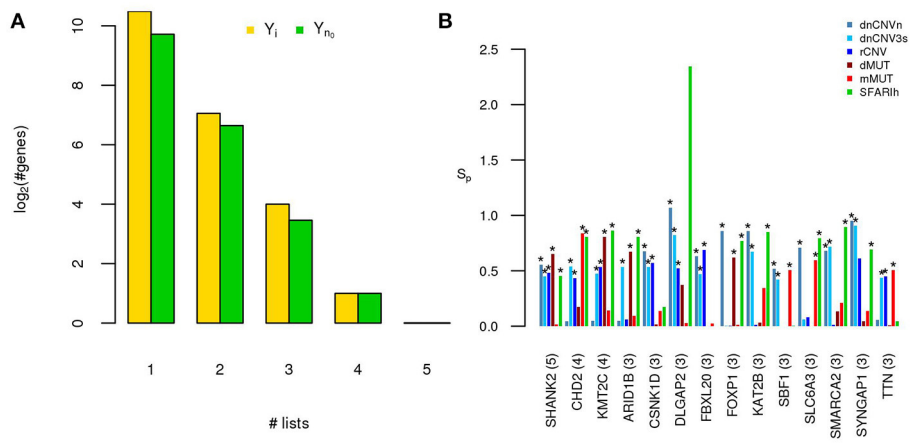


FIGURE 2 | A small number of ASD risk genes is found in more than one large study. **(A)** Number of risk genes found in 1 or more of the six gene lists on ASD; *i*: original gene lists; *n0*: original gene lists wit only genes occurring in the STRING network. **(B)** Permutation-adjusted network smoothing index of the 14 genes occurring in 3 or more original lists (the number is reported between parenthesis); the asterisk (*) indicates genes of the corresponding original list.

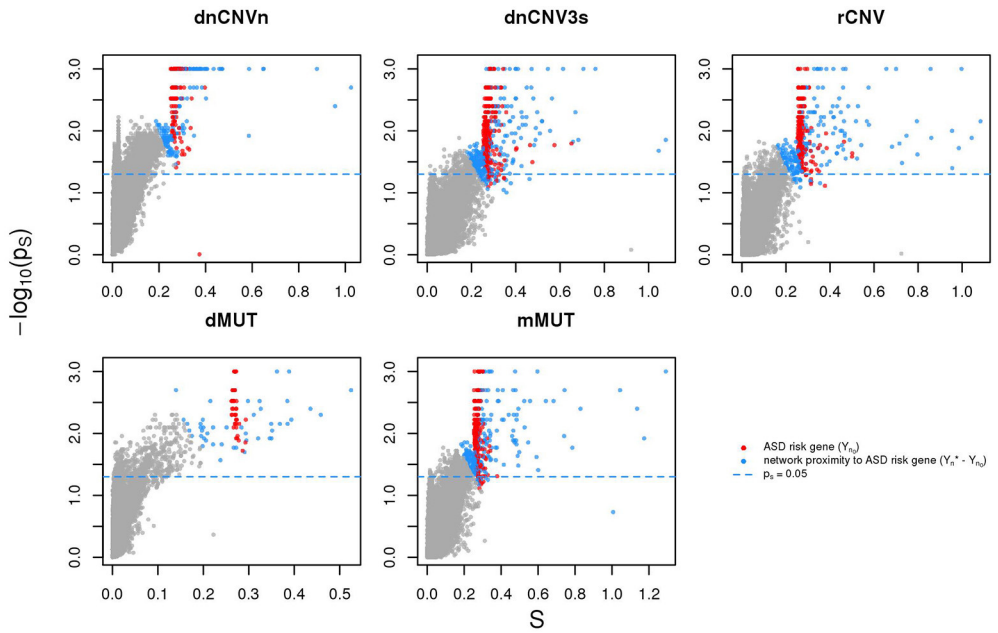


FIGURE 3 | Genes in network proximity to ASD risk genes. The network smoothing index (*S*) and the corresponding *p* value (p_S) are the two factors that make up *Sp*. Blue: genes belonging to the corresponding original list (sources); red: genes in network proximity to sources; gray: remaining genes.

TABLE 3 | Co-occurrence of genes in network proximity to ASD risk genes from one or more sources.

		Input (n_I)				
		0	**1**	**2**	**3**	**#**
Predicted (n_P)	**1**	691	68	14	2	775
	2	89	15	4	1	109
	3	18	2	0	0	20
	4	6	0	0	0	6
	5	3	0	0	0	3
	#	807	85	18	3	913

A total of 913 genes were predicted to be in network proximity to ASD risk genes of one or more studies (rows) where could appear as ASD risk genes (columns). For example, 15 genes were predicted in network proximity to ASD risk genes of 2 studies and were proposed as ASD risk genes in 1 study ($n_I = 1$, $n_P = 2$); #: row or column sum.

studies: for example, CTNNB1 (Catenin-Beta1) encoding a protein part of the adherens junctions complex, NRXN1 (Neurexin1), NLGN4X (Neuroligin4X), encoding a pre-synaptic and post-synaptic protein, respectively, and the tumor suppressor PTEN (Phosphatase And Tensin Homolog), were predicted as functionally related to 2 gene lists and included as risk genes in other 2 lists.

We have also found genes that were not included in any input gene list, but were predicted to be in relevant network proximity to multiple gene lists. For example, 27 were predicted as functionally related to three gene lists (**Figure 4**); among these, ADGRL2 (adhesion G protein-coupled receptor L2), LRTOM (leucine rich transmembrane and O-methyltransferase domain containing) and SRC (Proto-Oncogene Non-Receptor Tyrosine Kinase SRC) were predicted in functional relation to 5 lists.

Many of the 29 genes predicted as functionally related to three or more lists—27 genes not included in any input list and 2 included in one study—take part in many PPIs, implicating they are central in the PPI network (e.g., TP53, AKT1). The significance of their *Sp* suggests that these genes were not only selected in relation to their centrality, but also because their network distance to ASD risk genes is lower than expected by chance (**Figure 3** and Supplementary Table 3). A further observation that supports this hypothesis is that these genes establish a number of interactions with ASD risk genes that is higher than expected ($p < 0.05$, hypergeometric test) (**Table 4**). From a network point of view, these 29 genes are "surrounded" by 369 ASD risk genes (first order neighbors).

It is also worth mentioning that several genes resulting with the highest network proximity score to each list tend to be list specific (**Figure 5**).

A total of 956 unique ASD risk genes occur in the 6 input lists. We calculated the *Sp* of all genes relative to these 956 genes (**Figure 6A**) and, by means of NR, found a significantly connected component of 561 genes (**Figure 6B** and Supplementary Table 5). This gene module (M_{ASD}) includes all SFARIh genes, 70% of genes occurring in dnCNV_n list, approximately 50% of the other gene lists (**Figure 6C**), 26 genes in SFARIl and 8 genes that do not belong to the input list. Among these 8 genes, we find the already mentioned AKT1, TP53, and SRC, which occupy a central role in the PPI network and were also predicted during the independent analysis of each input list (**Figures 4**, **6D**).

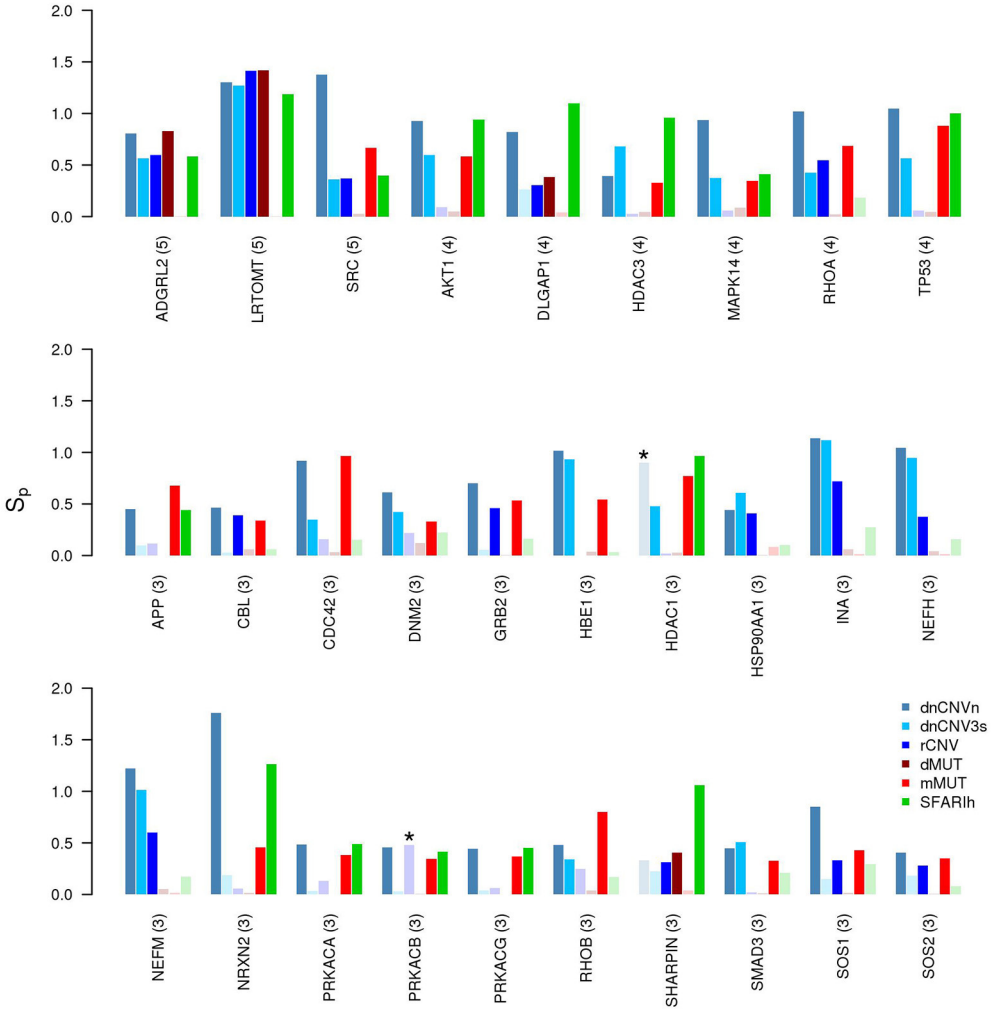


FIGURE 4 | Genes in network proximity to ASD risk genes from 3 or more studies. Permutation-adjusted network smoothing index (*Sp*) of the 29 genes predicted as functionally related to 3 or more ASD risk gene lists (the number is reported between parenthesis); the asterisk (*) indicates genes of the corresponding original list; faded colors indicate input ASD risk genes or genes with no significant *Sp* values.

Pathway Analysis of Gene Lists Associated with ASDs

We carried out an over-representation analysis to characterize original gene lists and network-based gene lists in terms of pathways. We observed significant pathways (at adjusted $p < 0.01$) in only two of the six original gene lists ($SFARIh_i$ and $dnCNVn_i$) and in five network-based lists ($SFARIh_{n^*}$, $mMUT_{n^*}$, $rCNV_{n^*}$, $dnCNV3s_{n^*}$, and $dnCNVn_{n^*}$) (Supplementary Table 6). Overall, we obtained a much higher number of pathways in network-based lists than in original ones, despite the number of tested genes was similar between the former and the latter ones. Therefore, the observed enrichment in pathways can be mainly brought back to the network-based analysis, since, by definition, it prioritizes genes functionally related to those considered in input. Further, apart from a single exception in SFARIh, pathways found only in original lists were similar, in terms of gene content, to pathways found also in network-based lists (**Figures 7A–F**). In other words, network-based analysis resulted in an enrichment at pathway level with a very limited loss of information. Interestingly, pathways found in original gene lists, and lost due to genes for which network-based analysis was not applicable, were recovered by network-based analysis (compare rows labeled with suffix "i", "n0", and "n*" in **Figure 7F**).

We summarized the significant pathways found in each lists in a unique pathway network, the so-called enrichment map (Merico et al., 2010). Specifically, we took into account up to 20 of the most significant pathways as representatives of each pathway cluster found in each list. This selection resulted in a total of 366 pathways, clustered in 11 groups: transmission across synapses (clusters 1 and 4), signal transduction (2), Rho GTPase and apoptosis (3), inositol phosphate metabolism (5), ER-associated degradation process (6), ion transport (7), chromatin remodeling (8), oxygen transport (9), proteoglycan biosynthesis (10) and Wnt signaling (11). While the majority

TABLE 4 | Hub genes predicted in network proximity to ASD risk genes of three or more studies establish a significant number of interactions with ASD risk genes.

Gene and function	Symbol	Band	# Studies	n_p	$\|I\|$	$\|A \cap I\|$	p
SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase Nonreceptor tyrosine kinase, frequently implicated in cancer	SRC	20q11.23	0	5	532	82	$2.00 \cdot 10^{-8}$
Tumor Protein P53 Involved in cell cycle regulation where negatively regulate cell division. Mutations in this gene are associated with a variety of human cancers	TP53	17p13.1	25	4	719	92	$1.36 \cdot 10^{-5}$
AKT Serine/Threonine Kinase 1 Implicated in the regulation of cell growth, proliferation, survival and differentiation (OMIM 164730)	AKT1	14q32.33	25	4	589	73	$2.90 \cdot 10^{-4}$
Ras Homolog Family Member A Regulates remodeling of the actin cytoskeleton during cell morphogenesis and motility. Overexpression of this gene is associated with tumor cell proliferation and metastasis	RHOA	3p21.31	13	4	406	65	$1.53 \cdot 10^{-7}$
Mitogen-Activated Protein Kinase 14 is a member of the MAP kinase family that are involved in cellular processes such as proliferation, differentiation, transcription regulation and development	MAPK14	6p21.31	11	4	333	44	$1.30 \cdot 10^{-3}$
Histone Deacetylase 3 belongs to the histone deacetylase family and represses transcription when tethered to a promoter; down-regulates p53 function and thus modulate cell growth and apoptosis	HDAC3	5q31.3	8	4	275	43	$3.54 \cdot 10^{-5}$
Heat Shock Protein 90 Alpha Family Class A Member 1 is a molecular chaperone involved in signal transduction, protein folding, protein degradation, and morphologic evolution.	HSP90AA1	14q32.31	10	3	589	65	$9.99 \cdot 10^{-3}$
Amyloid Beta Precursor Protein Is involved in promoting transcriptional activation; can participate in the formation of amyloid plaques of Alzheimer disease.	APP	21q21.3	15	3	363	42	$1.68 \cdot 10^{-2}$
Histone Deacetylase 1 Is a histone deacetylase and represses transcription; interacts with retinoblastoma tumor-suppressor protein to control cell proliferation and differentiation; modulates p53 effect on cell growth and apoptosis.	HDAC1	1p35.2	12	3	351	48	$3.69 \cdot 10^{-4}$
Cell Division Cycle 42 Acts in cell morphology, migration, endocytosis and cell cycle progression.	CDC42	1p36.12	9	3	336	56	$2.97 \cdot 10^{-7}$
Growth Factor Receptor Bound Protein 2 is involved in the signal transduction pathway.	GRB2	17q25.1	21	3	279	42	$1.06 \cdot 10^{-4}$
Protein Kinase CAMP-Activated Catalytic Subunit Alpha phosphorylates proteins and substrates, changing their activity; contributes to the control glucose metabolism, cell division, and contextual memory; developmental changes in synapse morphology.	PRKACA	19p13.12	14	3	278	41	$2.00 \cdot 10^{-4}$
SOS Ras/Rac Guanine Nucleotide Exchange Factor 1 participates in signal transduction pathways.	SOS1	2p22.1	14	3	260	50	$1.26 \cdot 10^{-8}$
SMAD Family Member 3 signal transducer and transcriptional modulator that mediates multiple signaling pathways probably involved in carcinogenesis	SMAD3	15q22.33	1	3	241	33	$2.83 \cdot 10^{-3}$
Protein Kinase CAMP-Activated Catalytic Subunit Beta is a member of the serine/threonine protein kinase family involved in cell proliferaton and differentiation	PRKACB	1p31.1	22	3	213	28	$9.81 \cdot 10^{-3}$
Cbl Proto-Oncogene targets substrates for degradation by the proteasome; is mutated or translocated in many cancers	CBL	11q23.3	12	3	203	37	$3.77 \cdot 10^{-6}$
Protein Kinase CAMP-Activated Catalytic Subunit Gamma is involved in the regulation of lipid and glucose metabolism and in the memory formation signaling cascade	PRKACG	9q21.11	7	3	185	27	$2.70 \cdot 10^{-3}$
Ras Homolog Family Member B Involved in intracellular protein trafficking of a number of proteins; plays a negative role in tumorigenesis	RHOB	2p24.1	8	3	174	38	$2.19 \cdot 10^{-8}$
SOS Ras/Rho Guanine Nucleotide Exchange Factor 2 is involved in the positive regulation of ras proteins	SOS2	14q21.3	12	3	109	24	$7.36 \cdot 10^{-6}$
Dynamin 2 produces microtubule bundles and binds and hydrolyzes GTP; regulates neuron morphology, axon growth; vesicular trafficking processes and cytokinesis	DNM2	19p13.2	19	3	84	21	$3.27 \cdot 10^{-6}$

Gene and function from GeneCards (Safran et al., 2003) and OMIM (Amberger et al., 2015); Band: cytogenetic band associated with genetic variations in SFARI (Abrahams et al., 2013); # Studies: number of studies supporting the genetic variations observed in the cytogenetic bands; I: interactors; A: ASD risk genes, $|A| = 956$; n_p: number of ASD risk gene sets the gene is network proximity to; I: interactors; p: probability that $|A \cap I|$ is greater than or equal to the observed value in a hypergeometric experiment; the total number of genes composing the network is 11,535.

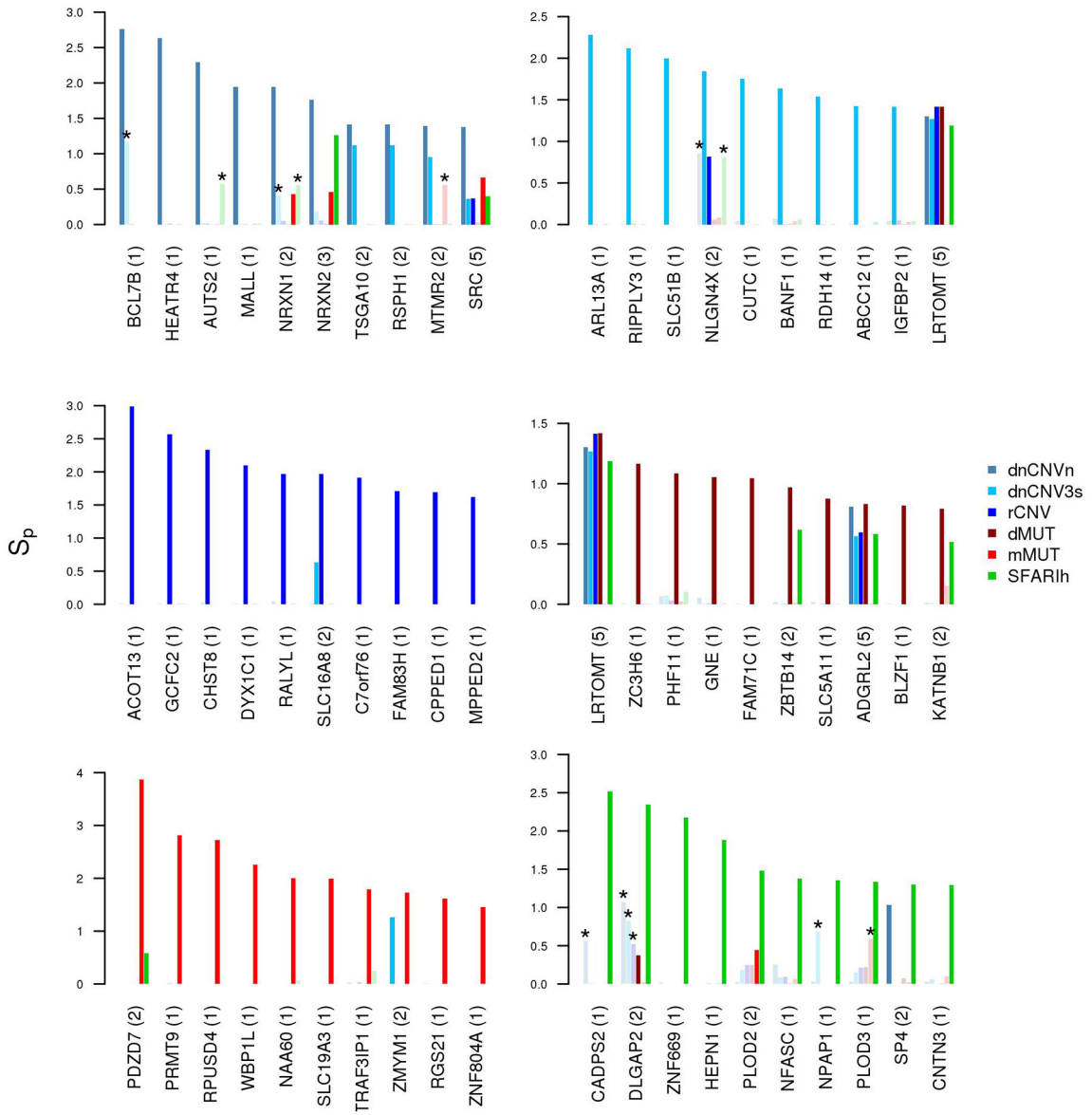


FIGURE 5 | Genes in network proximity to ASD risk genes from each study. Top 10 genes with the highest permutation-adjusted network smoothing index (*Sp*) calculated in the analysis of each ASD risk gene list; the number of lists to which the gene was predicted as network proximal is reported between parenthesis; the asterisk (*) indicates genes of the corresponding original list; faded colors indicate input ASD risk genes or genes with no significant *Sp* values.

of pathways were found in more than 1 list (**Figure 7G**, yellow circles), some pathway clusters were composed of pathways uniquely associated with one list. For instance, proteoglycan biosynthesis was specifically associated with rCNV, ion transport with mMUT, oxygen transport with dnCNV3s and dnCNVn.

DISCUSSION

Recently, the knowledge of molecular interactions has been used for the interpretation of genetic data on ASDs. In comparison to previous works, we analyzed multiple ASD risk gene lists proposed in large studies, for a total of approximately 1,000 genes. We observed a low overlap between ASD risk gene lists. Whether this heterogeneity reflects the biology of ASD or is the result of confounding factors, the analysis of network proximity between genes underlines the ASD risk genes that are also in functional relation and lead to the identification of modules of functionally related genes hit by genetic variations. The main limitation of a network-based analysis such as ours is the availability of *a priori* annotations required for the definition of the genome-scale network. In this work we considered both direct (physical) and indirect (functional) high confidence PPI from STRING, which allowed us to analyze 11,535 human genes. Note that the use of direct and indirect STRING interactions

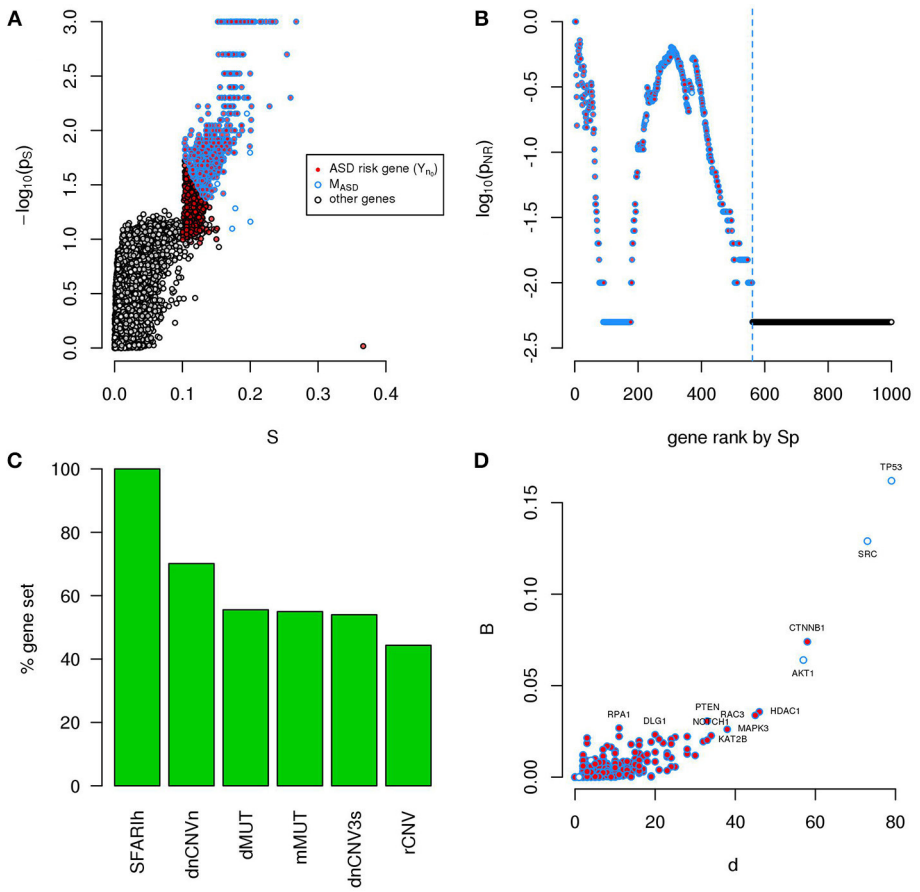


FIGURE 6 | A significantly connected gene module based on ASD risk genes from 6 sources. The network-diffusion based analysis of 956 ASD risk genes from 6 sources (red) leads to the definition of a significantly connected gene module (M_{ASD}) of 561 genes (blue border). **(A)** Network smoothing index (*S*) and corresponding *p*-value (p_S). **(B)** For each rank *n* (horizontal axis), the estimated *p*-value (p_{NR}) that quantifies the significance of the gene network defined by the genes ranked up to the *n*-th rank is reported on the vertical axis. **(C)** Percent of ASD genes of original lists included in M_{ASD}. **(D)** Number of interactions (*d*) and betweenness (B) of genes in M_{ASD}. **(A–D)** *G*: all genes included in the PPI network.

showed good performances in prioritizing candidate disease genes (Köhler et al., 2008). Moreover, unlike other network-based works on ASD genetic data, we used network diffusion to quantify network proximity between ASD risk genes and other genes. Network diffusion (a global approach) considers the whole network topology in its full complexity and, therefore, has better performances than local approaches (e.g., direct neighborhood or shortest path length; Wang et al., 2011). Lastly, we underlined ASD risk gene modules without constraining the search to topological communities. In fact, there is no guarantee that topological communities are able to capture disease modules (Ghiassian et al., 2015). Hence, we quantified the significance of the observed network proximity scores in comparison to random networks of the same degree distribution (Bersanelli et al., 2016).

Our work provides a network-based prioritization of human genes associated with ASD by previous studies. We extracted a module of 244 genes in network proximity to genes reported in SFARI as strongly associated with ASD. Interestingly, the module contains a significant number of genes proposed as possibly involved in ASD (categorized as "minimal evidence" and "hypothesized" in SFARI) and another 93 genes not scored in SFARI (Supplementary Table 2). While this manuscript was under review, 3 of these 93 genes were included in SFARI independently from our study.

From the 93 genes, 16 genes are involved in synaptogenesis and synaptic plasticity or transmission, and alterations in structure and function of neuronal synapses are well known causes of ASD. Among these, APLP2 also regulates proper progression of neuronal differentiation program during cortical development (Shariati et al., 2013), is involved in Alzheimer disease and interacts with CNTN in neurodevelopment and diseases (Osterfield et al., 2008). Then again, the 3 genes, CACNA2D1, CACNB1, CACNG1 induce the repression of the downstream regulatory element antagonist modulator (DREAM) and the expression of the neuropeptide dynorphin (DYN). DREAM plays a role in synaptic plasticity and behavioral memory (Wu et al., 2010), while DYN is involved in behavioral symptoms characteristic of human depressive disorders (Knoll and Carlezon, 2010). Also CACNG3, a calcium channel protein, regulates the function of AMPA-selective glutamate

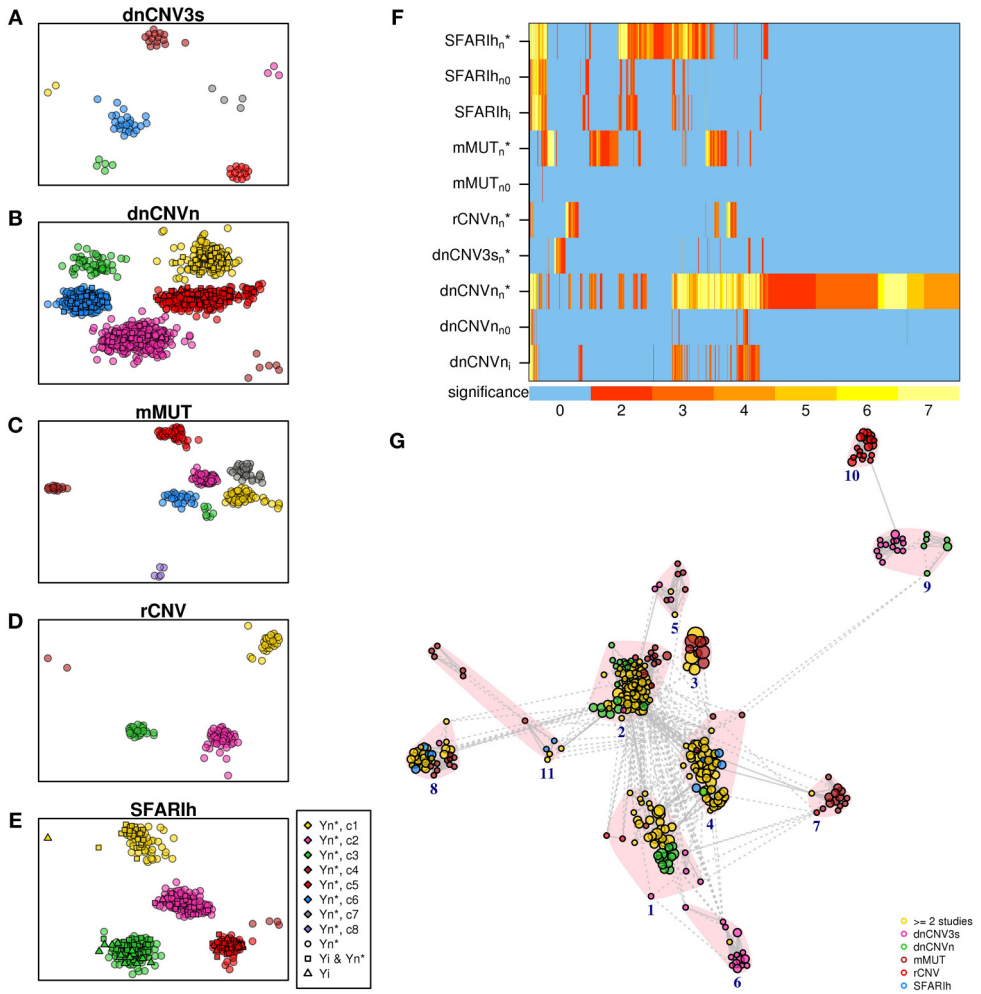


FIGURE 7 | Significant pathways enriched in genes associated with ASD. **(A–E)** Pathways found in the analysis of each original list (*i*) and corresponding predicted genes (n*). **(F)** Heatmap of all the pathways found in the analysis of the lists reported on the rows; also pathways found analyzing the original lists without genes not occurring in the STRING network (n_0) are represented; the color bar indicates the $-\log_{10}(p)$ of the hypergeomtric *p*-value, adjusted for multiple hypothesis testing; the number 7 indicates $p \leq 10^{-7}$. **(G)** Enrichment map; vertex size is proportional to pathway significance (adjusted *p*-value); links are reported only for overlap coefficients >0.5; for each pathway, only the links with the top 5 most similar pathways are drawn. **(A–G)** See Supplementary Table 6.

receptors and mediates synaptic transmission in CNS while FLRT3 takes part in a trans-synaptic complex (Lu et al., 2015). Eight genes are involved in neuronal differentiation, neurodevelopment and neuronal function. More specifically, AKT1 is a downstream mediator of the PI3K pathway that regulates synaptic formation and plasticity and which imbalance leads to autism and schizophrenia (Enriquez-Barreto and Morales, 2016); genetic variations in contactins (CNTN) have been described in association with neurodevelopmental disorders, including autism. Specifically, CNTN1 and CNTN2 are members of the presynaptic *NRXN* superfamily and 13 rare non-synonymous variants of CNTN2 have been found in ASDs patients while mice with *Cntn5* mutations show an abnormal audiogenic response due to defects in the formation of synapse in auditory neurons (Cottrell et al., 2011; Chen et al., 2014).

Many of the 93 genes are involved in syndromic comorbidities, including auditory and visual senses deficit, epilepsy, mental retardation and psychiatric conditions that affect nearly three-quarters of children with ASD. For instance, CAMK2G and PDZD7 are involved in Usher Syndrome the most common condition leading to deafness and blindness, as well as DNMT1 has a role in DNMT1-Related Dementia, Deafness, and Sensory Neuropathy (Vernon and Rhodes, 2009) and LRTOMT in deafness. Again, MYL7 mutations are associated with Fechtner Syndrome which features include hearing loss and eye abnormalities. SP4 is involved in bipolar disorder and schizophrenia while ANK3, ACSL4, DLG3 are associated with mental retardation and, interestingly, recalling the high male prevalence of ASD, the latter two map on X-cromosome; NHS also maps on X-cromosome and mutations in this gene cause Nance-Horan Syndrome characterized by congenital

cataract leading to vision loss; in males mild or moderate mental retardation may also occur and ASD have also been described in few patients (Toutain et al., 1997). Mutations in NAGLU and HGSNAT cause the Sanfilippo Syndrome (also called mucopolysaccharidosis Type III) often misdiagnosed with idiopathic developmental delay, attention deficit/hyperactivity disorder and/or ASD (Wijburg et al., 2013). QDPR mutations provoke hyperphenylalaninemia (Trujillano et al., 2014), (also called atypical phenylchetonuria (PKU), a genetic metabolic disease provoking postnatal cognitive deficit due to the neurotoxic effect of hyperphenylalaninemia; interestingly, PKU could be a comorbid condition of ASD, although with low prevalence (Baieli et al., 2003). MKRN3 is associated with Prader Willy Syndrome, NPAP1 both with Prader Willy Syndrome and Angelman Syndrome while DSCAML1 with Down Syndrome. These syndromes are characterized by mental retardation and can have co-occurring ASDs (Peters et al., 2004; Capone et al., 2005; Dykens et al., 2011). KMT2D and WDR5 defects are involved in Kabuki Syndrome characterized by multiple congenital abnormalities, from mild to severe developmental delay and intellectual disability. People suffering from this syndrome may also manifest seizures, hypotonia, strabism, hear infections, hearing loss and autism (Parisi et al., 2015). The very rare mutations in MANBA results in β-mannosidosis with a severe neurological disorder that can include mental retardation, cerebellar ataxia along with visual and hearing deficits (Sabourdy et al., 2009). CACNG3 is involved in Childhood Absence Epilepsy and is also associated with some cases of ASD (Danielsson et al., 2005) while mutations of KATNB1 cause complex cerebral malformations (Mishra-Gorur et al., 2014).

The remaining genes (among the 93) are mostly involved in epigenetics, cell cycle and cell adhesion and some of them are also implicated in tumor development as already reported by Crespi (2011) and Crawley et al. (2016).

The network-based analysis of genes from SFARI and other 5 previous studies resulted in the definition of a gene module that involves 561 ASD risk genes in significant functional relation. The module contains all the considered SFARI genes (strongly associated with ASD) and from 40% to 70% of genes from each of the other lists of ASD risk genes. Therefore, this module can be seen as a further screening of the genes proposed by such studies, which underlined those in significant functional relation from a network perspective.

More generally, the network-based scores that we calculated for every gene in the considered STRING network can be used to quantify the functional relation between any gene and ASD risk genes found in one or more previous studies.

Biological pathways enriched in genes in network proximity to ASD risk genes encompass several functions already proposed to be associated with ASD (see, for example, Pinto et al., 2010). Network-based analysis, through the prioritization of functionally related genes, enriched the number of significant pathways found by ORA in comparison to the analysis of original gene lists. Despite not all genes occurring in original lists underwent network-based analysis, the latter was not affected by a loss of information at pathway level.

The predicted genes in network proximity to ASD risk genes that have a central role in the PPI networks, but SRC, mapped in ASD risk loci. SFARI Gene database lists all the studies reporting CNV at the chromosome bands where predicted genes are localized (**Table 4**). In many reports, the CNV of interest was subsequently confirmed or validated by an independent method following its discovery. Additionally, from a functional point of view, most of the predicted genes are involved in epigenetics, cell cycle, growth-, proliferation- and differentiation-signaling and are often implicated in cancer development. This finding indicates pleotropic effects of some autism-associated genes on cancer risk and is supported by previous discussions that highlight a wide overlap in risk genes and pathways for cancer and autism (Crespi, 2011; Crawley et al., 2016). Advances in pharmacological therapies to ameliorate autism symptoms could be resulted from cancer drugs that target the same growth-signaling pathways (Crespi, 2011).

AUTHOR CONTRIBUTIONS

EM collected the ASD data from the literature, setup and run the analyses; MB curated the physical mathematical modeling of network diffusion; MG and MM managed the high performance computing infrastructure; GC supported the physical mathematical modeling; LM coordinated the research; AM analyzed the ASD literature, interpreted the biological results, coordinated the research. All authors discussed the results, contributed to manuscript writing and revision.

FUNDING

The work has been supported by: the EU FP7 project "MIMOmics" (305280); the Italian Ministry of Education, University and Research (MIUR) projects "INTEROMICS" (PB05) and "PRIN 2015"; the Lombardy Region Foundation FRRB project "LYRA" (2015-0010).

ACKNOWLEDGMENTS

We would like to thank John Hatton (CNR-ITB) for proofreading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00129/full#supplementary-material

REFERENCES

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., et al. (2013). SFARI gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* 4:36. doi: 10.1186/2040-2392-4-36

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). Omim.org: online mendelian inheritance in man (omim), an online

catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Anderson, G. R., Galfin, T., Xu, W., Aoto, J., Malenka, R. C., and Sdhof, T. C. (2012). Candidate autism gene screen identifies critical role for cell-adhesion molecule caspr2 in dendritic arborization and spine development. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18120–18125. doi: 10.1073/pnas.1216398109

Baieli, S., Pavone, L., Meli, C., Fiumara, A., and Coleman, M. (2003). Autism and phenylketonuria. *J. Autism Dev. Disord.* 33, 201–204 doi: 10.1023/A:1022999712639

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Berryer, M. H., Hamdan, F. F., Klitten, L. L., Mller, R. S., Carmant, L., Schwartzentruber, J., et al. (2013). Mutations in syngap1 cause intellectual disability, autism, and a specific form of epilepsy by inducing haploinsufficiency. *Hum. Mutat.* 34, 385–394. doi: 10.1002/humu.22248

Bersanelli, M., Mosca, E., Remondini, D., Castellani, G., and Milanesi, L. (2016). Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Sci. Rep.* 6:34841. doi: 10.1038/srep34841

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43, D36–D42. doi: 10.1093/nar/gku1055

Capone, G. T., Grados, M. A., Kaufmann, W. E., Bernad-Ripoll, S., and Jewell, A. (2005). Down syndrome and comorbid autism-spectrum disorder: characterization using the aberrant behavior checklist. *Am. J. Med. Genet. A* 134, 373–380. doi: 10.1002/ajmg.a.30622

Chen, J., Yu, S., Fu, Y., and Li, X. (2014). Synaptic proteins and receptors defects in autism spectrum disorders. *Front. Cell. Neurosci.* 8:276. doi: 10.3389/fncel.2014.00276

Chien, W. H., Gau, S. S. F., Liao, H. M., Chiu, Y. N., Wu, Y. Y., Huang, Y. S., et al. (2013). Deep exon resequencing of dlgap2 as a candidate gene of autism spectrum disorders. *Mol. Autism* 4:26. doi: 10.1186/2040-2392-4-26

Cottrell, C. E., Bir, N., Varga, E., Alvarez, C. E., Bouyain, S., Zernzach, R., et al. (2011). Contactin 4 as an autism susceptibility locus. *Autism Res.* 4, 189–199. doi: 10.1002/aur.184

Crawley, J. N., Heyer, W.-D., and LaSalle, J. M. (2016). Autism and cancer share risk genes, pathways, and drug targets. *Trends Genet.* 32, 139–146. doi: 10.1016/j.tig.2016.01.001

Crespi, B. (2011). Autism and cancer risk. *Autism Res.* 4, 302–310. doi: 10.1002/aur.208

Cristino, A., Williams, S., Hawi, Z., An, J., Bellgrove, M., Schwartz, C., et al. (2014). Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. *Mol. Psychiatry* 19, 294–301 doi: 10.1038/mp.2013.16

Danielsson, S., Gillberg, I. C., Billstedt, E., Gillberg, C., and Olsson, I. (2005). Epilepsy in young adults with autism: a prospective population-based follow-up study of 120 individuals diagnosed in childhood. *Epilepsia* 46, 918–923. doi: 10.1111/j.1528-1167.2005.57504.x

Devlin, B., and Scherer, S. W. (2012). Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* 22, 229–237. doi: 10.1016/j.gde.2012.03.002

Dykens, E. M., Lee, E., and Roof, E. (2011). Prader-willi syndrome and autism spectrum disorders: an evolving story. *J. Neurodev. Disord.* 3, 225–237. doi: 10.1007/s11689-011-9092-5

Enriquez-Barreto, L., and Morales, M. (2016). The pi3k signaling pathway as a pharmacological target in Autism related disorders and Schizophrenia. *Mol. Cell. Ther.* 4, 2. doi: 10.1186/s40591-016-0047-9

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., et al. (2010). The NCBI biosystems database. *Nucleic Acids Res.* 38, D492–D496. doi: 10.1093/nar/gkp858

Ghiassian, S. D., Menche, J., and Barabási, A. L. (2015). A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare *de novo* variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70, 898–907 doi: 10.1016/j.neuron.2011.05.021

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Knoll, A. T., and Carlezon, W. A. (2010). Dynorphin, stress, and depression. *Brain Res.* 1314, 56–73. doi: 10.1016/j.brainres.2009.09.074

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-,h., Leotta, A., Kendall, J., et al. (2011). Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* 70, 886–897. doi: 10.1016/j.neuron.2011.05.015

Li, J., Shi, M., Ma, Z., Zhao, S., Euskirchen, G., Ziskin, J., et al. (2014). Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol. Syst. Biol.* 10:774 doi: 10.15252/msb.20145487

Lu, Y. C., Nazarko, O. V., Sando, R., Salzman, G. S., Sdhof, T. C., and Ara. (2015). Structural basis of latrophilin-flrt-unc5 interaction in cell adhesion. *Structure* 23, 1678–1691. doi: 10.1016/j.str.2015.06.024

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* 5:e13984. doi: 10.1371/journal.pone.0013984

Miles, J. H. (2011). Autism spectrum disorders–a genetics review. *Genet. Med.* 13, 278–294. doi: 10.1097/GIM.0b013e3181ff67ba

Mishra-Gorur, K., Caglayan, A. O., Schaffer, A. E., Chabu, C., Henegariu, O., Vonhoff, F., et al. (2014). Mutations in KATNB1 cause complex cerebral malformations by disrupting asymmetrically dividing neural progenitors. *Neuron* 84, 1226–1239. doi: 10.1016/j.neuron.2014.12.014

Mitra, K., Carvunis, A. R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732. doi: 10.1038/nrg3552

Monaco, A. P., and Bailey, A. J. (2001). Autism. The search for susceptibility genes. *Lancet* 358(Suppl. S3). doi: 10.1016/S0140-6736(01)07016-7

National Institute of Mental Health (2013). *Autism Spectrum Disorder.* Available online at: https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242–245. doi: 10.1038/nature11011

Noh, H. J., Ponting, C. P., Boulding, H. C., Meader, S., Betancur, C., Buxbaum, J. D., et al. (2013). Network topologies and convergent aetiologies arising from deletions and duplications observed in individuals with autism. *PLoS Genet.* 9:e1003523 doi: 10.1371/journal.pgen.1003523

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* 485, 246–250. doi: 10.1038/nature10989

Osterfield, M., Egelund, R., Young, L. M., and Flanagan, J. G. (2008). Interaction of amyloid precursor protein with contactins and NgCAM in the retinotectal system. *Development* 135, 1189–1199. doi: 10.1242/dev.007401

Parisi, L., Di Filippo, T., and Roccella, M. (2015). Autism spectrum disorder in kabuki syndrome: clinical, diagnostic and rehabilitative aspects assessed through the presentation of three cases. *Minerva Pediatr.* 67, 369–375.

Peters, S. U., Beaudet, A. L., Madduri, N., and Bacino, C. A. (2004). Autism in angelman syndrome: implications for autism research. *Clin. Genet.* 66, 530–536. doi: 10.1111/j.1399-0004.2004.00362.x

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* 94, 677–694. doi: 10.1016/j.ajhg.2014.03.018

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466, 368–372. doi: 10.1038/nature09146

Sabourdy, F., Labauge, P., Stensland, H. M. F. R., Nieto, M., Garc, V. L., Renard, D., et al. (2009). A manba mutation resulting in residual beta-mannosidase activity associated with severe leukoencephalopathy: a possible pseudodeficiency variant. *BMC Med. Genet.* 10:84. doi: 10.1186/1471-2350-10-84

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., et al. (2003). Human gene-centric databases at the Weizmann institute of science: genecards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.* 31, 142–146. doi: 10.1093/nar/gkg050

Sanders, S. J., Ercan-Sencicek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., et al. (2011). Multiple recurrent *de novo* cnvs, including duplications of the 7q11.23 williams syndrome region, are strongly associated with Autism. *Neuron* 70, 863–885. doi: 10.1016/j.neuron.2011.05.002

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241. doi: 10.1038/nature10945

Shariati, S. A. M., Lau, P., Hassan, B. A., Mller, U., Dotti, C. G., De Strooper, B., et al. (2013). APLP2 regulates neuronal stem cell differentiation during cortical development. *J. Cell Sci.* 126, 1268–1277. doi: 10.1242/jcs.122440

Smalley, S. L. (1998). Autism and tuberous sclerosis. *J. Autism Dev. Disord.* 28, 407–414. doi: 10.1023/A:1026052421693

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Toutain, A., Ronce, N., Dessay, B., Robb, L., Francannet, C., Le Merrer, M., et al. (1997). Nance-horan syndrome: linkage analysis in 4 families refines localization in Xp22.31-p22.13 region. *Hum. Genet.* 99, 256–261. doi: 10.1007/s004390050349

Trujillano, D., Perez, B., Gonzz, J., Tornador, C., Navarrete, R., Escaramis, G., et al. (2014). Accurate molecular diagnosis of phenylketonuria and tetrahydrobiopterin-deficient hyperphenylalaninemias using high-throughput targeted sequencing. *Eur. J. Hum. Genet.* 22, 528–534. doi: 10.1038/ejhg.2013.175

Vernon, M., and Rhodes, A. (2009). Deafness and autistic spectrum disorders. *Am. Ann. Deaf* 154, 5–14.

Volmar, C.-H., and Wahlestedt, C. (2015). Histone deacetylases (HDACs) and brain function. *Neuroepigenetics* 1, 20–27. doi: 10.1016/j.nepig.2014.10.002

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* 10, 280–293. doi: 10.1093/bfgp/elr024

Wijburg, F. A., Wegrzyn, G., Burton, B. K., and Tylki-Szymanska, A. (2013). Mucopolysaccharidosis type III (Sanfilippo Syndrome) and misdiagnosis of idiopathic developmental delay, attention deficit/hyperactivity disorder or autism spectrum disorder. *Acta Paediatr.* 102, 462–470. doi: 10.1111/apa.12169

Wu, L.-J., Mellstrm, B., Wang, H., Ren, M., Domingo, S., Kim, S. S., et al. (2010). Dream (downstream regulatory element antagonist modulator) contributes to synaptic depression and contextual fear memory. *Mol. Brain* 3:3. doi: 10.1186/1756-6606-3-3

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. *Adv. Neural Inf. Process. Syst.* 16, 321–328.

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 Genetics

ORIGINAL RESEARCH
published: 20 April 2017
doi: 10.3389/fgene.2017.00043

Pancreatic Islet Protein Complexes and Their Dysregulation in Type 2 Diabetes

Helle Krogh Pedersen[1†], *Valborg Gudmundsdottir*[1†] *and Søren Brunak*[1,2*]

[1] Department of Bio and Health Informatics, Technical University of Denmark, Kgs Lyngby, Denmark, [2] Disease Systems Biology, Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

Type 2 diabetes (T2D) is a complex disease that involves multiple genes. Numerous risk loci have already been associated with T2D, although many susceptibility genes remain to be identified given heritability estimates. Systems biology approaches hold potential for discovering novel T2D genes by considering their biological context, such as tissue-specific protein interaction partners. Pancreatic islets are a key T2D tissue and many of the known genetic risk variants lead to impaired islet function, hence a better understanding of the islet-specific dysregulation in the disease-state is essential to unveil the full potential of person-specific profiles. Here we identify 3,692 overlapping pancreatic islet protein complexes (containing 10,805 genes) by integrating islet gene and protein expression data with protein interactions. We found 24 of these complexes to be significantly enriched for genes associated with diabetic phenotypes through heterogeneous evidence sources, including genetic variation, methylation, and gene expression in islets. The analysis specifically revealed ten T2D candidate genes with probable roles in islets (*ANPEP, HADH, FAM105A, PDLIM4, PDLIM5, MAP2K4, PPP2R5E, SNX13, GNAS,* and *FRS2*), of which the last six are novel in the context of T2D and the data that went into the analysis. Fifteen of the twenty-four complexes were further enriched for combined genetic associations with glycemic traits, exemplifying how perturbation of protein complexes by multiple small effects can give rise to diabetic phenotypes. The complex nature of T2D ultimately prompts an understanding of the individual patients at the network biology level. We present the foundation for such work by exposing a subset of the global interactome that is dysregulated in T2D and consequently provides a good starting point when evaluating an individual's alterations at the genome, transcriptome, or proteome level in relation to T2D in clinical settings.

Keywords: diabetes, data integration, protein complexes, tissue specificity, pancreatic islets, patient network biology

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Christophe Magnan,
Paris Diderot University, France
Piero Marchetti,
University of Pisa, Italy

***Correspondence:**
Søren Brunak
soren.brunak@cpr.ku.dk

[†] *These authors have contributed equally to this work.*

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

Received: *01 February 2017*
Accepted: *27 March 2017*
Published: *20 April 2017*

Citation:
Pedersen HK, Gudmundsdottir V and Brunak S (2017) Pancreatic Islet Protein Complexes and Their Dysregulation in Type 2 Diabetes. Front. Genet. 8:43. doi: 10.3389/fgene.2017.00043

INTRODUCTION

Diabetes is a multi-tissue metabolic disease caused by defects in insulin action, insulin secretion, or both, resulting in hyperglycemia. The heritability of type 2 diabetes (T2D) has been estimated to range from 25 to 80% (Prasad and Groop, 2015). Despite that more than 120 T2D risk loci have been identified so far (Prasad and Groop, 2015) their combined effect explains only a fraction of the

heritability. The unexplained heritability of complex traits is expected to mainly reside in a large number of common and rare variants across the human genome (Yang et al., 2015). Identifying the remaining variants involved in T2D through traditional single-variant association analyses will require greatly increased sample sizes compared to current studies for improving statistical power (Morris et al., 2012). Integrative systems biology approaches hold the promise to facilitate this process by considering gene products in the context of cellular networks rather than in isolation, thus improving power through the use of existing biological knowledge.

Genome-wide analyses, such as genome-wide association studies (GWAS) and studies of differential expression or methylation, often rank thousands of genes for phenotype associations. Integrating such data is a powerful way to identify genes important in the disease pathogenesis that are not identifiable in any single dataset but become evident when considering the different evidence sources collectively (Kodama et al., 2012; Pers et al., 2013). Combining such integrative evidence with protein complexes provides additional insight into the biological context and has the potential to reveal novel therapeutic targets (Lage et al., 2012).

The subset of protein complexes active in a given tissue is restricted by the tissue-specific proteome, which is important to consider because disease-associated genes have a tendency to exhibit tissue-specific gene expression in affected tissues (Lage et al., 2008). Previous studies have shown that disease-gene prioritization is improved when using tissue-specific networks compared to tissue-naive protein interaction networks (Magger et al., 2012; Ganegoda et al., 2014). Consequently, considering disease associated genes in the appropriate context is a promising avenue for making further inroads into disease understanding (Gross and Ideker, 2015). Such tissue-specific analyses are now enabled by the increasing amount of large-scale tissue and cell type specific data sets (Lonsdale et al., 2013; Kim et al., 2014; Uhlén et al., 2015), making it possible to disentangle or deconvolute tissue and cell type-specific processes.

A key diabetes tissue is the islet of Langerhans, which plays an important role in diabetes pathology. Islets are scattered around in the pancreas where they only constitute 1–2% of the total organ mass. They consist of a number of different highly specialized endocrine cell-types with the insulin-producing beta-cells and glucagon-producing alpha-cells being of the highest relevance to diabetes (Danielsson et al., 2014). Utilizing tissue-specific data, one major aim of this study was to create a pancreatic and beta-cell specific resource of protein complexes to serve as an integration scaffold in this and future studies. Previous work on tissue-specific protein interaction networks did either not include human pancreatic islets (Guan et al., 2012; Barshir et al., 2013; Basha et al., 2015) or were restricted to tissue-specific gene expression data (Bossi and Lehner, 2009; Magger et al., 2012; Greene et al., 2015). By focusing on the pancreatic islet, we supplement these resources by integrating high-confidence physical protein interaction network data with islet-specific gene expression data from both microarray and RNAseq studies, as well as protein expression from immunohistochemistry-based protein profiling.

Another major aim of the study was to identify a set of islet protein complexes that are likely dysregulated or dysfunctioning in T2D. To investigate this, we searched for complexes that were enriched for genes implicated in diabetic phenotypes through heterogeneous sources of evidence, ranging from genetic variation to methylation and gene expression in islets. The resulting complexes thus represent functional units whose perturbation can give rise to a diabetic phenotype and at the same time provide insight into the genetic heterogeneity that contributes to the pathogenesis of T2D in pancreatic islets.

RESULTS

Defining a Catalog of 3,692 Islet Protein Complexes

We generated an islet-specific protein interaction network using gene and protein expression data combined with high-confidence protein interactions (see Section Methods and **Figure 1A**). This network was further decomposed into 3,692 overlapping protein complexes (10,805 genes) using the two complementary methods, ClusterOne (Nepusz et al., 2012) and spoke-hub, focusing on high internal connectivity and hub-topology, respectively (see Section Methods for details). We specifically chose network decomposition algorithms that allow for overlapping complexes as many proteins participate in multiple processes, making it difficult to decide on a single partition that closely reflects biological reality. These complexes, ranging in size from 6 to 50 proteins, captured different regions in network topology space, some being sparsely connected whereas others showed complete internal connectivity with all nodes having a physical interaction with all other nodes (Supplementary Table 1). This set of complexes represents a catalog of islet protein complexes and their constituents.

Coordinated Expression of Islet Protein Complexes

Tissue-specific coordination of gene expression among members of a protein complex may indicate an important function of the complex in the respective tissue (Han et al., 2004; Taylor et al., 2009; Börnigen et al., 2013). To investigate the status of the islet complexes, we calculated the degree of coordinated expression of each of the 3,692 complexes across a range of 34 tissues as the normalized average Pearson correlation coefficient of interacting proteins, using data from the GTEx consortium (Ardlie et al., 2015) and the study by Nica et al. (2013) (see Section Methods for details and **Figure 1B**). To evaluate the importance of coordinated expression for islet relevant complexes, we defined a set of 76 benchmarking islet complexes, each constituted by 10% of genes known to be of major importance for islet function and identity (Pasquali et al., 2014; **Figure 1C**). These benchmarking complexes had significantly higher coordinated expression in either islets, beta-cells, or non-beta islet cells compared to the background distribution of all other complexes (MWU, $P = 9.6 \times 10^{-4}$, Supplementary Figure 1). These results suggest that coordinated

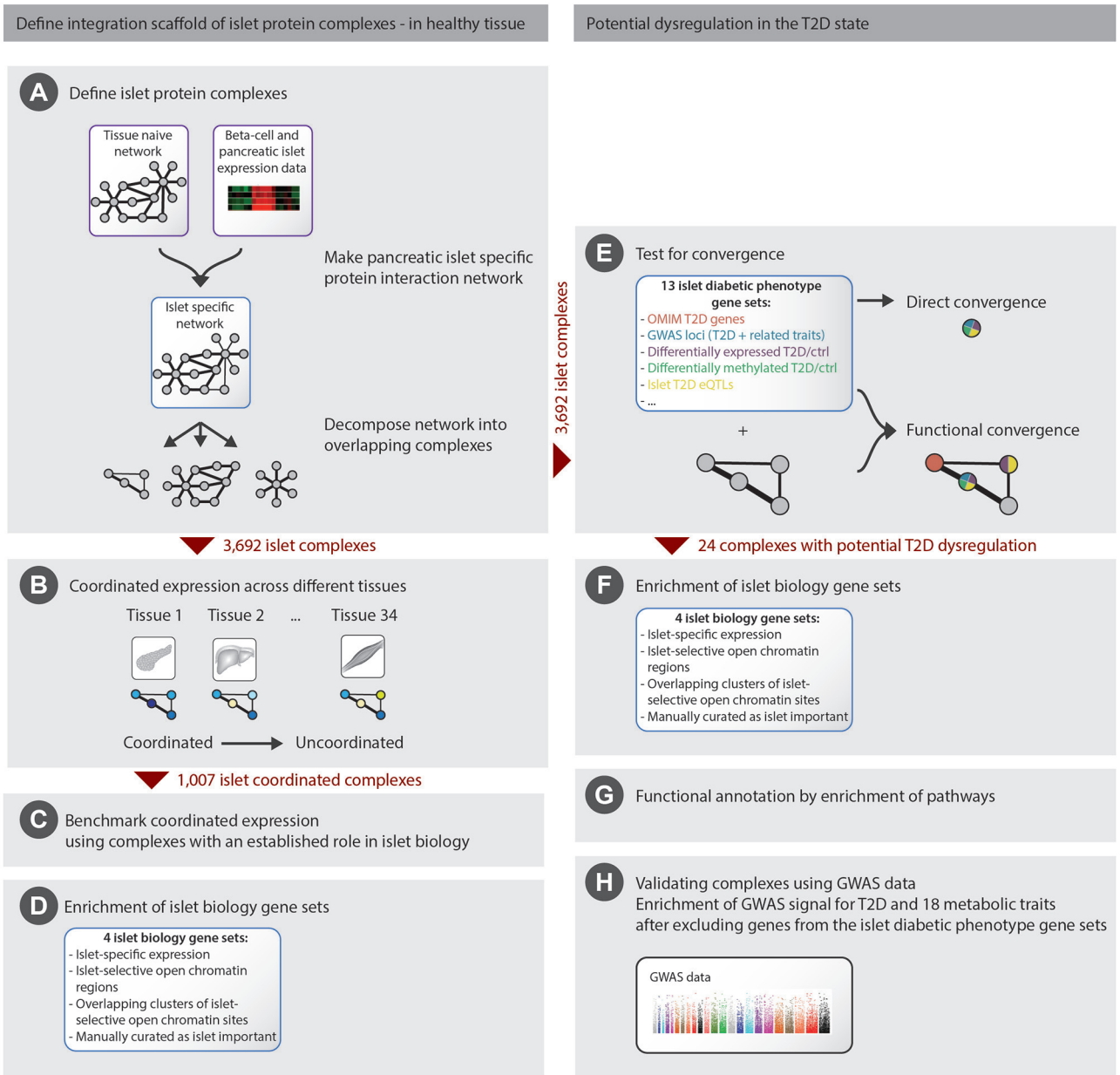


FIGURE 1 | Overview of the methodology employed. We first generated an integration scaffold of islet protein complexes in healthy tissue **(A)** and defined a subset of complexes with coordinated expression in islets **(B)** that were further benchmarked **(C,D)**. We then identified the subset of islet complexes with potential dysregulation in the T2D state by functional convergence of 13 islet diabetic phenotype gene sets **(E)**, followed by functional annotation and validation **(F–H)**. For comparison, direct convergence of the islet diabetic phenotype gene sets was evaluated **(E)**.

islet gene expression of protein complex members can indicate an important role in islet biology. We therefore defined a subset of 1,007 islet-coordinated complexes where at least one of the islet tissue components (whole islets, beta, or non-beta cells) was among the three highest ranked across the 34 tissues tested (see Section Methods). Moreover, the 1,007 complexes were enriched (MWU, $P = 2.8 \times 10^{-4}$) for genes residing in islet regulatory regions defined as having islet-selective open chromatin in the transcription start site or gene-body (Supplementary Table 2; **Figure 1D**).

While these 1,007 complexes are of special interest in the context of islet function, previous work related to the cell cycle (de Lichtenberg et al., 2005) has illustrated that protein complexes can be functional even though not fully coordinated due to sophisticated, temporal regulation. We therefore included all 3,692 complexes in the further analyses on T2D dysregulation.

Limited *Direct* Overlap of Islet Diabetes Gene Sets

Having a catalog of 3,692 islet relevant protein complexes we next turned to investigate which of those were most likely to be implicated in T2D (**Figure 1E**). The underlying hypothesis is that complexes exhibiting pronounced convergence of genes originating from different evidence sources related to diabetes are likely to play a role in the disease.

We thus compiled 13 sets of genes associated with T2D, monogenic forms of diabetes and related metabolic phenotypes (**Table 1**), hereafter termed islet diabetic phenotype gene sets. Despite all gene sets being related to diabetes, they generally showed surprisingly little direct overlap, although many pairwise overlaps were still larger than expected by chance (**Figure 2**). The largest overlaps, ranging from 11 to 55% relative to the size of the shortest list, were observed between gene sets based on genetic variation (Monogenic, OMIM, T2D GWAS/rare variant, Glycemic GWAS/rare variant, and Glycemic gene-based), which is to some extent expected as many genes causing monogenic forms of diabetes also harbor variants associated with T2D and

TABLE 1 | Description of the thirteen islet diabetic phenotype gene sets and the four islet biology related gene sets.

Name	Description	References	# Genes (# genes in network)
ISLET DIABETIC PHENOTYPE GENE SETS			
GWAS LOCI AND RARE VARIANT GENES			
T2D GWAS/rare variant	Genes in the vicinity of T2D GWAS SNPs, using a boundary of 110 kb upstream and 40 kb downstream of each gene, as well as genes harboring rare variants associated with T2D.	Morris et al., 2012; Albrechtsen et al., 2013; Flannick et al., 2014; Mahajan et al., 2014; Steinthorsdottir et al., 2014; Wessel et al., 2015	235 (162)
Glycemic GWAS/ rare variant	Genes in the vicinity of GWAS SNPs (FG, BMI-adjusted FG, 2 h Glu, BMI-adjusted 2 h Glu, insulinogenic index, disposition index, proinsulin), using a boundary of 110 kb upstream and 40 kb downstream of each gene, as well as genes harboring rare variants associated with FG, proinsulin, or insulinogenic index.	Strawbridge et al., 2011; Scott et al., 2012; Huyghe et al., 2013	135 (107)
GWAS GENES (GENE-BASED TEST)			
Glycemic gene-based	Genes associated with FG, 2 h Glu, or proinsulin using a gene-based analysis.	Scott et al., 2012; Huyghe et al., 2013	146 (130)
OMIM T2D GENES			
OMIM	Genes associated with "Diabetes mellitus, noninsulin-dependent; NIDDM" in the OMIM database (accession #125853)		26 (24)
MONOGENIC DIABETES GENES			
Monogenic	MODY and other monogenic diabetes genes.	McCarthy, 2010; Scott et al., 2012	28 (28)
ISLET eQTL GENES For 47 T2D SNPs (CIS AND TRANS)			
T2D eQTL	Five cis and 176 trans eQTLs in islets, based on 47 SNPs associated with T2D.	Taneera et al., 2012	163 (129)
GENES DIFFERENTIALLY METHYLATED IN ISLETS (T2D vs. CTRL)			
T2D methylation (A)	Genes in differentially methylated regions that are also differentially expressed.	Dayeh et al., 2014	113 (88)
T2D methylation (B)	Genes in differentially methylated regions.	Volkmar et al., 2012	221 (169)
GENES CO-EXPRESSED WITH 2+ T2D GENES			
Co-expression	Genes that are co-expressed in islets with 2 or more of 48 T2D genes.	Taneera et al., 2012	231 (197)
GENES DIFFERENTIALLY EXPRESSED IN ISLETS (T2D OR HYPERGLYCEMIC vs. CTRL)			
Hyperglycemia expression	Differentially expressed genes in islets, in hyperglycemic vs. normoglycemic individuals.	Taneera et al., 2012	121 (109)
T2D expression (A)	Differentially expressed genes in islets, in T2D patients vs. controls.	Taneera et al., 2012	106 (90)
T2D expression (B)	Differentially expressed genes in islets, in T2D patients vs. controls.	Dominguez et al., 2011	174 (150)
T2D expression (C)	Differentially expressed genes in beta-cells, in T2D patients vs. controls.	Marselli et al., 2010	281 (237)
ISLET BIOLOGY GENE SETS			
Islet specific	Top 30 islet specific genes.	Morán et al., 2012	33 (28)
Open chromatin	Genes with islet-selective (compared to five non-islet cell lines) open chromatin in the transcription start sites or gene-body.	Gaulton et al., 2010	319 (226)
Open chromatin clusters	Genes overlapping *clusters* of islet-selective open chromatin sites.	Gaulton et al., 2010	1,512 (1,340)
Islet biology	Sixty-seven genes curated as important for islet cell identity and function, Supplementary Table 2.	Pasquali et al., 2014	67 (57)

2 h Glu, 2 hour glucose; BMI, body mass index; FG, fasting glucose; T2D, type 2 diabetes.

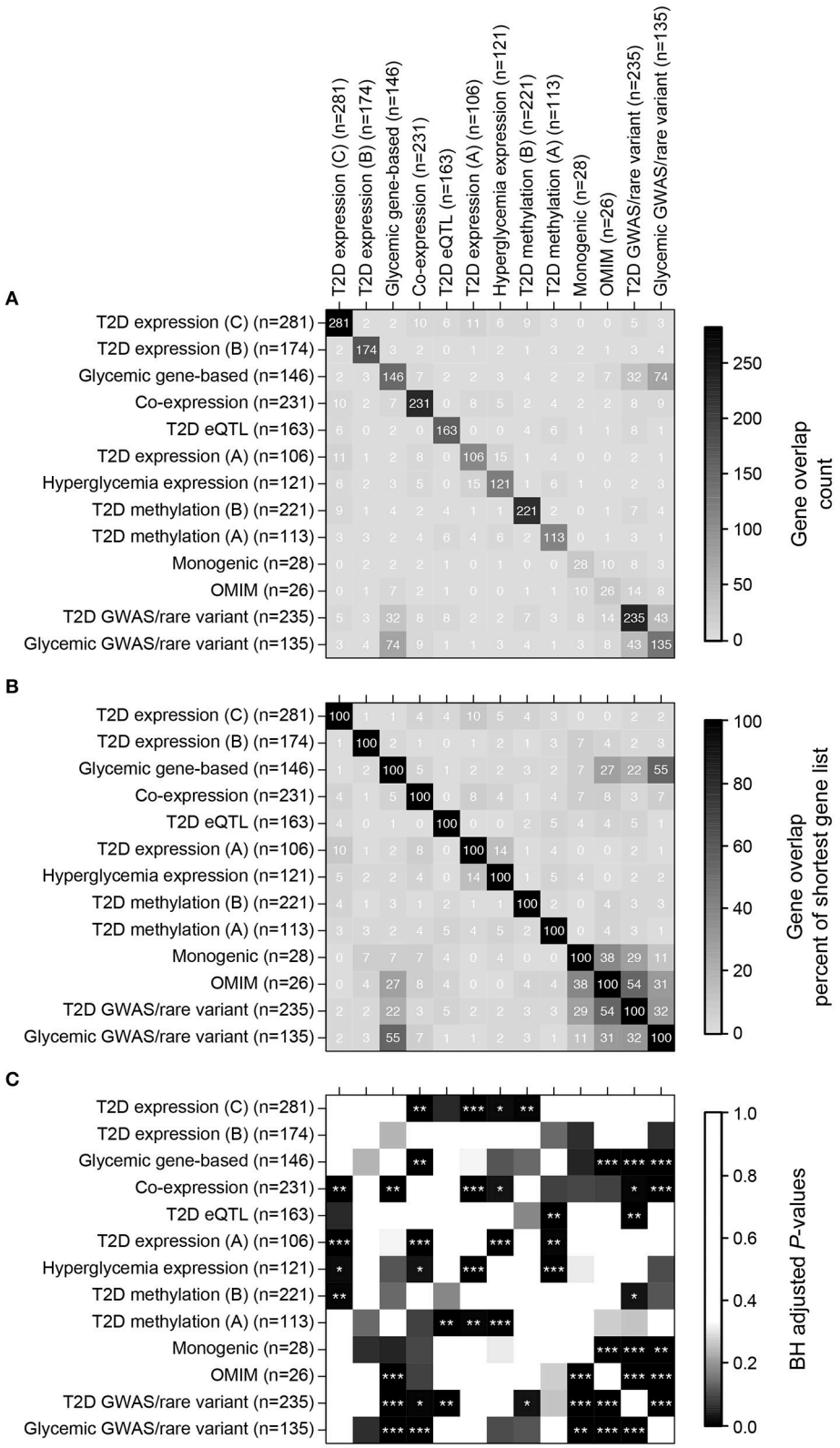


FIGURE 2 | Direct overlap of the thirteen islet diabetic phenotype gene sets. (A) Overlap in terms of gene counts. **(B)** Overlap in terms of percent overlap relative to the size of the shortest gene sets. **(C)** BH-adjusted *P*-values for testing significance of overlap (hypergeometric test using all 22,766 genes as background), stars are as follows: $^{***}P \leq 0.001$, $^{**}P \leq 0.01$, $^{*}P \leq 0.05$.

glycemic traits (Bonnefond and Froguel, 2015). Twenty genes were found to be part of four or more of the 13 gene sets (Supplementary Table 3), many of which are well-known T2D susceptibility genes while others are less well-established in the context of diabetes, some of those examples are highlighted in **Box 1**.

Complexes Showing *Functional* Overlap of Islet Diabetes Gene Sets

We next investigated if the 13 islet diabetic phenotype gene sets functionally converged on any of the 3,692 islet protein complexes, by calculating the combined enrichment for the 13 gene sets for each complex (**Figure 1E**). We found that the 1,007 complexes with coordinated expression in islets were enriched for small *P*-values (MWU, $P = 1.66 \times 10^{-5}$) and we furthermore observed significant convergence of the islet diabetic phenotype gene sets in 24 complexes (9 coordinated, 15 un-coordinated) after adjusting for multiple hypothesis testing (BH adjusted $P < 0.05$; Supplementary Table 4, Data Sheet 1). All of these 24 complexes contained one or more gene supported by genetic evidence (GWAS, rare variants or monogenic forms of diabetes), suggesting that the majority are likely to play a causal role in the development of T2D (Supplementary Table 4).

The 24 complexes were additionally enriched for genes in all four islet biology gene sets (Supplementary Table 2; **Figure 1F**), suggesting an important role in pancreatic islets. The complexes largely showed limited gene-overlap (Supplementary Figure 2), which indicates that they span different parts of the islet interactome.

We next investigated the biological functions of the 24 diabetic phenotype associated complexes (**Figure 1G**), and found that the complexes segregate into functional distinctive groups based on their pathway enrichment patterns (**Figure 3**). A number of these groups were characterized by molecular processes well-known to be dysregulated in diabetic islets—such as potassium channels, glucokinase, incretin signaling, and Wnt signaling—while others were enriched for processes less established in the islet pathogenesis of T2D, such as insulin-, interleukin-, and ephrin-signaling, cell and adherens junctions and neurotransmitter release. Interestingly, seven of the 24 complexes contained one or more target of FDA-approved drugs, many of which are not anti-diabetic agents (Data Sheet 1).

Leveraging the Complexes to Propose Novel T2D Genes

The 294 genes constituting the 24 complexes are all interesting in the context of diabetes (Supplementary Table 5). Obviously, many of them already have an established role in T2D. By contrast, the subset of 217 genes that were not part of any of the 13 islet diabetic phenotype gene sets comprise an interesting set for further prioritization. In particular, we identified six genes (*MAP2K4*, *PDLIM5*, *PPP2R5E*, *SNX13*, *GNAS,* and *FRS2*) of high interest as novel T2D associated genes, as they all have additional support for being of relevance for islet biology or function from the islet biology gene sets and furthermore SNPs in the vicinity of these genes are associated with T2D or glycemic traits with $P < 1 \times 10^{-4}$ (**Table 2**).

Interestingly, after our analysis was completed, a targeted study of variants in the *PDLIM5* gene reported an association with T2D (rs11097432, $P = 1.07 \times 10^{-3}$; Owusu et al., 2017). Additional support for the prioritized genes emerges from the recent wave of single-cell transcriptomics studies of human islets that were published after our analysis was finished (Segerstolpe et al., 2016; Wang et al., 2016; Xin et al., 2016; Lawlor et al., 2017). Remarkably, *GNAS* is among the 11 genes showing consistent differentially expression in diabetic cell types (compared to non-diabetic) with same direction of effect in beta-cells (higher in T2D) in the first three studies and, furthermore, one (of 41 genes) found by both Lawlor et al. and Segerstolpe et al. with same direction of effect in alpha-cells (lower in T2D; Lawlor et al., 2017). In addition, Xin et al. (2016), reports *GNAS* to be abundant in all four major islet endocrine cell types (alpha, beta, delta, PP) in both non-diabetic and T2D donors (but not

BOX 1 | T2D CANDIDATE GENES PRIORITISED BY DIRECT CONVERGENCE.

The following genes were supported by four or more of the thirteen islet diabetic phenotype evidence sources, many across different levels of molecular regulation, but have not been strongly established in the context of T2D.

The alanyl (membrane) aminopeptidase (*ANPEP*) gene resides in a locus on chromosome 15 containing variants associated with T2D in South Asian individuals (Kooner et al., 2011) and its expression levels are furthermore associated with the T2D associated SNP rs560887 (*G6PC2* locus on chromosome 2), thus, representing a trans-eQTL (Taneera et al., 2012). In addition, the *ANPEP* gene promoter is located in a region that is hypomethylated in T2D islets (Volkmar et al., 2012), and finally the gene itself is differentially expressed in T2D beta-cells (Marselli et al., 2010). Collectively, these heterogeneous data types indicate together a plausible role of *ANPEP* in the pathogenesis of T2D in pancreatic islets. Supporting our observation, this gene has been proposed as the causal gene in this GWAS locus through a study of allelic expression profiling (Locke et al., 2015). A variant in this gene is associated with the levels of a peptide derived from the C3 complement protein that plays a role in the innate immune system (Shin et al., 2014).

Hydroxyacyl-CoA dehydrogenase (*HADH*) was differentially expressed in islets in three independent data sets comparing T2D patients and controls, as well as being co-expressed in islets with two or more T2D candidate genes. Mutations in *HADH* are known to cause familial hyperinsulinism (Glaser, 2013), which motivated a targeted study of common variants in the gene that however did not find any association with T2D (van Hove et al., 2006). Yet, our observations suggest that the expression of the gene is affected in pancreatic islets in T2D and that it may play a role in the disease.

The islet expression of Family with sequence similarity 105, member A (*FAM105A*) and PDZ and LIM domain 4 (*PDLIM4*) was associated with both T2D (Marselli et al., 2010; Taneera et al., 2012) and hyperglycemia (Taneera et al., 2012). *FAM105A* was furthermore coexpressed with the T2D genes *SLC30A8*, *G6PC2* and *KCNJ11* (Taneera et al., 2012) while *PDLIM4* resides in a region of the genome that was differentially methylated in islets when comparing T2D patients and controls (Dayeh et al., 2014). A variant upstream of *PDLIM4* (rs7727038) shows a nominal association ($P = 5.2 \times 10^{-5}$) with fasting glucose in the MAGIC consortium (Dupuis et al., 2010). Both of these genes encode for proteins with relatively unknown functions.

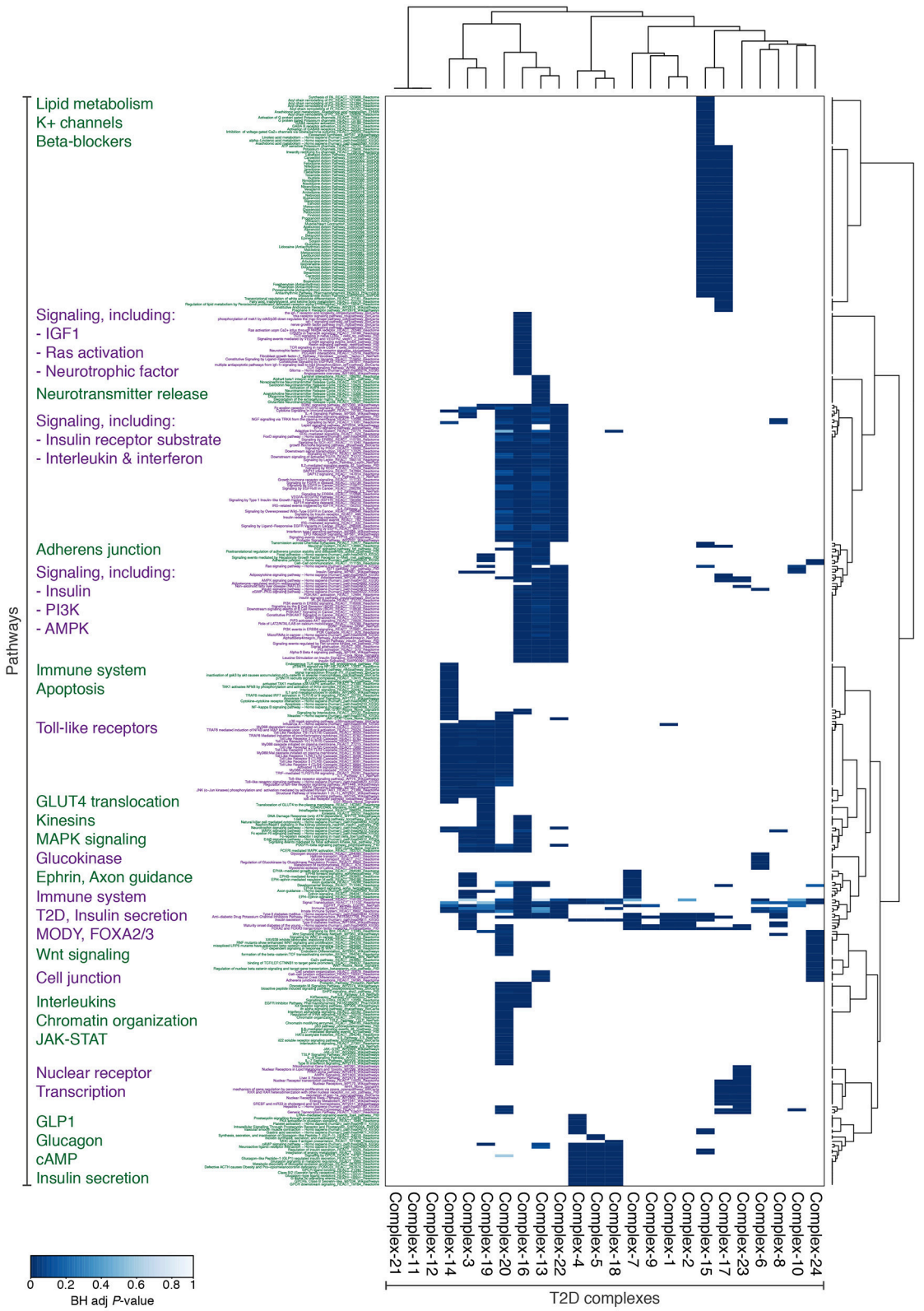


FIGURE 3 | The 24 complexes with potential T2D dysregulation are enriched for diverse and relevant functions. Subset of Consensus PathDB-pathways, for which at least one protein complex is enriched with BH-adjusted $P < 0.001$. The pathways and complexes are clustered with Ward's hierarchical clustering using an asymmetric binary similarity measure.

TABLE 2 | Plausible novel T2D genes prioritized from the complexes with potential T2D dysregulation.

Gene symbol	Gene name	# Islet diabetic phenotype gene sets	# Islet biology gene sets	Minimum *P*-value for associated SNPs	Corresponding GWAS trait
MAP2K4	Mitogen-activated protein kinase kinase 4	0	1	7.83×10^{-6} (rs929441)	AUCIns/AUCGluc
PDLIM5	PDZ and LIM domain 5	0	1	9.87×10^{-5} (rs17021900)	Fasting glucose
PPP2R5E	Protein phosphatase 2, regulatory subunit B, epsilon isoform	0	1	7.05×10^{-5} (rs10151995)	Fasting glucose
SNX13	Sorting nexin 13	0	1	4.02×10^{-6} (rs2723517)	HbA1c
GNAS	GNAS complex locus	0	1	4.73×10^{-5} (rs6026565)	Fasting glucose, Manning
FRS2	Fibroblast growth factor receptor substrate 2	0	1	9.76×10^{-6} (rs12425398)	Fasting glucose, Manning

Genes are prioritized if they, besides being part of a protein complex showing potential T2D dysregulation, are part of at least one of the four islet biology gene sets and harbor at least one SNP with $P < 1 \times 10^{-4}$ in one or more of the 19 GWAS described in Supplementary Table 6. Only the best SNP P-value and corresponding GWAS trait are shown. To focus on novel T2D genes, genes in any of the 13 islet diabetic phenotype gene sets are excluded.

significantly differentially expressed). *SNX13* also exhibits cell type specific differential expression in T2D, being lower in delta cells of diabetic donors (fold change $= -13.02$, FDR $= 4.93 \times 10^{-2}$; Xin et al., 2016). Whole islet gene expression (profiled with microarrays and RNA-seq) is further nominally associated with lower HbA1c levels for both *GNAS* ($P = 2.14 \times 10^{-3}$, FDR $= 4.30 \times 10^{-2}$) and *SNX13* ($P = 1.61 \times 10^{-2}$, FDR $= 1.02 \times 10^{-1}$; Fadista et al., 2014). In mice, disruption of the G protein α-subunit (one of the *GNAS* gene products) maternal (but not paternal) allele leads to severe obesity, hypertriglyceridemia, impaired glucose tolerance and insulin resistance (Xie et al., 2008). Together, these observations add support for the genes being important for shaping the diabetic phenotype in one or more islet cell types.

Both *MAP2K4* and *GNAS* are known to be involved in pancreatic cancer [Cancer Gene Census (Forbes et al., 2017) and Intogen (Gonzalez-Perez et al., 2013) databases]. The *MAP2K4* gene encodes the mitogen-activated kinase kinase (MKK)4, which constitutes a part of the apoptotic-effect mediating MEKK1-MKK4-JNK pathway (Xia et al., 1995) and is inhibited in pancreatic beta-cells by the glucagon-like peptide-1 analog exending-4, resulting in protection from palmitate-induced apoptosis (Natalicchio et al., 2013). *MAP2K4* is furthermore a proposed tumor suppressor gene and is significantly under-expressed in metastatic compared to benign pancreatic endocrine tumors (or islet cell tumors; Couvelard et al., 2006). These results point to an important role of *MAP2K4* in the survival of pancreatic islet cells, which is a process central to the etiology of both diabetes and pancreatic carcinomas. Further studies of the potential dual role of *MAP2K4* and *GNAS* might help elucidating the molecular basis for the complex bidirectional relationship observed between diabetes and pancreatic cancer (Li, 2012).

Verification of Potential T2D Dysregulation of Complexes Using GWAS Data

As the 24 complexes were enriched for genes associated with diabetes and glycemic traits (input genes), it is likely that their disruption gives rise to these phenotypes. Thus, the remaining (non-input) genes in the complexes have a high likelihood of also contributing to the same traits. We tested this hypothesis by investigating the enrichment of GWAS signals for T2D and glycemic traits from the DIAGRAM and MAGIC consortiums in each of the 24 complexes (**Figure 1H**). Using the Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) tool to test the enrichment, we identified 30 significant ($P < 0.05$) complex-trait combinations, spanning 15 complexes and 13 traits, of which 25 remained significant after excluding any genes that were used as input in the corresponding gene sets used for discovery of the complexes (Supplementary Figure 3). The last definition was applied to avoid any circularity, as the different GWAS might be the source of the association leading to the gene being in the islet diabetic phenotype gene sets that were used to define the 24 complexes with potential T2D dysregulation. These results indicate that the non-input genes in the complexes indeed harbor variants that are associated with the same phenotypes, although not so strongly that they could be discovered by the GWAS analysis alone.

We further investigated if the GWAS enrichment within the complexes was driven by many genes in loci with modest associations converging in the same functional context or mainly by one or a few genes with low *P*-values (minimum *P*-value for the SNPs mapping to the respective genes). We therefore repeated the analysis after excluding all genes with genome-wide significant *P*-values ($P < 5 \times 10^{-8}$) in the respective GWAS and found that the enrichment for 27 out of 30 complex-trait combinations remained significant (**Figure 4**, Group 1). This suggests that the majority of the complexes represent examples where many small effects collectively perturb their function, leading to a molecular phenotype that gives rise to disturbed glucose homeostasis. All of the three complex-trait combinations that became non-significant (**Figure 4**, Group 2) contained one or more gene with a genome-wide significant signal ($P < 5 \times 10^{-8}$), indicating that these genes were the main driver of the enrichment.

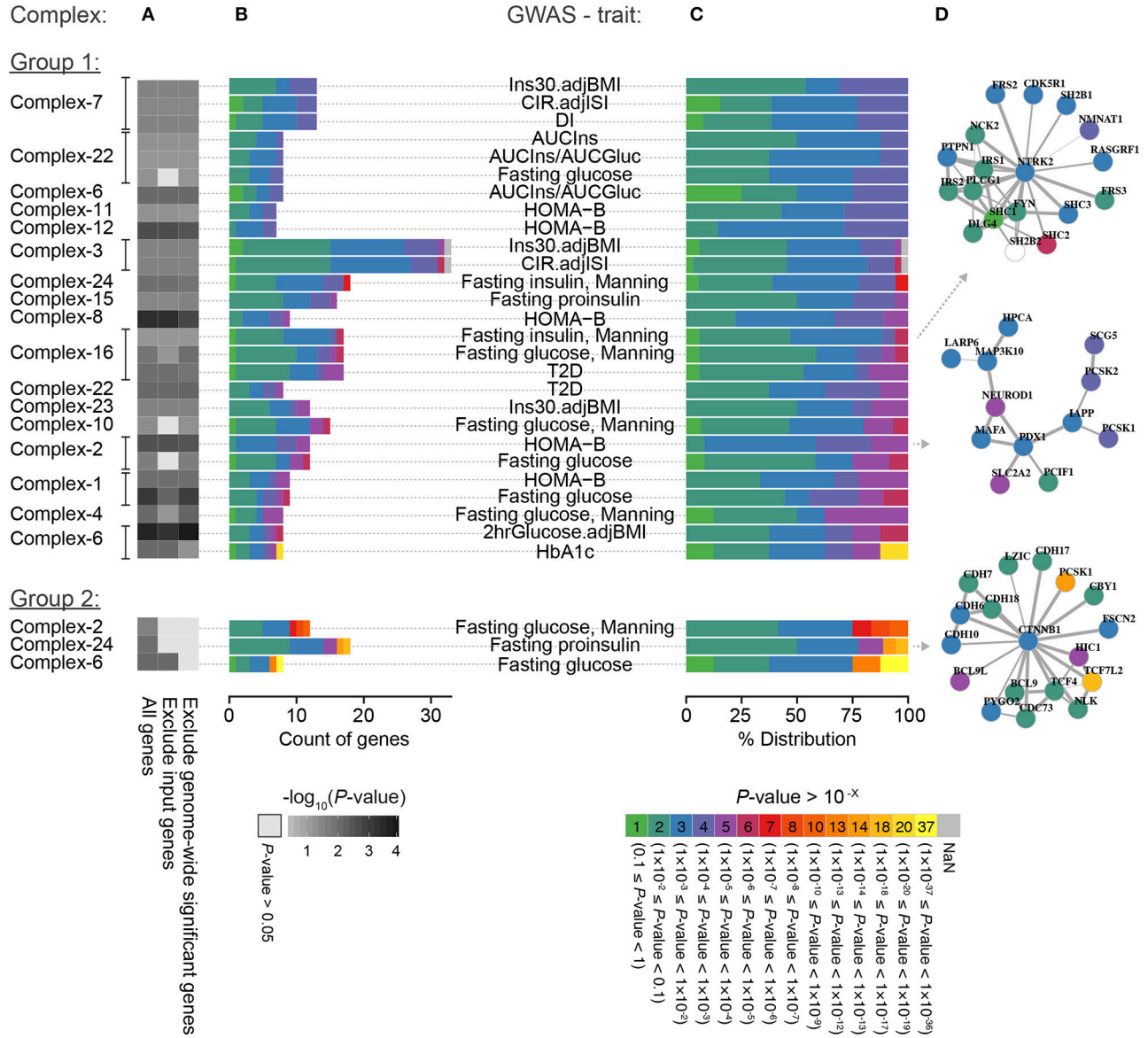


FIGURE 4 | Breakdown of significantly enriched complex-trait combinations. (A) The enrichment of GWAS signals for each of the 30 significant complex-trait combinations when including all genes, excluding input genes, and excluding genes with genome-wide significant association in the given GWAS (see Section Methods for details). The genes in each complex-trait combination are colored based on *P*-value (i.e., minimum *P*-value for the SNPs mapping to the respective gene) partitioned into factor-10 groups; **(B)** actual count and **(C)** percentage distribution of gene *P*-values within a complex in the GWAS for the given glycemic trait. **(D)** Example of complexes.

The Nature of the Evidence Sources behind the Enrichment

The 24 diabetic phenotype associated complexes could further be characterized by the diversity of supporting data driving their enrichment, such as the proportion of genes in the complex supported by multiple gene sets and the total number of gene sets supporting each complex. More specifically, we observed three notable trends (**Figure 5**) where the enrichment of a complex was mainly driven by (a) genes supported by multiple sources each, (b) genes supported by one or few sources each and few in total, and (c) genes supported by one or few sources each but many in total. A representative example from each of these three groups of complexes is shown in **Figure 5**. In group (A), the complex Complex-2 consisted of many genes that are associated with multiple diabetic phenotypes each and are well-established in the context of diabetes, including the transcription factor *NEUROD1*, which is required for normal beta-cell development, and *SLC2A2*, which encodes GLUT2—the main glucose sensor in rodent beta-cells (but not human; McCulloch et al., 2011). Furthermore, the complex contained a

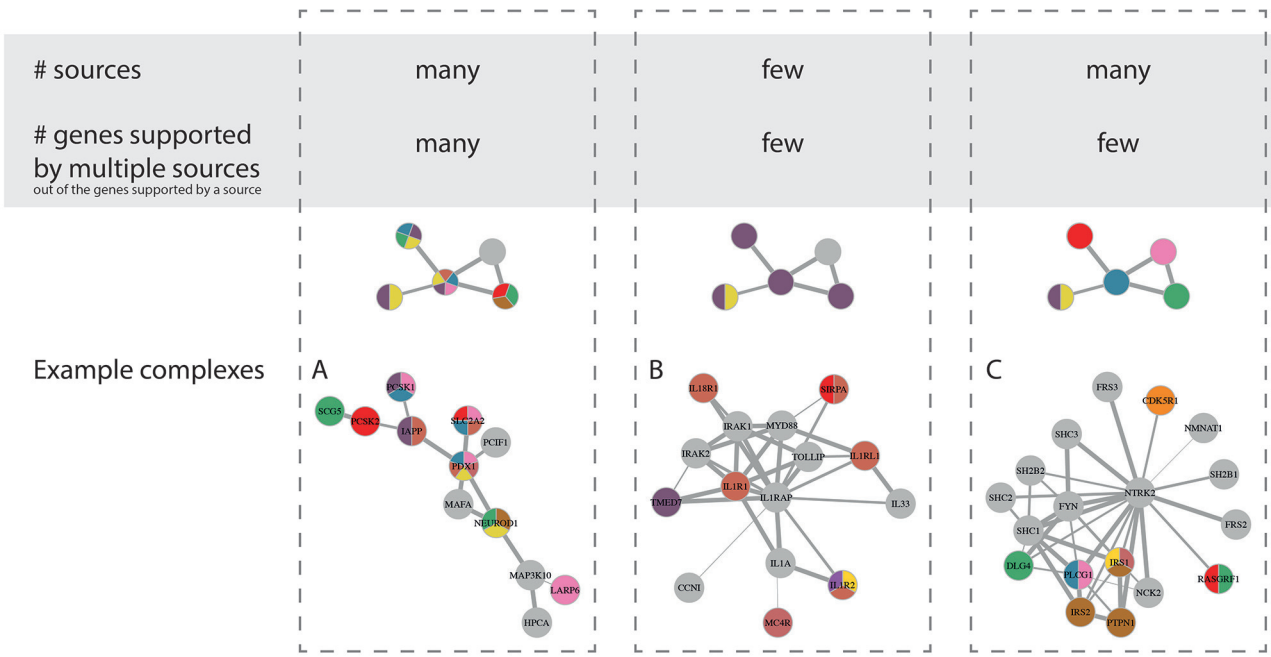


FIGURE 5 | Highlevel grouping of complexes by nature of evidence driving their enrichment. Schematic visualization (top) and representative examples (bottom) for the three overall groups. The fourth theoretical category with few sources but a high percentage of genes supported by multiple sources is excluded here, as we did not observe any good examples. Group A, Complex-2; group B, Complex-14; and group C, Complex-16.

number of genes directly involved in insulin transcription and secretion, such as the insulin regulating transcription factors *PDX1* and *MAFA*, *PCSK1* and *PCSK2*, which are known to localize with insulin in islets, *IAPP*, which is co-secreted with insulin and *SCG5*, which is a marker of insulin secreting tumors. Interestingly, the *LARP6* gene in the complex was included in the islet diabetic phenotype gene sets because of its proximity to the fasting proinsulin associated SNP rs1549318 (Strawbridge et al., 2011). Its presence in the complex suggests that *LARP6* may play an important role in beta-cell function and insulin secretion. In line with the function of the genes in the complex, the overall complex was enriched for genetic associations with HOMA-B based on MAGIC data.

Complex-14 is an example from group (B), where the enrichment was driven by genes mainly supported by the same gene set (5/7 genes), namely the "Hyperglycemia expression" data. The additional supporting gene sets were mainly from gene expression or methylation sources, while it only contained one gene (*MC4R*) supported by genetic evidence that was only weakly connected to the remainder of the complex. Furthermore, no enrichment was found for low SNP *P*-values in MAGIC and DIAGRAM data. This might as such be an example of a complex that is rather involved in a response to the diabetic state in the islets than playing a causal role. This is fitting with it being mainly composed of interleukins and toll-like receptors and enriched for inflammatory response and apoptosis pathways that have a clear relevance to the beta-cell mass deterioration in T2D pathogenesis.

Finally, Complex-16 is an example of a complex where the enrichment was supported by multiple sources, but few consensus support genes. Such complexes are interesting because they could not have been revealed using any data type alone, but constitute a functionally related group of genes that are identified by multiple types of diabetes-associated evidence. Complex-16 was strongly enriched for brain-derived neurotrophin (BDFN) signaling. BDFN has indeed been shown to affect the histological organization of beta and non-beta cells in the pancreatic islets (Yamanaka et al., 2006). This complex was furthermore enriched for GWAS signals for fasting glucose levels, fasting insulin levels and T2D.

DISCUSSION

To harvest the power of data integration, we have brought together results from genetic studies of islet-relevant phenotypes and human islet studies spanning different levels of molecular regulation. We identify 24 protein complexes with strong supporting evidence for being implicated in diabetes pathogenesis in pancreatic islets and show how they are enriched for multiple modest effects of genetic variants associated with glycemic traits. Furthermore, we specifically prioritize ten candidate genes for T2D, of which six are novel, based on the investigation of either direct or functional convergence of the evidence sources. Additionally, we compose a set of 3,692 islet protein complexes that can serve as an integration scaffold for future studies.

By comparing the direct overlap between the heterogeneous islet diabetes-related gene sets we identified genes such as *ANPEP* and *HADH* that are currently not well-established as diabetes susceptibility genes but had consensus support across evidence sources. These observations highlight that such a straightforward

data integration approach is able to pinpoint potentially new disease genes. Apart from these few, but interesting, examples of genes that were part of multiple gene sets, the generally limited direct overlap between the gene sets emphasizes the necessity of integrative systems biology approaches focusing on functional entities rather than single genes for further understanding of the dysregulation and dysfunctioning occurring in diabetic islets.

Previous work on congenital heart disease (Lage et al., 2012) has shown similar results, where a limited overlap was observed between genes identified in different types of genetic studies whereas they converged significantly in protein networks related to heart development. Here we extended this approach to T2D, where we found the prioritized complexes to mainly be involved in signaling cascades, immune functions, apoptosis and cell-cell communication in addition to the expected insulin secretion pathway. We thus show that these particular molecular mechanisms are consistently supported by complementary types of molecular data from human islets to form a major component of the T2D etiology. These results reduce the many previously observed pathways related to T2D pathogenesis in human and animal islets from single omics studies to a set of highly credible pathways.

A previous systems genetics study of the T2D state in human islets (Taneera et al., 2012) identified a set of 20 genes that collectively explained a significant portion of HbA1c variation. Here we add to those results by combining multiple independent data sets to identify nine additional T2D candidate genes that likely play a role in pancreatic islets. Furthermore, we prioritized specific protein complexes and their associated pathways that provide biological insight into T2D pathogenesis.

The majority of the 24 protein complexes found in this study were enriched for modest GWAS signals, suggesting that multiple small effects collectively perturb the complexes and give rise to variation in glycemic phenotypes. We thus provide insight into the mechanisms by which common genetic variation translates into a disease phenotype, which supports that the multifactorial genetic architecture of complex traits is constituted by a large number of variants disrupting cellular networks (Schadt, 2009).

An advantage to investigating functional convergence on protein complexes is that not all genes in the complex need to have prior diabetes-related evidence for the complex to be significant. Consequently, this approach concurrently prioritizes genes without prior diabetes-related evidence, but whose products interact with other diabetes relevant proteins in the islet, such as the six T2D candidate genes highlighted in **Table 2**. Furthermore, complexes containing both genes from GWAS loci and genes supported by other evidence sources, provide support for the GWAS gene mediating the signal in that locus, such as *LARP6* in the complex Complex-2 that resides in a proinsulin associated GWAS locus. Lastly, the complexes provide a functional context for the disease genes. Many genes naturally participate in several functions, reflected by the overlap of many of the complexes. For such multifunctional genes, the approach outlined here prioritizes the subset of disease relevant complexes and thus the disease relevant functions.

A major goal for T2D and other common diseases is to identify causal pathways and network modules underlying disease pathogenesis to enable precise risk prediction and development of new therapeutic strategies (McCarthy, 2015). Furthermore, such pathways and network modules need to be identified in a tissue-specific context (Gross and Ideker, 2015). Here we provide causal network modules for T2D in the form of tissue-specific protein complexes that provide more biological insight into the disease pathogenesis than disease genes in isolation and furthermore form a basis for integrating person-specific genetic, transcriptomic, or proteomic profiles in a clinical setting. Dissecting these complexes can moreover reveal new drug-targets, such as genes interacting with targets of currently used anti-diabetic medications, genes supported by multiple evidence sources or their more druggable interaction partners. Furthermore, complexes that contain targets of FDA-approved drugs may highlight opportunities for drug repurposing in the search for new diabetes treatments.

METHODS

Construction of a Pancreatic Islet-Specific Protein Interaction Network

Previous tissue-specific protein interaction networks mainly fall into three categories: node-removal, where interactions between proteins absent in the given tissue are excluded (Bossi and Lehner, 2009; Barshir et al., 2013; Basha et al., 2015), edge-reweight, where interactions between absent proteins are down-weighted (Magger et al., 2012), and data-driven Bayesian methodologies (Guan et al., 2012; Basha et al., 2015; Greene et al., 2015). Here we created both an edge-reweighted as well as a node-removal islet-specific protein interaction network, to accommodate downstream network analysis approaches that did or did not consider edge-weights, respectively.

The islet-specific protein interaction networks were constructed by pruning high confidence protein interaction from an updated version (2014) of the InWeb database (Lage et al., 2007; 14,536 proteins with 337,951 high-confidence interactions) using the data sets described in Supplementary Table 7. More specifically, for the node-removal protein interaction network, genes not passing the specified cutoffs in all of the data sets were considered less likely to be expressed in pancreatic islets and thus removed from the pruned islet network. For the edge-reweighted protein interaction network, lowly expressed genes were not removed but instead the confidence score of their interactions were down-weighted using the approach proposed by Magger et al. (2012):

$$w'_{ij} = w_{ij} * rw^n$$

where w_{ij} is the original edge weight between protein i and j, n is the number of lowly expressed genes in the tissue constituting the interaction (i.e., {0,1,2}), and rw is the probability that a gene is expressed in the tissue even though it does not pass the cut-offs listed below, which was chosen to be 0.1.

If genes were not covered by any of the data sets—or in the case of the Human Protein Atlas (HPA) data, annotated with uncertainty—a benefit-of-the-doubt approach was applied where such genes were considered present.

Included Data-Sources

Tissue-specific protein expression profiles based on immunohistochemistry using tissue microarrays were obtained from the HPA version 13, 11/6-2014, downloaded on 10/3-2015 from www.proteinatlas.org, with Ensembl version 75.37 (Uhlén et al., 2015). Proteins were categorized as present, absent, or uncertain based on the reliability and level of their expression value. Specifically, proteins with supportive expression values were categorized as absent if they were not detected and otherwise as present if they had low, medium or high expression values whereas proteins with uncertain expression values were categorized as uncertain.

Microarray gene expression data from the GNF Tissue Atlas (GNF) (GEO: GSE1133) was downloaded from BioGPS (http://biogps.org/downloads/; Su et al., 2004).

Defining Topology-Based Complexes within the Network

Many different methods with different objective functions have been proposed for defining clusters of genes in protein interaction networks. Here we applied two complementary approaches; one aiming at identifying tightly connected genes, and one centered on spoke-hub complexes as often applied in previous work (Lage et al., 2007; Börnigen et al., 2013).

Strongly connected components in the edge-weighted islet-specific network were identified by ClusterONE, a non-partitioning graph decomposition algorithm (Nepusz et al., 2012), using a minimum density of 0.2, which is calculated as the average edge weight within the complex if missing edges are assumed to have a weight of zero, and a maximum overlap of 0.3 between two complexes before they were merged using the multi-merge option, and otherwise default parameters. ClusterONE uses the matching score as default for calculating the overlap between two complexes, which is defined as the intersection size squared, divided by the product of the sizes of the two complexes.

A three-step approach was applied to define spoke-hub-complexes. First, for each gene in the network a complex was defined by all its first order interaction partners. Next a topology filter was applied to prune complexes for interaction partners that tend to interact with many proteins in an unspecific way, due to either experimental artifacts or for biological reasons. In brief, genes were removed from the complex if <5% of its interaction partners were within the given complex. Lastly, overlapping clusters were merged using the same approach as for ClusterONE. Since this approach ignores edge-weights it was applied to the node-removal version of the islet-specific protein interaction network.

Finally, overlapping complexes resulting from the two approaches were merged using the same approach as before. Complexes with fewer than 6 or more than 50 nodes were discarded in the downstream analysis, resulting in 3,692 islet complexes. Diameter and average degree, clustering coefficient and betweenness-centrality were calculated for each complex using the igraph R-package (Csardi and Nepusz, 2006).

Coordinated Expression of Protein Complexes

The TissueRanker approach (Börnigen et al., 2013) utilizes the assumption that a mutation in a hub-spoke complex is likely to have an affect in tissues where the proteins within the complex show high degree of coordinated expression and thus, that the degree of coordinated expression may aid in prioritizing tissues in which the complex is active and where deregulation of the complex could be detrimental. Here we extended the methodology to complexes with more complex topology. In brief the $PCC.mean_c^t$ for complex c in tissue t is defined as the average pairwise Pearson correlation coefficient (PCC) of gene expression (RPKM values) between any two interacting genes within the complex for the given tissue:

$$PCC_{xy}^t = \frac{\sum_{i=1}^{N_s}\left(x_i^t - \overline{x^t}\right)\left(y_i^t - \overline{y^t}\right)}{\sqrt{\sum_{i=1}^{N_s}\left(x_i^t - \overline{x^t}\right)^2}\sqrt{\sum_{i=1}^{N_s}\left(y_i^t - \overline{y^t}\right)^2}}$$

$$PCC.mean_c^t = \frac{\sum_{x=1}^{N_g}\sum_{y\in I_x} PCC_{xy}^t}{2\cdot N_e}$$

where N_s is the number of samples for tissue t, N_g is the number of genes in protein complex c, N_e is the number of edges in protein complex c, and I_x is the interaction partners of gene x excluding any self-loops.

To alleviate any potential bias arising from different numbers of tissues samples (Börnigen et al., 2013) we further standardized the $PCC.mean_c^t$ values within a tissue by first converting the average correlation coefficients to an approximately normal distribution using Fisher transformation:

$$z_c^t = \frac{1}{2}\ln\frac{1 + PCC.mean_c^t}{1 - PCC.mean_c^t}$$

$$CE_c^t = \frac{z_c^t - \mu^t}{\sigma^t}$$

The resulting z-scores are here referred to as coordinated expression (CE) and used to compare tissue relevance across tissues for a given complex.

RPKM values from RNAseq data for 31 tissues from the Genotype-Tissue Expression (GTEx) project were obtained through the database of Genotypes and Phenotypes (dbGaP) (study accession phs000424.v4.p1, version from 17/1-2014; Mailman et al., 2007). However, since the GTEx data does not include pancreatic islets, RNAseq data for whole islets, beta cells, and non-beta cells (from pancreatic islets; Nica et al., 2013) were combined with the GTEx data.

We defined 1,007 islet complexes with coordinated expression as the subset of the 3,692 islet complexes where at least one of the islet tissue components (whole islet, beta, and non-beta cells) was among the three tissues with highest coordinated expression level among the 34 included tissues.

Compiling Islet Biology and Islet Diabetic Phenotype Related Gene Sets

We compiled a set of 13 complementary sets of genes associated with T2D, monogenic forms of diabetes and related metabolic

phenotypes (**Table 1**). These 13 gene sets are collectively referred to as islet diabetic phenotype gene sets and were chosen because of their relevance to the islet tissue.

We obtained GWAS SNPs and genes supported by gene-based tests for T2D (Morris et al., 2012; Mahajan et al., 2014), fasting glucose (Dupuis et al., 2010; Scott et al., 2012), 2 hour glucose (2 h glu) during an oral-glucose tolerance test (Dupuis et al., 2010; Scott et al., 2012), and proinsulin (Strawbridge et al., 2011). SNPs in GWAS loci were mapped to a gene if they fell within 110 kb upstream or 40 kb downstream of its transcription start and stop sites respectively, as these boundaries have been shown to capture the majority of *cis*-eQTLs associations (Veyrieras et al., 2008; Ardlie et al., 2015). We additionally included all genes that were reported in eQTL associations for the GWAS SNPs from the respective publications (Dupuis et al., 2010; Voight et al., 2010; Strawbridge et al., 2011; Morris et al., 2012; Scott et al., 2012; Mahajan et al., 2014). We also included genes harboring rare variants associated with either fasting glucose and T2D (Albrechtsen et al., 2013; Flannick et al., 2014; Steinthorsdottir et al., 2014; Wessel et al., 2015) or insulin processing and secretion (Huyghe et al., 2013). Genes associated with monogenic forms of diabetes were obtained from a literature review (McCarthy, 2010) and a curated list from a previous study (Morris et al., 2012).

Genes differentially expressed in islets were obtained from a study by Taneera et al. (2012). In addition, two other microarray datasets of beta-cell and islet gene expression, respectively, were obtained from the Gene Expression Omnibus database (accession IDs: GSE20966 and GSE25724) and differential gene expression between T2D patients and non-diabetic controls evaluated using the "limma" R package as implemented in the NCBI GEO2R tool. Genes with $P < 0.001$ were included in the gene sets, except for the dataset by Dominguez et al. (2011) where a stricter cutoff of $P < 0.0001$ was applied due to inflated significance values. We further included additional gene sets defined by the islet gene expression study from Taneera et al. (2012), namely genes that showed *cis*- or *trans*-eQTL associations with T2D associated SNPs and genes that were co-expressed with >2 T2D candidate genes. Finally, we included genes that were differentially methylated in islets in T2D patients compared to non-diabetic controls (Volkmar et al., 2012) or were both differentially methylated and differentially expressed (Dayeh et al., 2014).

We furthermore constructed four gene sets related to islet function, referred to as islet biology gene sets (**Table 1**). These sets included genes with islet-specific expression (Morán et al., 2012), genes in islet-selective open chromatin regions or genes overlapping clusters of islet-selective open chromatin sites (Gaulton et al., 2010) and genes manually curated as islet important (Pasquali et al., 2014).

Finally, we obtained at list of proteins that are targets of FDA approved drugs from the druggable human proteome (Uhlén et al., 2015).

The direct overlap of the gene sets was tested using a hypergeometric test with all 22,766 human genes as background.

Functional Convergence Testing

To test the protein complexes for potential dysregulation in T2D, the degree of functional convergence of diabetes related genes was assessed. For each complex, the enrichment of each of the 13 islet diabetic phenotype gene sets was first calculated using a hypergeometric test and the corresponding *P*-values were next combined using Fisher's combined probability test ($P_{combined}$).

The likelihood of observing a similar degree of functional convergence by chance was estimated for each complex by randomly sampling 100,000 sets of the same number of genes from the whole network. An empirical *P*-value (P_{emp}) was then calculated by counting how many of these 100,000 random sets had a $P_{combined} \leq$ the real case divided by the number of random sets ($n = 100{,}000$). P_{emp} was adjusted for multiple hypotheses testing across complexes using a Benjamini–Hochberg correction and complexes with $P_{emp.adj} < 0.05$ were considered significant and thus showing potential T2D dysregulation. Genes in the gene sets without any interaction partners were excluded from the test.

Functional Annotation of Protein Complexes

We downloaded 3,906 biological pathways from ConsensusPathDB release 30 (Kamburov et al., 2013). Over-representation analysis of pathways was tested using a hypergeometric test. In short, all gene sets with at least two candidate genes were tested. The background was restricted to the subset of all genes within the protein interaction network that participate in at least one pathway and similarly, only input genes that were part of the background were included for testing.

Testing for Enrichment of Diabetes-Related GWAS Signal

We further investigated whether the complexes with potential T2D dysregulation were enriched for association with T2D or glycemic traits in 19 different GWA-studies (Supplementary Table 6) using the MAGENTA method (Segrè et al., 2010).

The analysis was repeated using three definitions of complexes: (1) including all genes in the complexes, (2) excluding genes from the complex that had genome-wide significant *P*-values ($P < 5 \times 10^{-8}$) in the respective GWAS, i.e., different genes are excluded from the complexes when testing enrichment in the different GWAS studies, and (3) excluding genes that were used for input in the corresponding gene set. For example, all fasting glucose associated genes were excluded from the complexes when testing for enrichment using "Fasting glucose" and "Fasting glucose, Manning" but not when testing for enrichment of e.g., "Fasting insulin".

In the MAGENTA analysis we used the 95th percentile of all gene *P*-values as the P^{cutoff} and SNPs in GWAS loci were mapped to a gene if they fell within 110 kb upstream or 40 kb downstream of its transcription start and stop sites, respectively.

Statistical Analysis and Visualization

Statistical analyses were performed in the statistical computing language R (R Core Team, 2014) and network visualizations were made in R using the igraph package (Csardi and Nepusz, 2006). Tissue depictions in figures were adapted from Stumvoll et al. (2010).

AUTHOR CONTRIBUTIONS

HP, VG, and SB conceived the study and provided the initial design and data analysis framework. HP and VG performed the analysis and drafted the original manuscript. HP, VG, and SB contributed to the interpretation and corresponding text. All authors approved the version to be published.

FUNDING

The Technical University of Denmark has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115317 (DIRECT), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies in kind contribution. The Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, is supported financially by the Novo Nordisk Foundation (Grant agreement NNF14CC0001).

ACKNOWLEDGMENTS

The authors wish to thank Kirstine GI Belling for critical comments to the manuscript and Jose MG Izarzugaza and Jessica Xin Hu for providing cancer gene lists. The Genotype-Tissue Expression (GTEx) dataset used for the analyses described in this manuscript was obtained from dbGaP at www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v4.p1. Data on glycemic traits has been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00043/full#supplementary-material

Data Availability

The 24 complexes with potential T2D dysregulation are provided in supplementary Data Sheet 2 and further visualized in Data Sheet 1.

REFERENCES

Albrechtsen, A., Grarup, N., Li, Y., Sparsø, T., Tian, G., Cao, H., et al. (2013). Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* 56, 298–310. doi: 10.1007/s00125-012-2756-1

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Barshir, R., Basha, O., Eluk, A., Smoly, I. Y., Lan, A., and Yeger-Lotem, E. (2013). The tissuenet database of human tissue protein-protein interactions. *Nucleic Acids Res.* 41, D841–D844. doi: 10.1093/nar/gks1198

Basha, O., Flom, D., Barshir, R., Smoly, I., Tirman, S., and Yeger-Lotem, E. (2015). MyProteinNet: build up-to-date protein interaction networks for organisms, tissues and user-defined contexts. *Nucleic Acids Res.* 43, W258–W263. doi: 10.1093/nar/gkv515

Bonnefond, A., and Froguel, P. (2015). Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell Metab.* 21, 357–368. doi: 10.1016/j.cmet.2014.12.020

Börnigen, D., Pers, T. H., Thorrez, L., Huttenhower, C., Moreau, Y., and Brunak S. (2013). Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic Acids Res.* 41:e171. doi: 10.1093/nar/gkt661

Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 5, 260. doi: 10.1038/msb.2009.17

Couvelard, A., Hu, J., Steers, G., O'Toole, D., Sauvanet, A., Belghiti, J., et al. (2006). Identification of potential therapeutic targets by gene-expression profiling in pancreatic endocrine tumors. *Gastroenterology* 131, 1597–1610. doi: 10.1053/j.gastro.2006.09.007

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Inter J. Complex Syst.* 1695.

Danielsson, A., Pontén, F., Fagerberg, L., Hallström, B. M., Schwenk, J. M., Uhlén, M., et al. (2014). The human pancreas proteome defined by transcriptomics and antibody-based profiling. *PLoS ONE* 9:e115421. doi: 10.1371/journal.pone.0115421

Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A. H., et al. (2014). Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genet.* 10:e1004160. doi: 10.1371/journal.pgen.1004160

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science* 307, 724–727. doi: 10.1126/science.1105103

Dominguez, V., Raimondi, C., Somanath, S., Bugliani, M., Loder, M. K., Edling, C. E., et al. (2011). Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic β cells. *J. Biol. Chem.* 286, 4216–4225. doi: 10.1074/jbc.M110.200295

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* 42, 105–116. doi: 10.1038/ng.520

Fadista, J., Vikman, P., Ottosson, E., Guerra, I., Lou, J., and Taneera, J. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13924–13929. doi: 10.1073/pnas.1402665111

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B. R., Grarup, N., Burtt, N. P., et al. (2014). Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* 46, 357–363. doi: 10.1038/ng.2915

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783. doi: 10.1093/nar/gkw1121

Ganegoda, G., Wang, J., Wu, F.-X., and Li, M. (2014). Prediction of disease genes using tissue-specified gene-gene network. *BMC Syst. Biol.* 8(Suppl. 3):S3. doi: 10.1186/1752-0509-8-S3-S3

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., et al. (2010). A map of open chromatin in human pancreatic islets. *Nat. Genet.* 42, 255–259. doi: 10.1038/ng.530

Glaser, B. (2013). *Familial Hyperinsulinism*. Seattle, WA: University of Washington. Available online at: http://www.ncbi.nlm.nih.gov/books/NBK1375/

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., et al. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* 10, 1081–1082. doi: 10.1038/nmeth.2642

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. doi: 10.1038/ng.3259

Gross, A. M., and Ideker, T. (2015). Molecular networks in context. *Nat. Biotechnol.* 33, 720–721. doi: 10.1038/nbt.3283

Guan, Y., Gorenshteyn, D., Burmeister, M., Wong, A. K., Schimenti, J. C., Handel, M. A., et al. (2012). Tissue-specific functional networks for

prioritizing phenotype and disease genes. *PLoS Comput. Biol.* 8:e1002694. doi: 10.1371/journal.pcbi.1002694

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi: 10.1038/nature02555

Huyghe, J. R., Jackson, A. U., Fogarty, M. P., Buchkovich, M. L., Stančáková, A., Stringham, H. M., et al. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* 45, 197–201. doi: 10.1038/ng.2507

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* 41, D793–D800. doi: 10.1093/nar/gks1055

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., et al. (2014). A draft map of the human proteome. *Nature* 509, 575–581. doi: 10.1038/nature13302

Kodama, K., Horikoshi, M., Toda, K., Yamada, S., Hara, K., Irie, J., et al. (2012). Expression-based genome-wide association study links the receptor *CD44* in adipose tissue with type 2 diabetes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7049–7054. doi: 10.1073/pnas.1114513109

Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., et al. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* 43, 984–989. doi: 10.1038/ng.921

Lage, K., Greenway, S. C., Rosenfeld, J. A., Wakimoto, H., Gorham, J. M., Segrè, A. V., et al. (2012). Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14035–14040. doi: 10.1073/pnas.1210730109

Lage, K., Hansen, N. T., Karlberg, E. O., Eklund, A. C., Roque, F. S., Donahoe, P. K., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20870–20875. doi: 10.1073/pnas.0810772105

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25, 309–316. doi: 10.1038/nbt1295

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., et al. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* 27, 208–222. doi: 10.1101/gr.212720.116

Li, D. (2012). Diabetes and pancreatic cancer. *Mol. Carcinog.* 51, 64–74. doi: 10.1002/mc.20771

Locke, J. M., Hysenaj, G., Wood, A. R., Weedon, M. N., and Harries, L. W. (2015). Targeted allelic expression profiling in human islets identifies cis-regulatory effects for multiple variants identified by type 2 diabetes genome-wide association studies. *Diabetes* 64, 1484–1491. doi: 10.2337/db14-0957

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Magger, O., Waldman, Y. Y., Ruppin, E., and Sharan, R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.* 8:e1002690. doi: 10.1371/journal.pcbi.1002690

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* 46, 234–244. doi: 10.1038/ng.2897

Mailman, M., Feolo, M., Jin, Y., and Kimura, M. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186. doi: 10.1038/ng1007-1181

Marselli, L., Thorne, J., Dahiya, S., Sgroi, D. C., Sharma, A., Bonner-Weir, S., et al. (2010). Gene expression profiles of Beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. *PLoS ONE* 5:e11499. doi: 10.1371/journal.pone.0011499

McCarthy, M. I. (2010). Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* 363, 2339–2350. doi: 10.1056/NEJMra0906948

McCarthy, M. I. (2015). Genomic medicine at the heart of diabetes management. *Diabetologia* 58, 1725–1729. doi: 10.1007/s00125-015-3588-6

McCulloch, L. J., van de Bunt, M., Braun, M., Frayn, K. N., Clark, A., and Gloyn, A. L. (2011). GLUT2 (*SLC2A2*) is not the principal glucose transporter in human pancreatic beta cells: implications for understanding genetic association signals at this locus. *Mol. Genet. Metab.* 104, 648–653. doi: 10.1016/j.ymgme.2011.08.026

Morán, I., Akerman, I., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., et al. (2012). Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.* 16, 435–448. doi: 10.1016/j.cmet.2012.08.010

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* 44, 981–990. doi: 10.1038/ng.2383

Natalicchio, A., Labarbuta, R., Tortosa, F., Biondi, G., Marrano, N., Peschechera, A., et al. (2013). Exendin-4 protects pancreatic beta cells from palmitate-induced apoptosis by interfering with GPR40 and the MKK4/7 stress kinase signalling pathway. *Diabetologia* 56, 2456–2466. doi: 10.1007/s00125-013-3028-4

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9, 471–472. doi: 10.1038/nmeth.1938

Nica, A. C., Ongen, H., Irminger, J.-C., Bosco, D., Berney, T., Antonarakis, S. E., et al. (2013). Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Res.* 23, 1554–1562. doi: 10.1101/gr.150706.112

Owusu, D., Pan, Y., Xie, C., Harirforoosh, S., and Wang, K.-S. (2017). Polymorphisms in *PDLIM5* gene are associated with alcohol dependence, type 2 diabetes, and hypertension. *J. Psychiatr. Res.* 84, 27–34. doi: 10.1016/j.jpsychires.2016.09.015

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* 46, 136–143. doi: 10.1038/ng.2870

Pers, T. H., Dworzynski, P., Thomas, C. E., Lage, K., and Brunak, S. (2013). MetaRanker 2.0: a web server for prioritization of genetic variation data. *Nucleic Acids Res.* 41, W104–W108. doi: 10.1093/nar/gkt387

Prasad, R., and Groop, L. (2015). Genetics of type 2 diabetespitfalls and possibilities. *Genes* 6, 87–123. doi: 10.3390/genes6010087

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi: 10.1038/nature08454

Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., et al. (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* 44, 991–1005. doi: 10.1038/ng.2385

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and Type 2 diabetes. *Cell Metab.* 24, 593–607. doi: 10.1016/j.cmet.2016.08.020

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., and Altshuler, D. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* 6:58. doi: 10.1371/journal.pgen.1001058

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550. doi: 10.1038/ng.2982

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* 46, 294–298. doi: 10.1038/ng.2882

Strawbridge, R. J., Dupuis, J., Prokopenko, I., Barker, A., Ahlqvist, E., Rybin, D., et al. (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 60, 2624–2634. doi: 10.2337/db11-0415

Stumvoll, M., Goldstein, B. J., and van Haeften, T. W. (2010). Type 2 diabetes: principles of pathogenesis and therapy. *Lancet* 365, 1333–1346. doi: 10.1016/S0140-6736(05)61032-X

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6062–6067. doi: 10.1073/pnas.0400782101

Taneera, J., Lang, S., Sharma, A., Fadista, J., Zhou, Y., Ahlqvist, E., et al. (2012). A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. *Cell Metab.* 16, 122–134. doi: 10.1016/j.cmet.2012.06.006

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., et al. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* 27, 199–204. doi: 10.1038/nbt.1522

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347:394. doi: 10.1126/science.1260419

van Hove, E. C., Hansen, T., Dekker, J. M., Reiling, E., Nijpels, G., Jorgensen, T., et al. (2006). The *HADHSC* gene encoding short-chain L-3-hydroxyacyl-CoA dehydrogenase (SCHAD) and type 2 diabetes susceptibility: the Damage study. *Diabetes* 55, 3193–3196. doi: 10.2337/db06-0414

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., et al. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4:e1000214. doi: 10.1371/journal.pgen.1000214

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* 42, 579–589. doi: 10.1038/ng.609

Volkmar, M., Dedeurwaerder, S., Cunha, D. A., Ndlovu, M. N., Defrance, M., Deplus, R., et al. (2012). DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO J.* 31, 1405–1426. doi: 10.1038/emboj.2011.503

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

Wessel, J., Chu, A. Y., Willems, S. M., Wang, S., Yaghootkar, H., Brody, J. A., et al. (2015). Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* 6, 5897. doi: 10.1038/ncomms6897

Xia, Z., Dickens, M., Raingeaud, J., Davis, R. J., and Greenberg, M. E. (1995). Opposing effects of ERK and JNK-p38 MAP kinases on apoptosis. *Science* 270, 1326–1331.

Xie, T., Chen, M., Gavrilova, O., Lai, E. W., Liu, J., and Weinstein, L. S. (2008). Severe obesity and insulin resistance due to deletion of the maternal Gsα allele is reversed by paternal deletion of the Gsα imprint control region. *Endocrinology* 149, 2443–2450. doi: 10.1210/en.2007-1458

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA Sequencing of single human islet cells reveals Type 2 diabetes genes. *Cell Metab.* 24, 608–615. doi: 10.1016/j.cmet.2016.08.018

Yamanaka, M., Itakura, Y., Inoue, T., Tsuchida, A., Nakagawa, T., Noguchi, H., et al. (2006). Protective effect of brain-derived neurotrophic factor on pancreatic islets in obese diabetic mice. *Metab. Clin. Exp.* 55, 1286–1292. doi: 10.1016/j.metabol.2006.04.017

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120. doi: 10.1038/ng.3390

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 Genetics

ORIGINAL RESEARCH
published: 13 October 2016
doi: 10.3389/fgene.2016.00183

Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans

Benjamin T. Mayne[1,2], *Tina Bianco-Miotto*[1,3], *Sam Buckberry*[4,5], *James Breen*[1,6], *Vicki Clifton*[7], *Cheryl Shoubridge*[1,2] *and Claire T. Roberts*[1,2]*

[1] Robinson Research Institute, University of Adelaide, Adelaide, SA, Australia, [2] Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia, [3] School of Agriculture, Food and Wine, Waite Research Institute, University of Adelaide, Adelaide, SA, Australia, [4] Harry Perkins Institute of Medical Research, The University of Western Australia, Perth, WA, Australia, [5] Plant Energy Biology, Australian Research Council Centre of Excellence, The University of Western Australia, Perth, WA, Australia, [6] Bioinformatics Hub, School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia, [7] Mater Research Institute, University of Queensland, Brisbane, QLD, Australia

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Francesco Russo,
University of Copenhagen, Denmark
Matteo Benelli,
University of Trento, Italy

***Correspondence:**
Claire T. Roberts
claire.roberts@adelaide.edu.au

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

Received: 12 August 2016
Accepted: 27 September 2016
Published: 13 October 2016

Citation:
Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C and Roberts CT (2016) Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. Front. Genet. 7:183. doi: 10.3389/fgene.2016.00183

The severity and prevalence of many diseases are known to differ between the sexes. Organ specific sex-biased gene expression may underpin these and other sexually dimorphic traits. To further our understanding of sex differences in transcriptional regulation, we performed meta-analyses of sex biased gene expression in multiple human tissues. We analyzed 22 publicly available human gene expression microarray data sets including over 2500 samples from 15 different tissues and 9 different organs. Briefly, by using an inverse-variance method we determined the effect size difference of gene expression between males and females. We found the greatest sex differences in gene expression in the brain, specifically in the anterior cingulate cortex, (1818 genes), followed by the heart (375 genes), kidney (224 genes), colon (218 genes), and thyroid (163 genes). More interestingly, we found different parts of the brain with varying numbers and identity of sex-biased genes, indicating that specific cortical regions may influence sexually dimorphic traits. The majority of sex-biased genes in other tissues such as the bladder, liver, lungs, and pancreas were on the sex chromosomes or involved in sex hormone production. On average in each tissue, 32% of autosomal genes that were expressed in a sex-biased fashion contained androgen or estrogen hormone response elements. Interestingly, across all tissues, we found approximately two-thirds of autosomal genes that were sex-biased were not under direct influence of sex hormones. To our knowledge this is the largest analysis of sex-biased gene expression in human tissues to date. We identified many sex-biased genes that were not under the direct influence of sex chromosome genes or sex hormones. These may provide targets for future development of sex-specific treatments for diseases.

Keywords: sex-biased gene expression, meta-analysis, microarray, human, organs

INTRODUCTION

Differences in both disease severity, prevalence, symptoms, and age of onset vary greatly between males and females (Morrow, 2015). For example, cardiovascular disease is one of the leading causes of death, affecting up to 55% of females but only 44% of males in Europe (Möller-Leimkühler, 2007). Sex differences are also evident in the risk factors for cardiovascular disease, such as diabetes which increases the risk for cardiovascular disease 2–3 fold in males but 3–7 fold in females (Eastwood and Doering, 2005). Sex differences have also been identified in the age of onset of brain diseases such as schizophrenia, where males develop symptoms between 18 and 25 years of age whereas females develop symptoms between 25 and 35 years (Ochoa et al., 2012). Moreover, reported atonic seizures in epilepsy are more frequent in males compared to females (6.5 vs. 1.7%; Carlson et al., 2014). These sex differences in diseases may be the result of tissue-specific differential gene expression between males and females. In schizophrenia, genes relating to energy metabolism have been found to have altered expression in the prefrontal cortex of only males (Qin et al., 2016). Therefore, gene expression may have a role in orchestrating sex differences in the prevalence of diseases.

Many studies neglect to account for sample sex in the design and analysis of their experiments (Mogil and Chanda, 2005; Beery and Zucker, 2011). Historically, females have been excluded from biomedical studies, due to the assumption that their hormonal cycles are a confounding factor in experimental manipulations (Zucker and Beery, 2010; Beery and Zucker, 2011). Despite females and males sharing highly similar genomes, there are numerous sex-specific traits in phenotype, physiology, and pathology. Sexually dimorphic traits can be influenced by sex chromosome genes or sex hormones, but may extend beyond these influences. Sex differences may arise through alterations in autosomal gene regulation but the true extent of sex specific differential gene regulation is not fully known. Understanding these differences may dictate that future research should consider sex as a biological confounder (Zucker and Beery, 2010). Sex differences in many traits are often small and require large sample sizes for studies to be sufficiently powered. The substantial increase in the number of large publicly available genomic data sets could assist in determining the true extent of sex-biased gene expression but to date there are no large-scale meta-analyses investigating this in adult human tissues.

Previous studies have reported sex-biased gene expression in the human brain (Vawter et al., 2004; Reinius and Jazin, 2009; Weickert et al., 2009; Kang et al., 2011; Trabzuni et al., 2013), pancreas (Hall et al., 2014), heart (Fermin et al., 2008), and liver (Zhang et al., 2011). Most studies identify sex-biased genes as those located on the sex chromosomes and it is well-known that these are a source of differentially expressed genes between the sexes (Carrel and Willard, 2005). In mammalian, female, somatic cells, one X chromosome is randomly inactivated by a process referred to as X chromosome inactivation (XCI; Carrel and Willard, 2005; Yang et al., 2010). In normal human XX females, up to 15% of genes on the X chromosome escape XCI, unlike the case in mice where very few escape inactivation (Carrel and Willard, 2005; Yang et al., 2010). Escape from XCI results in a number of genes that are expressed more highly in females compared to males. In addition, autosomal genes have also been shown to be sex-biased in human tissues including the brain (Trabzuni et al., 2013), heart (Fermin et al., 2008) and placenta (Buckberry et al., 2014b). Furthermore, sex differences in the brain in diseases such as multiple sclerosis (MS) are related to autosomal genes and are not regulated by sex chromosome genes (Voskuhl and Palaszynski, 2001; Ebers et al., 2004). These studies highlight the importance of investigating sex differences outside the context of reproductive and sex chromosome factors. In order to characterize the true extent of sex-biased gene expression in humans, we performed a large meta-analysis of publicly available microarray data. We limited our analysis to tissue samples from healthy individuals, reducing the possible effect that diseases may have on gene expression. Our analysis revealed consistencies in sex differences that are widespread in a range of human tissues. Furthermore, we have identified sex-biased genes that are disease-related, suggesting possible mechanisms for the associations of sex with an increased risk of certain diseases.

MATERIALS AND METHODS

Data Collection

Data sets were from different microarray platforms and therefore pre-processing was tailored to each platform. Briefly, data from Illumina platforms were pre-processed using Beadarray prior to quantile normalization (Dunning et al., 2007). Data from Affymetrix platforms were pre-processed and quantile normalized using the robust multiarray average (RMA) or GeneChip-RMA (GC-RMA) where appropriate that is implemented in Simpleaffy (Wilson and Miller, 2005). Batch effects in data sets were corrected for using the "combat" function in the SVA package (Leek et al., 2012). Outliers were identified and removed using ArrayQualityMetrics by analysing MA plots (Kauffmann et al., 2009).

Sample Sex Identification

To identify sample sex in each data set we used the massiR Bioconductor package (Buckberry et al., 2014a). This R package uses unsupervised clustering of probes that target Y chromosome genes to identify sample sex. In data sets where sample sex was supplied, we found an agreement in all predicted and supplied sample sex identification.

Differential Gene Expression Analysis

Probes were re-annotated to Ensembl gene identifiers using biomaRt (Durinck et al., 2009). In tissues where only one data set was found to be useable, sex-biased gene expression was determined using the Empirical Bayes methods within limma (Ritchie et al., 2015). For tissues that were present in >1 data set, differential gene expression analysis was performed using the *meta*GEM package (https://spiral.imperial.ac.uk/handle/10044/1/4217) and using the inverse–variance method as previously described (Ramasamy et al., 2008). For each probe, study specific effect sizes were calculated, by determining the mean and standard deviation for each probe which was corrected using Hedges' g (accounts for the number of samples in each dataset). *Z*

statistics were calculated for each gene identifier which was used to calculate a nominal p-value to give a corrected p-value (false discovery rate, FDR).

Androgen and Estrogen Response Elements

To determine which genes contained androgen response elements (AREs), we firstly downloaded the coordinates of AREs from JASPAR (Hu et al., 2010; Mathelier et al., 2014) and determined the positions within the genome in relation to genes and genomic locations. This was performed using the matchGenes function in the bumphunter Bioconductor package (Jaffe et al., 2012) and UCSC hg19 annotation package (BP)[1]. For estrogen response elements (EREs) we used a previous study that lists genes that are targets of ERα (Jin et al., 2004).

Identifying Enriched Transcription Factors

Transcription factor (TF) binding sites within 10 kb upstream/downstream of sex-biased genes were analyzed using oPOSSUM-3 and the JASPAR vertebrate core profiles (Kwon et al., 2012; Mathelier et al., 2014). We chose 10 kb upstream/downstream of genes as this was the largest range the oPOSSUM-3 would allow. Thus, we sought to identify all possible TF binding sites enriched within sex-biased genes. For each sex-biased gene in each tissue, the TF binding site motifs were searched with a conservation cut-off of 0.4, an 85% threshold for the matrix score and minimum specificity of 8 bits. The resulting TF analysis was limited to the most enriched TFs which were defined as those with the highest Fisher's exact test and z-score rankings.

Gene Ontology

Gene ontology (GO) analysis was performed using all human genes in the Database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.7 (Huang da et al., 2009) and g:Profiler (Reimand et al., 2016). GO terms were considered significant if the corrected p-value (FDR) < 0.05.

A more detailed account of the methodology is provided in **File S1**.

RESULTS AND DISCUSSION

Overview of Publicly Available Microarray Data

Using the Gene Expression Omnibus (GEO; Barrett et al., 2013) and ArrayExpress (Brazma et al., 2003) we identified 22 microarray data sets containing a total of 2502 samples, in 15 different human tissues (**Table 1**). We excluded pooled samples and limited our analyses to data sets with >10 samples to allow better determination of sample sex. To increase the number of useable data sets we used massiR (Buckberry et al., 2014a) to identify and to verify the sample sex in all data sets. From the 22 chosen studies, 10 had sample sex metadata and within these we found concordance with all the predicted and supplied sample sex information. Female samples ($N = 803$) made up 32% of all samples across all data sets (**Table 1**).

Sex differences in autosomal gene expression are typically small so in order to increase statistical robustness, we performed multiple testing corrections in three different analyses for each tissue. We determined the adjusted p-value implemented by Benjamini and Hochberg (1995) for each autosomal gene, where (1) all the chromosomes were included, (2) the Y chromosome was excluded, and (3) both the X and Y chromosomes were excluded in the analysis (**Table 2**). In general, we observed a reduction in the number of autosomal genes that were significantly sex-biased when we removed sex chromosomes from the analysis. Since most genes located on the sex chromosomes had the smallest adjusted p-value, their removal from the analysis slightly increased the adjusted p-value for all other genes. Here we supply the adjusted p-values for all three analyses (**Tables S1– S3**) but discuss only autosomal genes that were significantly different in all three cases. Furthermore, the sample size in each tissue was not reflective of the total number of genes differentially expressed between males and females (**Figure 1**). For example, despite the frontal lobe of the cerebral cortex or frontal cortex (FC) and cerebellum (CB) data sets containing the greatest number of samples, with 455 and 553 samples, respectively, we detected only a small number of sex-biased genes compared to other tissues such as the anterior cingulate cortex (AnCg) and the heart which contained the greatest number of sex-biased genes with average sample sizes (**Figure 1**, **Table 2**).

Sex-Biased Gene Expression in the Human Brain

Previous studies have found sex-biased gene expression in the human brain (Vawter et al., 2004; Reinius and Jazin, 2009; Weickert et al., 2009; Kang et al., 2011). We identified five data sets for seven brain regions and our analyses showed that each region had different numbers of differentially expressed genes (**Tables 1, 2**). Our findings were consistent with previous studies (Reinius and Jazin, 2009; Weickert et al., 2009; Kang et al., 2011), whereby the most striking differences in gene expression between the sexes were sex chromosome genes. These comprised most of the sex-biased genes in the amygdala (65%; AMY) and FC (78%). However, a large proportion of sex-biased genes were autosomal in the nucleus accumbens (91%; NC), AnCg (95%), dorsolateral prefrontal cortex (91%; DLPFC), CB (60%) and the hippocampus (89%; HC). Of the 1690 autosomal sex-biased genes in AnCg, 65% were expressed more highly in males (**Figure 2A**, **Tables S1–S3**). Conversely, we observed a greater proportion of autosomal genes expressed more highly in females in the NC (75%), DLPFC (68%), and the HC (62%). We also found that each brain region was unique in its proportion of sex-biased genes, with as many sex-biased genes in one brain region that were not sex-biased in another (**Figure 2B**).

An increase in the expression of heat shock proteins (HSPs) has been shown to have protective roles in pro-inflammatory responses (Grundtman et al., 2011). Consistent with a previous study (Lin et al., 2011), we found genes that encode for HSPs

[1]BP., C. M. a. M., TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.2.

TABLE 1 | Gene expression data involving 15 healthy tissues.

Organ/tissue	GEO accession	Microarray manufacturer	Samples in data set	Control samples	Sample after pre-processing	Males	Females
Bladder	GSE13507	Affymetrix	256	68	68	48	20
Brain	GSE45642	Affymetrix	670	670	659	493	166
Brain	GSE11512	Affymetrix	80	44	44	29	15
Brain	GSE54572	Affymetrix	24	12	12	5	7
Brain	GSE36192	Illumina	911	911	911	622	289
Brain	GSE44456	Affymetrix	39	39	39	28	11
Colon	GSE8671	Affymetrix	62	25	23	15	8
Colon	GSE41328	Affymetrix	20	10	10	8	2
Heart	GSE55231	Illumina	129	129	118	69	49
Heart	GSE26887	Affymetrix	24	24	23	19	4
Heart	GSE57338	Affymetrix	313	136	136	97	39
Kidney	GSE43974	Illumina	554	118	118	73	45
Kidney	GSE50892	Affymetrix	17	17	15	9	6
Liver	GSE61276	Illumina	106	50	48	22	26
Liver	GSE23649	Illumina	69	69	68	42	26
Liver	GSE38941	Affymetrix	27	10	10	4	6
Lung	GSE10072	Affymetrix	107	49	46	32	14
Lung	GSE18995	Affymetrix	35	35	34	15	19
Lung	GSE51024	Affymetrix	96	41	39	34	5
Pancreas	GSE15471	Affymetrix	78	36	35	19	16
Thyroid	GSE33630	Affymetrix	105	45	35	10	25
Thyroid	GSE65144	Affymetrix	25	13	12	7	5
		Total	3747	2551	2502	1699	803

Each row corresponds to a data set where only healthy tissue was used within this analysis. The columns report the Microarray manufacturer, total number of samples, and which data sets supplied sample sex.

TABLE 2 | Total number of sex-biased genes in each tissue.

Organ/tissue	No. of sex-biased genes (All chromosomes)	No. of autosomal sex-biased genes (Sex chromosomes included in analysis)	No. of autosomal sex-biased genes (Sex chromosomes removed)	No. of autosomal sex-biased genes (Y chromosome removed)
Bladder	16	0	0	0
Brain (Nucleus Accumbens)	264	239	216	244
Brain (Amygdala)	17	6	0	0
Brain (Cerebellum)	98	59	45	52
Brain (Anterior Cingulate Cortex)	1818	1726	1690	1728
Brain (Dorsolateral Prefrontal Cortex)	198	180	165	169
Brain (Frontal Cortex)	45	10	27	7
Brain (Hippocampus)	205	183	174	180
Colon	218	199	162	190
Heart	375	348	334	346
Kidney	224	196	194	194
Liver	32	21	16	28
Lung	36	14	2	12
Pancreas	22	0	0	0
Thyroid	163	151	133	135

Each column corresponds to the total number of genes that were differentially expressed between males and females in each analysis.

to have sex-biased expression in the human brain. Our analyses also identified genes that are involved in pro-inflammatory responses, such as those encoding interleukins, that are more highly expressed in females in NC, AnCg, DLPFC, and HC tissues (**Tables S1**–**S3**). By contrast, genes expressed more highly in males within the brain were related to energy production

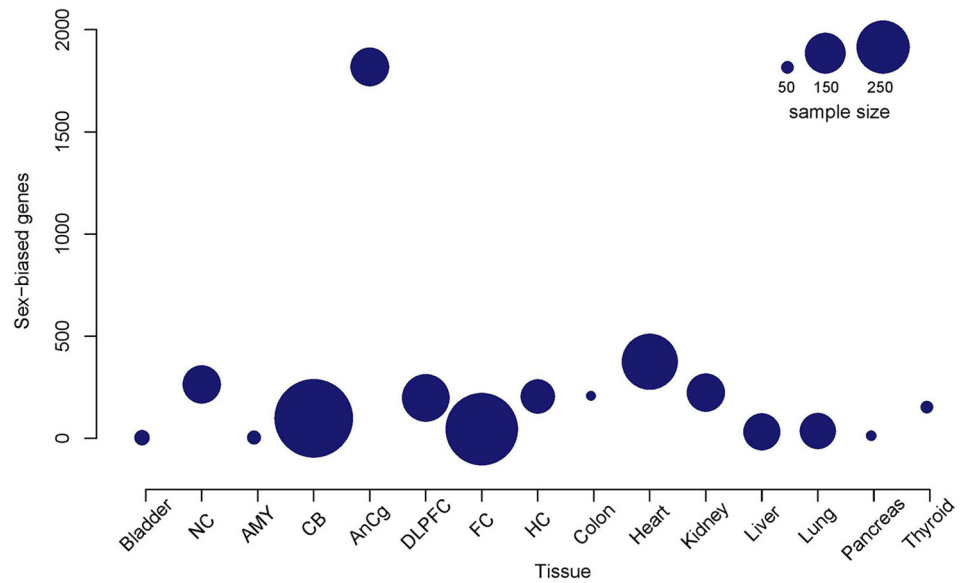


FIGURE 1 | Total number of detectable sex-biased genes relative to the sample size in each tissue. A bubble plot of each tissue where the size of the bubble is proportional to the sample size of the tissue. Bubbles that are higher on the y-axis are tissues that demonstrate a higher number of detectable sex-biased genes. Nucleus accumbens (NC); amygdala (AMY); cerebellum (CB); anterior cingulate cortex (AnCg); dorsolateral frontal cortex (DLPFC); frontal cortex (FC); hippocampus (HC).

and growth, including ATPase's and insulin-like growth factors in the HC and NC, respectively, and *GAPDH* in the AnCg. We found sex-biased genes in the NC, AnCg, and HC to be enriched for GO as defined by DAVID v6.7 for terms relating to cellular functions, the immune response and energy production (**Figures 2C–E**, **Table S4**). We also used g:Profiler (Reimand et al., 2016) for a comparison of GO terms and found similar results to what was found by DAVID v6.7. For example, in the NC, AnCg, and HC we found that the gene upregulated in females were enriched for those involved in the immune response (GO:0006955). Whereas, genes upregulated in males were found to be enriched for GO terms such as generation of precursor metabolites and energy (GO:0006091). Overall, varying proportions and types of sex-biased genes were identified within different locations of the brain, suggesting that specific cortical regions may influence sexually dimorphic traits. As mentioned above, the AnCg contained the largest number of genes differentially expressed between males and females. The AnCg is one of the most recently evolved parts of the mammalian brain (Allman et al., 2001) and also has been shown to regulate behavior and act in a sex-specific manner (Liu et al., 2012). Furthermore, previous studies have identified sex differences in mood disorders and the AnCg is known to have a role in regulating mood (Seney and Sibille, 2014; Yang et al., 2015). In mice, the AnCg has also been shown to have a critical role in sexual interest of males for females (Wu et al., 2009) and hence the large number of genes that were differentially expressed between sexes in the AnCg may assist in the explanation for sexual dimorphism in behavior.

Sex biased gene expression in the brain may potentially contribute to differences in certain neurological diseases between sexes, such as the previously mentioned epilepsy. Sex differences in gene expression may mediate these differences in susceptibility or comprise part of the mechanistic pathways involved in their pathology. Previously, sex biased gene expression in the brain has been proposed to underlie the sex differences in schizophrenia (Trabzuni et al., 2013) which has an incidence of 1.4:1 between males and females (Abel et al., 2010). We found several genes that have been associated with brain disorders to be sex-biased within specific locations of the brain. For example in the AnCg, *NOTCH3*, a gene associated with hereditary stroke disorder (Joutel et al., 1996), and *ALDH3B1*, a gene associated with schizophrenia (Wang et al., 2009), were more highly expressed in females than males. On the other hand, *KCNH3*, a gene associated with epilepsy (Zhang et al., 2010), *GABRB3*, a gene associated with schizophrenia (Huang et al., 2014), epilepsy (Gurba et al., 2012), and autism (Buxbaum et al., 2002), *SNCA*, a gene associated with Parkinson's disease (Wang et al., 2015), and *RGS4*, a gene associated with schizophrenia (Jönsson et al., 2012), were all expressed more highly in males. Recently, sex-biased gene expression has also been identified during developmental stages of the human brain (Shi et al., 2016). Furthermore, genes associated with schizophrenia have been found to be upregulated in male brains as opposed to females across different developmental stages (Shi et al., 2016). This demonstrates consistency in sex-biased genes within the human brain across different studies. Taken together, these findings suggest possible mechanisms by which sex-specific prevalence of brain disorders may occur.

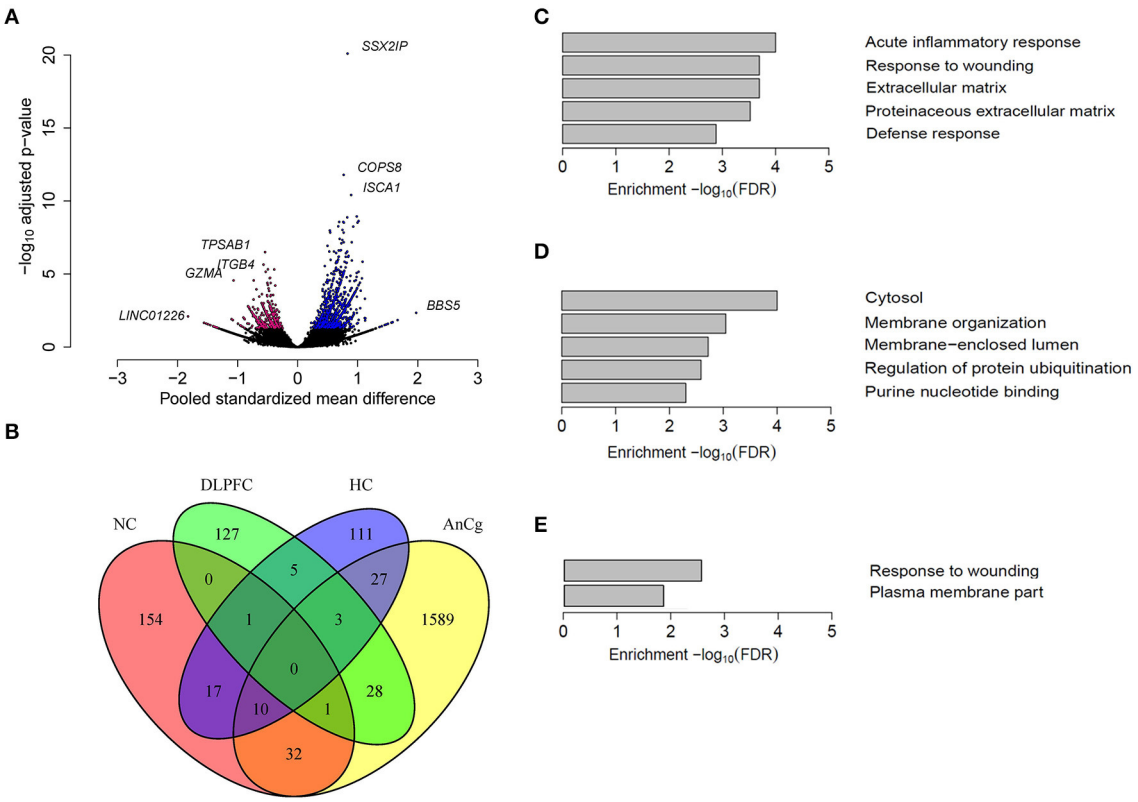


FIGURE 2 | Sex differences in autosomal gene expression in the human brain. (A) A volcano plot representing the autosomal genes that were sex-biased in the AnCg. Pink colored dots represent genes that were significantly expressed more highly in females and blue colored dots represent genes that were expressed more highly in males. **(B)** A four-way venn diagram showing the overlap of sex-biased autosomal gene expression in different regions of the human brain. Most genes that were found to be sex-biased in one region were not sex-biased in another region. The top GO terms that were enriched for sex-biased genes in **(C)** Nucleus accumbens (NC), **(D)** anterior cingulate cortex (AnCg) and **(E)** hippocampus (HC).

The Heart and Kidney Show Opposite Trends in Sex Differences in Gene Expression

Most of the heart gene expression data used in this study are from individuals with an average age of 47 years and we observed many sex differences in expression of genes associated with heart disease. It has been reported in elderly individuals (>75 years), isolated systolic hypertension can be up to 14% more prevalent in females than males (Maas and Appelman, 2010). We found *SCN10A*, a gene associated with hypertrophic cardiomyopathy (Iio et al., 2015), and *KCNE1,* a gene associated with long-QT syndrome (Splawski et al., 2000), to be expressed more highly in hearts from females. Interestingly, 62% of the 334 autosomal sex-biased genes in the heart were expressed more highly in females. The distribution of sex-biased genes across all chromosomes in the heart was similar to that in a previous study (Fermin et al., 2008). However, we report a much smaller number of sex-biased genes in the heart [375 genes in 277 samples (**Table 2**, **Table S1**) compared to 1800 genes in 102 samples in that study (Fermin et al., 2008)].

Conversely, compared to the heart, we found an opposite trend in the kidneys, with 72% of a total of 194 autosomal genes being expressed more highly in males. We also identified six genes located on chromosome 1 that were expressed more highly in females in the heart that were more abundantly expressed in males in the kidney (**Figure 3**). These genes are from the RNA U1 family (*RNU1-1*, *RNU1-2*, *RNU1-3*, *RNU1-4*, *RNVU1-7*, and *RNU1-18*) that includes genes that regulate transcription, elongation and pre-mRNA splicing events (O'Reilly et al., 2013; Guiro and O'Reilly, 2015). It has been suggested that the expression of these genes is different between tissues to regulate organ specific alternative splicing events (Guiro and O'Reilly, 2015). Sex differences in alternative splicing have also previously been detected in the brain, where it has been found to affect 2.5% of expressed genes (Trabzuni et al., 2013). Apart from RNA U1 family all other sex-biased genes were only found to be expressed more highly in one sex.

Sex Hormones and Gene Expression

Many of the sex-biased genes we identified encode enzymes that are known to regulate the production of sex hormones. In the AnCg, three genes from the sulfotransferase family that regulates sulfate conjugation in estrogen precursors (Adjei et al., 2003; *SULT2A1*, *SULT1B1*, and *SULT1C1*) were expressed more highly

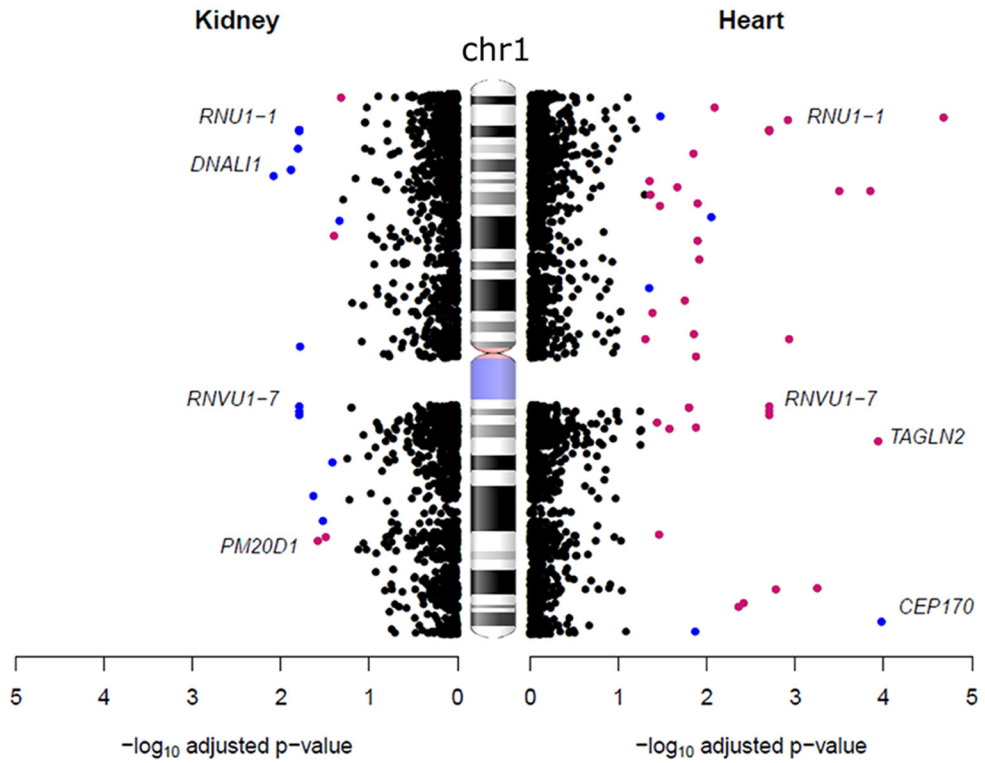


FIGURE 3 | Sex-biased gene expression differences on chromosome 1 in the heart and kidney. Each dot represents a gene, blue dots are genes that were expressed more highly in males and pink dots are those expressed more highly in females. The ideogram of chromosome 1 was obtained from the (National Centre for Biotechnology Information, 2016) (NCBI).

in females. In addition, we also found *STS* [a gene involved in the production of estrogen precursors (Miki et al., 2002)] to be expressed more highly in females in the FC and CB, as well as in the heart and lung. We did not find any major sex differences in gene expression in the bladder, liver, lung, or pancreas, apart from genes located on the sex chromosomes and those that are involved in sex hormone production. This can be contradictory to that which has been found in mouse studies where thousands of genes have been found to be sex-biased (Yang et al., 2006; van Nas et al., 2009). This may reflect an evolutionary difference between the species. Apart from the brain, we found the largest number of sex-biased gene expression differences in the heart, kidney, colon, and thyroid (**Table 2**). Thyroid hormones are known to regulate sex hormone-binding globulin (SHBG) production, which transports androgens and estrogens through the bloodstream (Selva and Hammond, 2009). In the thyroid, 133 autosomal genes were sex-biased, 75% of which were expressed more highly in males. Genes that encode for growth factors and signaling molecules were highly expressed in the thyroid of males, such as *CCL28*, a growth factor in hematopoietic stem cells (Karlsson et al., 2013), *CMTM4*, a chemokine that regulates the cell cycle (Plate et al., 2010), and *GH1*, a gene that encodes for growth hormone (Vakili et al., 2014). These findings suggest a functional role for the thyroid in influencing sexually dimorphic traits such as metabolism, as well as sex differences in thyroid hormone secretion (Ehrenkranz et al., 2015). There is also evidence to suggest that thyroid hormones significantly influence testosterone levels (Meikle, 2004).

To determine if the differentially expressed genes between sexes were regulated by sex hormones, we quantified the number of genes that contained either AREs or EREs. For AREs we downloaded the coordinates of AR binding sites from the JASPAR database (Hu et al., 2010; Mathelier et al., 2014) and for EREs we used a list of previously reported ERα targets (Jin et al., 2004). In total, we identified 3014 different genes that were expressed more highly in either sex in at least one tissue. Of the 3014 genes, 875 contained AREs, 239 contained EREs and 86 contained both. On average 32% of autosomal genes that were sex biased in tissues contained AREs or EREs. Therefore, across all tissues analyzed approximately two-thirds of autosomal genes did not contain either AREs or EREs. Four hundred and eighty-nine genes contained AREs within gene bodies such as introns and exons, 216 genes contained AREs upstream and within the promoters, and 170 genes contained AREs located downstream of the gene. The precise locations of EREs were unknown as we were using a list of previously defined ERα targets. GO enrichment for genes that contained both AREs and EREs in each individual tissue did not produce any significant enrichment, most likely due to the lists of genes being too small. We therefore found it advantageous to combine the list of genes

across different tissues since the list of genes in each tissue were too small to produce any significant results. The genes that contained either or both AREs or EREs and were expressed more highly in females were enriched for GO terms relating to response to wounding and inflammatory response. For example, we found genes related to interleukin signaling and inflammatory processes to be expressed more highly in females such as *TNFAIP6*, *IL10RB*, and *IFNA2* in the DLPFC, HC, and AnCg, respectively. On the other hand genes containing either or both AREs or EREs that were expressed more highly in males were enriched for GO terms relating to mitochondrion and generation of precursor metabolites and energy. As already mentioned, we found a variety of ATPase's to be expressed more highly in males in the AnCg, NC, DLPFC, CB, thyroid, colon, and kidney such as *ATP5G1*, *ATP6V1B2*, *ATP6V0B*, *ATP6V1C1*, and *ATP6V1A*. These results indicate that sex chromosome genes and sex hormones are key regulators of sex-biased gene expression across a range of tissues. However, our data also suggest a significant number of genes that have sex-biased expression may potentially be independent of direct influence by sex chromosomes or sex hormones.

Sex-Biased Epigenetic Modifications

Genes that are involved in the regulation of transcription and histone modifications also showed sex differences. In the colon, genes expressed more highly in males included those that encode for histones (*H3F3A*, *H3F3AP4*, *H3F3AP6*, and *H3F3BP1*) and ribosomal proteins (*RPS3A*, *RPS3AP26*, *RPS3AP6*, *RPL13A*, *RPL4*, *RPL4P4*, *RPL13AP5*, *RPS3AP5*, *RPS3AP47*, *RPL7A*, *RPL7AP6*, *RPL23AP74*, *RPL4P5*, *RPL3P4*, *RPL13AP20*, and *RPL13AP25*). These genes were also expressed more highly in males in other tissues such as the brain, heart, and kidney. It is worth mentioning that we also found other members of the RPL gene family to be more highly expressed in females in other tissues (**Tables S1**–**S3**). We also found sex bias in some genes that encode for enzymes that regulate histone modifications. For example, *SET*, a gene that inhibits nucleosome and histone H4 acetylation (Krajewski and Vassiliev, 2011) was expressed more highly in males in the DLPFC, *SMYD3*, a histone methyltransferase (Hamamoto et al., 2004), *PRMT2*, *PRMT5*, and *PRMT8* [histone arginine methyltransferases (Di Lorenzo and Bedford, 2011)] were more highly expressed in males in the AnCg and DLPFC (**Tables S1**–**S3**). Together these findings suggest that sex differences in tissue-specific gene expression extend from sex hormones and into genes that regulate gene expression and translation. Furthermore, our findings of sex bias in genes that encode for histones and histone modifying enzymes in most tissues suggest the possibility that sex-specific epigenetic modifications act on transcription that may result in phenotypic sex differences.

X-Linked Sex-Biased Gene Expression

As expected, a majority of X-linked, sex-biased genes were expressed more highly in females (**Figure 4**), with the exception of those in the AnCg in which 75% were more abundantly expressed in males. The mechanism by which genes on the single copy X chromosome in males could be expressed more highly than in females with two copies is obviously likely to be associated with XCI but another mechanism is likely to be active and requires investigation. Although we do report Y chromosome genes in our analysis (**Table 2**, **Table S1**), we do not consider these genes as differentially expressed between sexes, since females do not have a Y chromosome. We do, however, consider the reported Y chromosome genes as detectable in the analyzed tissues and act as a positive control and these genes may have potential roles in the male phenotype in these tissues. Many X-linked genes that were expressed more highly in females have been previously reported to escape XCI (Cotton et al., 2015). Not surprisingly, we consistently found *XIST* and *JPX* [genes that orchestrate XCI (Augui et al., 2011; Lee, 2011)] to be expressed more highly in females and interestingly, many sex-biased X-linked genes that are known to regulate gene expression have been defined previously (Bellott et al., 2014). For example, we found *KDM6A* (**Figure 5**), a gene that regulates chromatin modifications, to be expressed more highly in females in the liver, lung, DLPFC, NC, AMY, FC, bladder, and CB. In addition, forest plots (**Figure 5**) demonstrate consistency between individual data sets of *KDM6A* expression showing higher expression in females across different tissues. Furthermore, we also found *KDM5C* to be expressed more highly in females in the lung, FC, bladder, and CB. Genes that are involved in post-transcriptional processes and more highly expressed in females in the liver, thyroid, FC, and CB, include *ZRSR2*, *DDX3X* which are involved in alternative splicing. In addition, we also found translation regulators *EIF1AX* and *RPS4X*, to be expressed more highly in females in the lung, pancreas, HC, and colon.

Across all tissues, we found a total of 86 different genes on the X chromosome to be more highly expressed in males in at least one tissue. Twenty-two of the 86 X chromosome genes more highly expressed in males have homologous counterparts on the Y chromosome and are located within pseudoautosomal region 1 (PAR1; Ross et al., 2005), which may explain the differences in expression. However, not all X chromosome genes that were expressed more highly in males were within PAR1 or had homologous Y chromosome counterparts, such as *SMARCA2*, an ATPase and chromatin re-modeler (Takeshima et al., 2015). These findings suggest that X-linked sex-biased genes may potentially regulate autosomal gene expression such as the possible case of *SMARCA2*, through epigenetic modifications and post-transcriptional processes.

Enriched Transcription Factors

We next investigated which TFs were enriched in the sex-biased genes by running a TF binding site (TFBS) enrichment analysis using oPOSSUM-3 and the JASPAR core motifs (Kwon et al., 2012; Mathelier et al., 2014). Both the Sry-related HMG box (SOX) and the Forkhead-box (FOX) family of TFs were enriched within 10 kb of the transcription start site (TSS) of sex-biased genes across all tissues (**Table S5**). The SOX TFs are vital for sex determination (Huang et al., 2015) and the FOX TFs are essential for embryonic development and also have roles in regulating the immune system (Coffer and Burgering, 2004; Jackson et al., 2010; Lam et al., 2013). Sex chromosome derived TFs such as *ZFX* and *SRY* were also enriched within 10 kb of the TSS. We also found the androgen receptor (*AR*) as an enriched TF within the AMY,

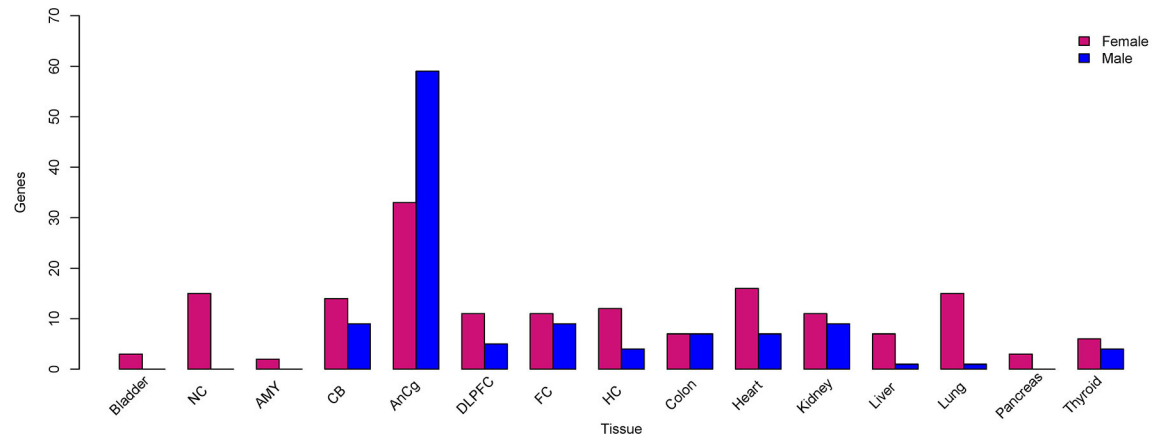


FIGURE 4 | X-linked sex-biased gene expression. The total number of genes located on the X chromosome that were expressed more highly in females (pink) and males (blue) compared to the opposite sex, respectively.

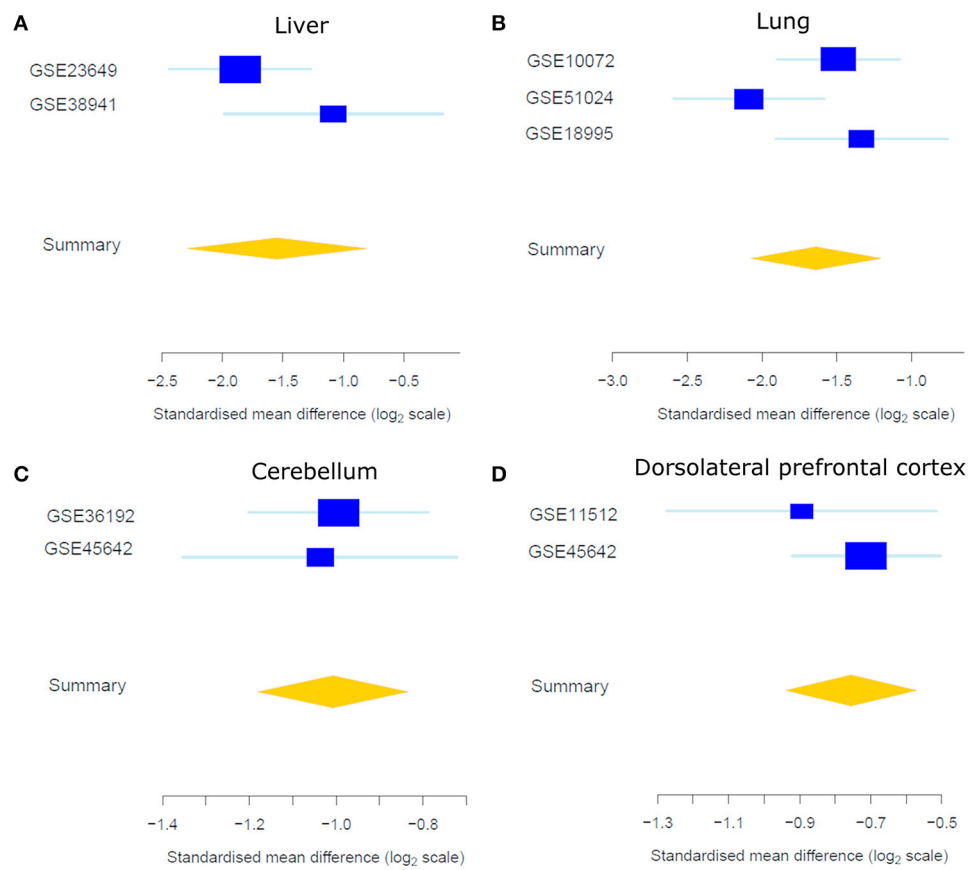


FIGURE 5 | Forest plots of the standardized mean difference of *KDM6A* expression showing higher expression in females in the liver (A), lung (B), CB (C), and DLPFC (D). Each blue box is representative of the study size in each data set and horizontal lines are standard error. The yellow diamond represents the overall gene summary for *KDM6A* in each tissue.

CB, FC, bladder, and lung. In addition, *HNF1A* and *HNF1B* were enriched in genes upregulated in both males and females within all tissues apart from the NC and DLPFC. *HNF1A* and *HNF1B* are homeobox TFs that are required for expression of specific liver genes (Shih et al., 2001). These findings reveal TFs that may have important roles in regulating sexually dimorphic gene expression such as *HNF1A* and *HNF1B*, which as mentioned earlier have only previously been shown to be required for expression of

specific liver genes (Shih et al., 2001). However, the genes that encode for the majority of the TFs that were enriched within sex-biased genes were not themselves differentially expressed between the sexes. Although in this study, we focus on gene expression, TFs undergo more processing post-transcription and therefore their protein abundance within tissues may differ between sexes.

Sex Differences in Other Tissues

In this study, we have analyzed sex-biased gene expression in 15 human tissues. However, we must acknowledge other studies that have also analyzed sex-biased gene expression. One of the largest studies that has analyzed sex-biased gene expression is the Genotype-Tissue Expression (GTEx) project (Melé et al., 2015). The GTEx project has used RNA-seq to analyse gene expression in a variety of different human tissues which would give a broader comparison of gene expression differences between tissues. In comparison to GTEx (Melé et al., 2015) we have analyzed sex-biased gene expression in five of the same human tissues which is represented as a Venn diagram (**Figure S1**). We found an overlap of sex chromosome genes as being sex-biased between this study and GTEx. However, there were many genes that we found to be sex-biased that were not in GTEx (Melé et al., 2015). A possible explanation for the difference between studies is that in GTEx only samples from175 individuals were used (Melé et al., 2015) as opposed to over 2500 in this study which provides much greater statistical power compared to GTEx (Melé et al., 2015). In addition, GTEx also used RNA-seq and were therefore able to quantify the expression of genes for which no probes were available in the microarrays used in this study.

Bias of Male Samples

To prevent any biases in our analyses we have performed differential gene expression in tissues from all publicly available data to our knowledge. However, since most studies neglect to account for samples sex (Mogil and Chanda, 2005; Beery and Zucker, 2011), we unfortunately had a ratio of 2.1:1 males to females on average across all tissues analyzed. Therefore, this in itself may create some biases in our analyses. Across all data sets (**Table 1**) the ratio of males to females was skewed toward males apart from one data set containing thyroid samples (GSE33630), where the ratio was 2.5 females for every male.

To determine if the ratio of males to females affects the differential expression analyses we conducted a 10-fold cross validation of the differential gene expression analyses in the tissue where the ratio of males to females was the greatest. The AMY gene expression data had a ratio of 4.5 males to every female. In this analysis we randomly removed male samples from the analysis to make the number of each sex the same and then assessed which genes were differentially expressed between males and females. We performed this analysis 10 times and then compared which genes were consistently identified as sex-biased to our original analysis where we did not sub-set any male samples. In the analysis with the sex chromosomes included we found the sex chromosome genes (*XIST*, *RPS4Y1*, *DDX3Y*, *KDM5D*, *USP9Y*, *EIF1AY*, and *TTTY15*) consistently classified as sex-biased in the 10-fold cross validation. However, in the original analysis we identified four autosomal genes to be sex-biased and upregulated in females (**Table S1**). However, these four autosomal genes were not found to be sex-biased in the 10-fold cross validation. By performing the 10-fold cross validation, we removed samples which would have decreased our statistical power and therefore increased the magnitude of the adjusted *p*-value which is what occurred. Therefore, caution should be taken when interpreting the results of genes that were found to be sex-biased with an adjusted *p*-value close to 0.05 and in tissues where there is a large ratio of males to females. However, this analysis does provide reassurance that the sex chromosome genes that were found to be sex-biased in the original analysis were not greatly affected by the bias in male samples.

STRENGTHS AND LIMITATIONS

While our analyses reveal many sex differences in gene expression within a variety of tissues, there are several limitations to this study. Firstly, most tissues (where age was provided) were from individuals who were post-reproductive age (average age = 47 years) which may not have captured the true extent of sex-biased gene expression that would otherwise be evident during early adulthood when sex hormones are at their peak production. Thus, using data from older individuals limited our ability to assess sex-biased gene expression in individuals of reproductive age. We also report a number of genes previously associated with diseases and disorders that were differentially expressed between sexes. RNA expression differences do not necessarily cause phenotypic variation, as there are multiple levels of gene and protein regulation that can occur post-transcription. Next-generation sequencing, as opposed to microarrays used in this study, would allow a more complete assessment of sex-dependent gene expression differences but there is currently more samples that have been analyzed using microarrays and therefore more statistical power can be achieved. Furthermore, on average, 64% of genes differentially expressed between sexes in each tissue had a magnitude $\log_2 FC < 1$. Most genes that were found to be sex-biased do not have large $\log_2 FC$ apart from genes located on the sex chromosomes. In addition, most genes that were found to be sex-biased across all tissues had a magnitude $\log_2 FC < 1.5$ (**Table S7**). Therefore, future studies would need to be adequately powered to replicate our findings. Despite these limitations, to our knowledge this is the largest analysis of sex differences in gene expression across a range of human tissues.

Despite the large amount of genomic data that was available for this study it was unfortunate not to consider clinical and lifestyle factors such as age, smoking status, sample heterogeneity and body mass index (BMI) which may potentially have an effect on gene expression. We were unable to correct for these potential confounding factors because, as detailed in **Table S6**, most studies provide little or no clinical information about the samples. Furthermore, only 32% of all the samples analyzed in this study were from females which may potentially create a bias for genes to be more highly expressed in males. However, by acknowledging this limitation we draw attention to the bias toward using only males in biomedical research. We therefore

urge future research in all fields of biomedical science to use an equal sex ratio in study design.

CONCLUSIONS

Our analyses have revealed substantial differences in the transcriptional landscape between sexes across a range of human organs and tissues and highlight possible mechanisms by which gene expression may contribute to sexually dimorphic traits. Improved understanding of these is fundamental to understanding diseases with different prevalence between the sexes. Our data show that sex differences in gene expression vary widely across different tissues. We identified a consistent trend for genes known to regulate the immune system to be more highly expressed in females and those involved in energy production and growth were more highly expressed in males. These may be the result of different evolutionary pressures between the sexes. The brain demonstrates the largest differences in sex-biased gene expression with several sex-biased genes associated with specific brain disorders, providing insight into possible mechanisms for the association of sex-specific prevalence of certain brain disorders.

Our findings also indicate that many sex biased genes within tissues are independent of sex chromosome genes or sex hormones. Approximately 32% of autosomal genes in each tissue contained an ARE or ERE, which suggests there are other mechanisms that underpin sex differences in gene expression. One potential mechanism is through epigenetic factors, such as chromatin modeling which has been suggested to have sex specific functional roles (Silkaitis and Lemos, 2014).

Finally, our data demonstrate why it is important to consider sex as a biological confounder in biomedical studies. Future studies should incorporate sex differences in their analyses which will help to provide new insights in health and disease. The sex-biased genes identified in this study provide a basis for determining the mechanism by which sexual dimorphism occurs and potential causal pathways for sexually biased disease susceptibility. More importantly however, they provide potential targets for novel sex specific treatments.

AUTHORS CONTRIBUTIONS

BM designed, conducted the study, analyzed and interpreted the data, and wrote the manuscript. SB conceived the initial part of the study and provided intellectual input into the manuscript. JB, TB, and CR were all involved in the study design, provided critical discussion and intellectual input into the manuscript. CS and VC provided critical discussion and intellectual input into the manuscript. All authors read and approved the final manuscript.

FUNDING

This project was funded in part by a National Health and Medical Research Council of Australia (NHMRC) Project Grant (GNT1059120) awarded to CR, CS, VC, and TB. CR is supported by a NHMRC Senior Research Fellowship GNT1020749. CS is supported by an Australian Research Council Future Fellowship (FT120100086). VC is supported by a NHMRC Senior Research Fellowship GNT1041918. SB is supported by an NHMRC-ARC Dementia Research Development Fellowship Grant (APP1111206). BM is supported by an Australian Post-graduate Award.

ACKNOWLEDGMENTS

The authors would like to thank the generosity of all individuals who were involved in the data creation of all data sets that were available for public analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00183

Table S1 | Differential gene expression analysis between males and females in each tissue, including the sex chromosomes. A list of differentially expressed genes between sexes in each tissue with all the chromosomes included in the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Table S2 | Differential gene expression analysis results with genes on the Y chromosome removed from the analysis. A list of sex-biased genes with the Y chromosome genes removed from the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Table S3 | Differential gene expression analysis results with genes on the X and Y chromosomes removed from the analysis. A list of sex-biased genes with the sex chromosome genes removed from the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Table S4 | Gene ontology results. This table lists all the GO terms that were found to be enriched within each tissue. Only significant GO terms were found for the NC, AnCg, and HC.

Table S5 | Transcription factors that were found to contain enriched motifs with 10 kb of the transcription start site of sex-biased genes in each tissue. A list of enriched transcription factors of the sex-biased genes in each tissue.

Table S6 | Clinical and lifestyle factors supplied by each data set. A table representing which data set supplied sample information such as age, ethnicity, sex, smoking status, and disease status.

Table S7 | Total number of sex-biased genes at different log_2FC cut-offs. A table listing the total number of genes that were found to be sex-biased in each tissue at different log_2FC cut-offs. This analysis was performed with the sex chromosomes included.

Figure S1 | Venn diagrams representing the overlap of defined sex-biased genes between this study and a previous study (Melé et al., 2015). Each Venn diagram represents an individual tissue and the overlap of genes that were found to be sex-biased between studies.

File S1 | Detailed methodology. A description of the precise methods used involved in data collection, data processing, normalization, batch correction, and differential expression.

REFERENCES

Abel, K. M., Drake, R., and Goldstein, J. M. (2010). Sex differences in schizophrenia. *Int. Rev. Psychiatry* 22, 417–428. doi: 10.3109/09540261.2010.515205

Adjei, A. A., Thomae, B. A., Prondzinski, J. L., Eckloff, B. W., Wieben, E. D., Weinshilboum, R. M., et al. (2003). Human estrogen sulfotransferase (SULT1E1) pharmacogenomics: gene resequencing and functional genomics. *Br. J. Pharmacol.* 139, 1373–1382. doi: 10.1038/sj.bjp.0705369

Allman, J. M., Hakeem, A., Erwin, J. M., Nimchinsky, E., and Hof, P. (2001). The anterior cingulate cortex. The evolution of an interface between emotion and cognition. *Ann. N.Y Acad. Sci.* 935, 107–117. doi: 10.1111/j.1749-6632.2001.tb03476.x

Augui, S., Nora, E. P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.* 12, 429–442. doi: 10.1038/nrg2987

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

Beery, A. K., and Zucker, I. (2011). Sex bias in neuroscience and biomedical research. *Neurosci. Biobehav. Rev.* 35, 565–572. doi: 10.1016/j.neubiorev.2010.07.002

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T. J., et al. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* 508, 494–499. doi: 10.1038/nature13206

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* 57, 289–300. doi: 10.2307/2346101

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J. Abeygunawardena, N., et al. (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68–71. doi: 10.1093/nar/gkg091

Buckberry, S., Bent, S. J., Bianco-Miotto, T., and Roberts, C. T. (2014a). massiR: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics* 30, 2084–2085. doi: 10.1093/bioinformatics/btu161

Buckberry, S., Bianco-Miotto, T., Bent, S. J., Dekker, G. A., and Roberts, C. T. (2014b). Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface. *Mol. Hum. Reprod.* 20, 810–819. doi: 10.1093/molehr/gau035

Buxbaum, J. D., Silverman, J. M., Smith, C. J., Greenberg, D. A., and Kilifarski, M., Reichert, J., et al. (2002). Association between a GABRB3 polymorphism and autism. *Mol. Psychiatry* 7, 311–316. doi: 10.1038/sj.mp.4001011

Carlson, C., Dugan, P., Kirsch, H. E., and Friedman, D. (2014). Sex differences in seizure types and symptoms. *Epilepsy Behav.* 41, 103–108. doi: 10.1016/j.yebeh.2014.09.051

Carrel, L., and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434, 400–404. doi: 10.1038/nature03479

Coffer, P. J., and Burgering, B. M. (2004). Forkhead-box transcription factors and their role in the immune system. *Nat. Rev. Immunol.* 4, 889–899. doi: 10.1038/nri1488

Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S., Brown, C. J., et al. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* 24, 1528–1539. doi: 10.1093/hmg/ddu564

Di Lorenzo, A., and Bedford, M. T. (2011). Histone arginine methylation. *FEBS Lett.* 585, 2024–2031. doi: 10.1016/j.febslet.2010.11.010

Dunning, M. J., Smith, M. L., Ritchie, M. E., and Tavaré, S. (2007). beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* 23, 2183–2184. doi: 10.1093/bioinformatics/btm311

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Eastwood, J. A., and Doering, L. V. (2005). Gender differences in coronary artery disease. *J. Cardiovasc. Nurs.* 20, 340–351. *quiz*: 352–343. doi: 10.1097/00005082-200509000-00008

Ebers, G. C., Sadovnick, A. D., Dyment, D. A., Yee, I. M., Willer, C. J., and Risch, N. (2004). Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *Lancet* 363, 1773–1774. doi: 10.1016/S0140-6736(04)16304-6

Ehrenkranz, J., Bach, P. R., Snow, G. L., Schneider, A., Lee, J. L., Ilstrup, S., et al. (2015). Circadian and circannual rhythms in thyroid hormones: determining the TSH and Free T4 reference intervals based upon time of day, age, and sex. *Thyroid* 25, 954–961. doi: 10.1089/thy.2014.0589

Fermin, D. R., Barac, A., Lee, S., Polster, S. P., Hannenhalli, S., Bergemann, T. L., et al. (2008). Sex and age dimorphism of myocardial gene expression in nonischemic human heart failure. *Circ. Cardiovasc. Genet.* 1, 117–125. doi: 10.1161/CIRCGENETICS.108.802652

Grundtman, C., Kreutmayer, S. B., Almanzar, G., Wick, M. C., and Wick, G. (2011). Heat shock protein 60 and immune inflammatory responses in atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 31, 960–968. doi: 10.1161/ATVBAHA.110.217877

Guiro, J., and O'Reilly, D. (2015). Insights into the U1 small nuclear ribonucleoprotein complex superfamily. *Wiley Interdiscip. Rev. RNA* 6, 79–92. doi: 10.1002/wrna.1257

Gurba, K. N., Hernandez, C. C., Hu, N., and Macdonald, R. L. (2012). GABRB3 mutation, G32R, associated with childhood absence epilepsy alters alpha1beta3gamma2L gamma-aminobutyric acid type A (GABAA) receptor expression and channel gating. *J. Biol. Chem.* 287, 12083–12097. doi: 10.1074/jbc.M111.332528

Hall, E., Volkov, P., Dayeh, T., Esguerra, J. L., Salö, S., Taneera, J., et al. (2014). Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biol.* 15:522. doi: 10.1186/s13059-014-0522-z

Hamamoto, R., Furukawa, Y., Morita, M., Iimura, Y., Silva, F. P., Li, M., et al. (2004). SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells. *Nat. Cell. Biol.* 6, 731–740. doi: 10.1038/ncb1151

Hu, S., Yao, G., Guan, X., Ni, Z., Ma, W., Wilson, E. M., et al. (2010). Research resource: genome-wide mapping of *in vivo* androgen receptor binding sites in mouse epididymis. *Mol. Endocrinol.* 24, 2392–2405. doi: 10.1210/me.2010-0226

Huang, C. C., Cheng, M. C., Tsai, H. M., Lai, C. H., and Chen, C. H. (2014). Genetic analysis of GABRB3 at 15q12 as a candidate gene of schizophrenia. *Psychiatr. Genet.* 24, 151–157. doi: 10.1097/YPG.0000000000000032

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, Y. H., Jankowski, A., Cheah, K. S., Prabhakar, S., and Jauch, R. (2015). SOXE transcription factors form selective dimers on non-compact DNA motifs through multifaceted interactions between dimerization and high-mobility group domains. *Sci. Rep.* 5:10398. doi: 10.1038/srep10398

Iio, C., Ogimoto, A., Nagai, T., Suzuki, J., Inoue, K., Nishimura, K., et al. (2015). Association between genetic variation in the scn10a gene and cardiac conduction abnormalities in patients with hypertrophic cardiomyopathy. *Int. Heart J.* 56, 421–427. doi: 10.1536/ihj.14-411

Jackson, B. C., Carpenter, C., Nebert, D. W., and Vasiliou, V. (2010). Update of human and mouse forkhead box (FOX) gene families. *Hum. Genomics* 4, 345–352. doi: 10.1186/1479-7364-4-5-345

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., et al. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41, 200–209. doi: 10.1093/ije/dyr238

Jin, V. X., Leu, Y.-W., Liyanarachchi, S., Sun, H., Fan, M., Andreassen, O. A., et al. (2004). Identifying estrogen receptor α target genes using integrated computational genomics and chromatin immunoprecipitation microarray. *Nucleic Acids Res.* 32, 6627–6635. doi: 10.1093/nar/gkh1005

Jönsson, E. G., Saetre, P., Nyholm, H., Djurovic, S., Melle, I., Andreassen, O. A., et al. (2012). Lack of association between the regulator of G-protein signaling 4 (RGS4) rs951436 polymorphism and schizophrenia. *Psychiatr. Genet.* 22, 263–264. doi: 10.1097/YPG.0b013e32834f3558

Joutel, A., Corpechot, C., Ducros, A., Vahedi, K., Chabriat, H., Mouton, P., et al. (1996). Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* 383, 707–710. doi: 10.1038/383707a0

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M.,et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature* 478, 483–489. doi: 10.1038/nature10523

Karlsson, C., Baudet, A., Miharada, N., Soneji, S., Gupta, R., Magnusson, M., et al. (2013). Identification of the chemokine CCL28 as a growth and survival factor

for human hematopoietic stem and progenitor cells. *Blood* 121, 3838–3842, S3831–S3815. doi: 10.1182/blood-2013-02-481192

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416. doi: 10.1093/bioinformatics/btn647

Krajewski, W. A., and Vassiliev, O. L. (2011). Interaction of SET domains with histones and nucleic acid structures in active chromatin. *Clin. Epigenetics* 2, 17–25. doi: 10.1007/s13148-010-0015-1

Kwon, A. T., Arenillas, D. J., Worsley Hunt, R., and Wasserman, W. W. (2012). oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda)* 2, 987–1002. doi: 10.1534/g3.112.003202

Lam, E. W. F., Brosens, J. J., Gomes, A. R., and Koo, C. Y. (2013). Forkhead box proteins: tuning forks for transcriptional harmony. *Nat. Rev. Cancer* 13, 482–495. doi: 10.1038/nrc3539

Lee, J. T. (2011). Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat. Rev. Mol. Cell. Biol.* 12, 815–826. doi: 10.1038/nrm3231

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Lin, L. C., Lewis, D. A., and Sibille, E. (2011). A human-mouse conserved sex bias in amygdala gene expression related to circadian clock and energy metabolism. *Mol. Brain* 4:18. doi: 10.1186/1756-6606-4-18

Liu, J., Zubieta, J. K., and Heitzeg, M. (2012). Sex differences in anterior cingulate cortex activation during impulse inhibition and behavioral correlates. *Psychiatry Res.* 201, 54–62. doi: 10.1016/j.pscychresns.2011.05.008

Maas, A. H. E. M., and Appelman, Y. E. A. (2010). Gender differences in coronary heart disease. *Neth. Heart J.* 18, 598–602. doi: 10.1007/s12471-010-0841-y

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., et al. (2014). JASPAR 2014, an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147. doi: 10.1093/nar/gkt997

Meikle, A. W. (2004). The interrelationships between thyroid dysfunction and hypogonadism in men and boys. *Thyroid* 14(Suppl 1.), S17–S25. doi: 10.1089/105072504323024552

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

Miki, Y., Nakata, T., Suzuki, T., Darnel, A. D., Moriya, T., Kaneko, C., et al. (2002). Systemic distribution of steroid sulfatase and estrogen sulfotransferase in human adult and fetal tissues. *J. Clin. Endocrinol. Metab.* 87, 5760–5768. doi: 10.1210/jc.2002-020670

Mogil, J. S., and Chanda, M. L. (2005). The case for the inclusion of female subjects in basic science studies of pain. *Pain* 117, 1–5. doi: 10.1016/j.pain.2005.06.020

Möller-Leimkühler, A. M. (2007). Gender differences in cardiovascular disease and comorbid depression. *Dialogues Clin. Neurosci.* 9, 71–83.

Morrow, E. H. (2015). The evolution of sex differences in disease. *Biol. Sex Dif.* 6:5. doi: 10.1186/s13293-015-0023-0

National Centre for Biotechnology Information (2016). U. S. N. L. o. M. w. n. n. n. g. g. t. g. A. D. Available online at: https://www.ncbi.nlm.nih.gov/genome/tools/gdp

O'Reilly, D., Dienstbier, M., Cowley, S. A., Vazquez, P., Drozdz, M., Taylor, S., et al. (2013). Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. *Genome Res.* 23, 281–291. doi: 10.1101/gr.142968.112

Ochoa, S., Usall, J., Cobo, J., Labad, X., and Kulkarni, J. (2012). Gender differences in schizophrenia and first-episode psychosis: a comprehensive literature review. *Schizophr. Res. Treat.* 2012:916198. doi: 10.1155/2012/916198

Plate, M., Li, T., Wang, Y., Mo, X., Zhang, Y., Ma, D., et al. (2010). Identification and characterization of CMTM4, a novel gene with inhibitory effects on HeLa cell growth through Inducing G2/M phase accumulation. *Mol. Cells* 29, 355–361. doi: 10.1007/s10059-010-0038-7

Qin, W., Liu, C., Sodhi, M., and Lu, H. (2016). Meta-analysis of sex differences in gene expression in schizophrenia. *BMC Syst. Biol.* 10(Suppl. 1), 9. doi: 10.1186/s12918-015-0250-3

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184. doi: 10.1371/journal.pmed.0050184

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., et al. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 44, W83–W89. doi: 10.1093/nar/gkw199

Reinius, B., and Jazin, E. (2009). Prenatal sex differences in the human brain. *Mol. Psychiatry* 14, 988–989. doi: 10.1038/mp.2009.79

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., et al. (2005). The DNA sequence of the human X chromosome. *Nature* 434, 325–337. doi: 10.1038/nature03440

Selva, D. M., and Hammond, G. L. (2009). Thyroid hormones act indirectly to increase sex hormone-binding globulin production by liver via hepatocyte nuclear factor-4alpha. *J. Mol. Endocrinol.* 43, 19–27. doi: 10.1677/JME-09-0025

Seney, M. L., and Sibille, E. (2014). Sex differences in mood disorders: perspectives from humans and rodent models. *Biol. Sex Dif.* 5:17. doi: 10.1186/s13293-014-0017-3

Shi, L., Zhang, Z., and Su, B. (2016). Sex biased gene expression profiling of human brains at major developmental stages. *Sci. Rep.* 6:21181. doi: 10.1038/srep21181

Shih, D. Q., Bussen, M., Sehayek, E., Ananthanarayanan, M., Shneider, B. L., Suchy, F. J., et al. (2001). Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism. *Nat. Genet.* 27, 375–382. doi: 10.1038/86871

Silkaitis, K., and Lemos, B. (2014). Sex-biased chromatin and regulatory cross-talk between sex chromosomes, autosomes, and mitochondria. *Biol. Sex Dif.* 5:2. doi: 10.1186/2042-6410-5-2

Splawski, I., Shen, J., Timothy, K. W., Lehmann, M. H., Priori, S., Robinson, J. L., et al. (2000). Spectrum of mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. *Circulation* 102, 1178–1185. doi: 10.1161/01.CIR.102.10.1178

Takeshima, H., Niwa, T., Takahashi, T., Wakabayashi, M., Yamashita, S., Ando, T., et al. (2015). Frequent involvement of chromatin remodeler alterations in gastric field cancerization. *Cancer Lett.* 357, 328–338. doi: 10.1016/j.canlet.2014.11.038

Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M. E, et al. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* 4:2771. doi: 10.1038/ncomms3771

Vakili, H., Jin, Y., and Cattini, P. A. (2014). Energy homeostasis targets chromosomal reconfiguration of the human GH1 locus. *J. Clin. Invest.* 124, 5002–5012. doi: 10.1172/JCI77126

van Nas, A., Guhathakurta, D., Wang, S. S., Yehya, N., Horvath, S., Zhang, B., et al. (2009). Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. *Endocrinology* 150, 1235–1249. doi: 10.1210/en.2008-0563

Vawter, M. P., Evans, S., Choudary, P., Tomita, H., Meador-Woodruff, J., Molnar, M., et al. (2004). Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology* 29, 373–384. doi: 10.1038/sj.npp.1300337

Voskuhl, R. R., and Palaszynski, K. (2001). Sex hormones in experimental autoimmune encephalomyelitis: implications for multiple sclerosis. *Neuroscientist* 7, 258–270. doi: 10.1177/107385840100700310

Wang, X., Yu, S., Li, F., and Feng, T. (2015). Detection of alpha-synuclein oligomers in red blood cells as a potential biomarker of Parkinson's disease. *Neurosci. Lett.* 599, 115–119. doi: 10.1016/j.neulet.2015.05.030

Wang, Y., Hu, Y., Fang, Y., Zhang, K., Yang, H., Ma, J., et al. (2009). Evidence of epistasis between the catechol-O-methyltransferase and aldehyde dehydrogenase 3B1 genes in paranoid schizophrenia. *Biol. Psychiatry* 65, 1048–1054. doi: 10.1016/j.biopsych.2008.11.027

Weickert, C. S., Elashoff, M., Richards, A. B., Sinclair, D., Bahn, S., Paabo, S., et al. (2009). Transcriptome analysis of male-female differences in prefrontal cortical development. *Mol. Psychiatry* 14, 558–561. doi: 10.1038/mp.2009.5

Wilson, C. L., and Miller, C. J. (2005). Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* 21, 3683–3685. doi: 10.1093/bioinformatics/bti605

Wu, L. J., Kim, S. S., Li, X., Zhang, F., and Zhuo, M. (2009). Sexual attraction enhances glutamate transmission in mammalian anterior cingulate cortex. *Mol. Brain* 2:9. doi: 10.1186/1756-6606-2-9

Yang, F., Babak, T., Shendure, J., and Disteche, C. M. (2010). Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* 20, 614–622. doi: 10.1101/gr.103200.109

Yang, X., Schadt, E. E., Wang, S., Wang, H., Arnold, A. P., Ingram-Drake, L., et al. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* 16, 995–1004. doi: 10.1101/gr.5217506

Yang, X., Wang, S., Kendrick, K. M., Wu, X., Yao, L., Lei, D., et al. (2015). Sex differences in intrinsic brain functional connectivity underlying human shyness. *Soc. Cogn. Affect. Neurosci.* 10, 1634–1643. doi: 10.1093/scan/nsv052

Zhang, X., Bertaso, F., Yoo, J. W., Baumgärtel, K., Clancy, S. M., Lee, V., et al. (2010). Deletion of the potassium channel Kv12.2 causes hippocampal hyperexcitability and epilepsy. *Nat. Neurosci.* 13, 1056–1058. doi: 10.1038/nn.2610

Zhang, Y., Klein, K., Sugathan, A., Nassery, N., Dombkowski, A., Zanger, U. M., et al. (2011). Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease. *PLoS ONE* 6:e23506. doi: 10.1371/journal.pone.0023506

Zucker, I., and Beery, A. K. (2010). Males still dominate animal studies. *Nature* 465:690. doi: 10.1038/465690a

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 Genetics

ORIGINAL RESEARCH
published: 23 May 2017
doi: 10.3389/fgene.2017.00062

Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods

Alessandra Dal Molin, Giacomo Baruzzo and Barbara Di Camillo *

Department of Information Engineering, University of Padova, Padova, Italy

The sequencing of the transcriptomes of single-cells, or single-cell RNA-sequencing, has now become the dominant technology for the identification of novel cell types and for the study of stochastic gene expression. In recent years, various tools for analyzing single-cell RNA-sequencing data have been proposed, many of them with the purpose of performing differentially expression analysis. In this work, we compare four different tools for single-cell RNA-sequencing differential expression, together with two popular methods originally developed for the analysis of bulk RNA-sequencing data, but largely applied to single-cell data. We discuss results obtained on two real and one synthetic dataset, along with considerations about the perspectives of single-cell differential expression analysis. In particular, we explore the methods performance in four different scenarios, mimicking different unimodal or bimodal distributions of the data, as characteristic of single-cell transcriptomics. We observed marked differences between the selected methods in terms of precision and recall, the number of detected differentially expressed genes and the overall performance. Globally, the results obtained in our study suggest that is difficult to identify a best performing tool and that efforts are needed to improve the methodologies for single-cell RNA-sequencing data analysis and gain better accuracy of results.

Keywords: single-cell RNA-seq, differential expression, differential distributions, benchmark, assessment

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Zlatko Trajanoski,
Innsbruck Medical University, Austria
Fabio Iannelli,
IFOM - The FIRC Institute of Molecular Oncology, Italy

***Correspondence:**
Barbara Di Camillo
barbara.dicamillo@unipd.it

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

Received: *18 March 2017*
Accepted: *08 May 2017*
Published: *23 May 2017*

Citation:
Dal Molin A, Baruzzo G and Di Camillo B (2017) Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods. Front. Genet. 8:62. doi: 10.3389/fgene.2017.00062

INTRODUCTION

Single-cell RNA-sequencing (scRNA-seq) has emerged a decade ago as a powerful technology for identifying and monitoring cells with distinct expression signatures in a population, and for studying the stochastic nature of gene expression; a task, this latter, possible only at single-cell level. Compared to bulk RNA-seq, scRNA-seq data are affected by higher noise deriving from both technical and biological factors. Technical variability mostly originates from the low amount of available mRNAs that need to be amplified in order to get the quantity suitable for sequencing. This process may lead to amplification biases or "dropout events," when the amplification or the capture are not successful (Kolodziejczyk et al., 2015; Stegle et al., 2015; Bacher and Kendziorski, 2016). Biological variability, instead, rises mainly from the stochastic nature of transcription (Chubb et al., 2006; Raj et al., 2006). Moreover, scRNA-seq has revealed multimodality in gene expression (Shalek et al., 2013) originating from the presence of multiple possible cell states within a cell population. The high variability of scRNA-seq data, the presence of dropout events that leads to zero expression measurements, and the multimodality of expression of a number of transcripts,

create some challenges for the detection of differentially expressed genes (DEGs), which is one of the main applications of scRNA-seq and the focus of the present work.

Many single-cell studies make use of methods for differential expression analysis originally developed for handling bulk RNA-seq data, e.g., (Brennecke et al., 2015; Tasic et al., 2016; Wang et al., 2016), which do not explicitly address the above challenges. A variety of methods has been recently proposed to analyze differential expression in scRNA-seq data (Bacher and Kendziorski, 2016). Most of them explicitly model the probability of dropout events, consider the multimodal nature of scRNA-seq data, or include a model of transcriptional burst.

Among the most popular scRNA-seq methods, Model-based Analysis of Single-cell Transcriptomics, MAST (Finak et al., 2015), explicitly considers the dropouts using a bimodal distribution with expression strongly different from zero or "non-detectable," and proposes a generalized linear model (GLM) to fit the data. Single-Cell Differential Expression, (SCDE; Kharchenko et al., 2014), models the counts of each cell as a mixture of a zero-inflated Negative Binomial distribution and a dropout component. Last, it uses a Bayesian model to estimate the posterior probability that a gene is differentially expressed in one group with respect to another. Monocle (Trapnell et al., 2014) is a tool originally designed for scRNA-seq data analysis for ordering cells based on their differentiation stage and extended to identify genes that are differentially expressed across different conditions. Data are fitted with a generalized additive model (GAM) and a Tobit model is used to account for dropout events. Another recently developed tool, Discrete Distributional Differential Expression, D^3E (Delmans and Hemberg, 2016), fits the bursting model of transcriptional regulation (Chubb et al., 2006; Raj et al., 2006) to the data and compares the gene expression distribution in one group with respect to another giving estimates of burst size, duty cycle, frequency, and mean of transcription. Single-cell Differential Distributions, scDD (Korthauer et al., 2016), is based on a multimodal Bayesian modeling framework for explicitly modeling the multimodal distributions of single cells and testing for differentially distributed genes associated with this multimodality. Bayesian Analysis of Single-Cell Sequencing Data, BASiCS (Vallejos et al., 2016), estimates the normalization parameters jointly across all genes by modeling spike-ins and endogenous genes as two Poisson-Gamma hierarchical models with shared parameters, and determines gene-specific posterior probabilities to identify highly variable genes.

Although a number of methods for the detection of DEGs in scRNA-seq have been developed, their performance on common benchmarks remains largely unclear. One recent study (Jaakkola et al., 2016), compared two scRNA-seq tools, MAST (Finak et al., 2015) and SCDE (Kharchenko et al., 2014), together with three tools traditionally used for the analysis of bulk RNA-seq data, Differential Expression analysis for Sequence count data, DESeq (Anders and Huber, 2010), Linear models for microarray and RNA-Seq data (Limma; Smyth, 2004), and Reproducibility-Optimized Test Statistic ROTS (Seyednasrollah et al., 2015), using three real datasets to assess their performance. In this study, we extended this comparison to four tools specifically developed for scRNA-seq data analysis (**Table 1**), MAST (Finak et al., 2015), SCDE (Kharchenko et al., 2014), Monocle (Trapnell et al., 2014), and D^3E (Delmans and Hemberg, 2016). Together with these tools, we also evaluated two of the most popular tools originally developed for DE analysis of bulk RNA-seq data (**Table 1**), DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010).

In addition to real scRNA-seq datasets (Islam et al., 2011; Grün et al., 2014), we used simulated datasets for our assessment. Using simulated data gives some advantages over the use of real data. Namely: (i) it provides a complete knowledge of positive, i.e., truly differentially expressed, and negative, i.e., truly not differentially expressed, genes; (ii) it gives the possibility to run replicated experiments, thus statistically testing the difference of the assessment scores; (iii) it allows testing different data scenarios. In this work, we specifically addressed the multimodality of scRNA-seq data, assessing methods performance on four different scenarios, as defined in Korthauer et al. (2016), related to different data distributions of the two conditions to be compared (**Figure 1**):

1. Unimodal distributions with different means (DE);
2. Bimodal distribution with different proportions of cells in the two components and equal component means across conditions (DP);
3. Unimodal distribution for one condition and bimodal distribution for the other, with one overlapping component and with equal component means across conditions (DM);
4. Unimodal distribution for one condition and bimodal distribution for the other, with different component means across conditions (DB).

Among the above listed scRNA-seq tools, BASiCS (Vallejos et al., 2016) and scDD (Korthauer et al., 2016) were not included in our comparison. BASiCS requires as input a set of spike-ins expression values, therefore it was not applicable to all the datasets used in our study. On the other side, scDD requires R version 3.4, which is a version of R under development and not stable.

MATERIALS AND METHODS

Real Datasets

To assess the performance of the selected methods we used the dataset published by Islam et al. (2011) consisting of 48 mouse Embryonic Stem Cells and 44 mouse Embryonic Fibroblasts analyzed using scRNA-seq, in parallel with a study by Moliner et al. (2008), conducted using the same cell types and culturing conditions, and followed by the validation of microarray expression measurements with qRT-PCR. Similarly to what was previously done by others (Kharchenko et al., 2014; Jaakkola et al., 2016), we used the top 1,000 DEGs from Moliner et al. as "positive control" to test the ability of the benchmarked tools to detect true positive genes. ScRNA-seq data, containing raw counts for 22,928 genes (excluded 8 spike-ins), were retrieved from GEO database with accession number GSE29087.

We used a second scRNA-seq dataset, published by Grün et al. (2014), as negative control. This dataset consists of 80 single cells

TABLE 1 | Tools compared in this study.

Tool	Model	Programming language	Operating system	Parallel execution
MAST; Finak et al., 2015	Generalized linear hurdle model	$R \geq 3.3$	Unix/Linux, Mac OS, Windows	Yes
SCDE; Kharchenko et al., 2014	Mixture of a negative binomial distribution and low-level Poisson distribution	$R \geq 3.0.0$	Unix/Linux, Mac OS, Windows	Yes
Monocle; Trapnell et al., 2014	Generalized additive model	$R \geq 2.10.0$	Unix/Linux, Mac OS, Windows	Yes
D^3E; Delmans and Hemberg, 2016	Transcriptional bursting model	Python*	Unix/Linux, Mac OS, Windows	No
DESeq; Anders and Huber, 2010	Negative binomial distribution	*R**	Unix/Linux, Mac OS, Windows	No
edgeR; Robinson et al., 2010	Negative binomial distribution	$R \geq 2.15.0$	Unix/Linux, Mac OS, Windows	No

MAST, SCDE, Monocle, and D^3E have been specifically developed for the analysis of scRNA-seq data. DESeq and edgeR have been originally designed for bulk RNA-seq data analysis. (*) No information available about the version.

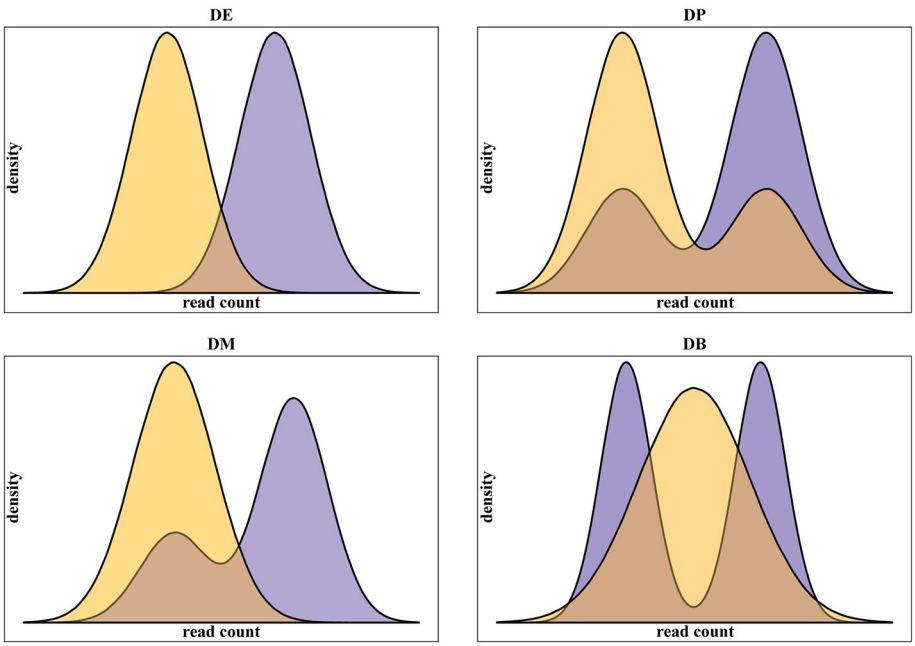


FIGURE 1 | Examples of the four classes of differential distributions, as defined in Korthauer et al. (2016), including on top-left the traditional differential expression (DE), the differential proportions of cells in multimodal distributions (DP) on top-right, the differential modality (DM) on bottom-left and both differential modality and proportions (DB) on bottom-right.

and 80 pool-and-split (P&S) samples cultured both in serum and two-inhibitor (2i) media. Briefly, P&S samples were generated by pooling ∼1 million single cells, splitting them into single-cell equivalents (∼20 pg) of RNA and then sequencing in the same way as single cells. Starting from the 80 P&S samples, we randomly sampled 10 times the 40 samples as control condition and the other 40 samples as testing condition, thus generating 10 independent datasets. These datasets were used as "negative control" for differential expression analysis, as no DEGs are expected in any of these comparisons. The raw counts of scRNA-seq data, for a total of 12,476 genes (excluded 59 spike-ins), were retrieved from GEO database with accession number GSE54695. Data were converted to UMI counts as described in the original publication (Islam et al., 2011): the total number of sequenced transcripts was calculated as $-K \ln\left(1 - {}^{k_{o,i}}/_{K}\right)$, where K denotes the total number of UMIs and $k_{o,i}$ denotes the number of observed UMIs for gene i.

Simulated Datasets

The simulated datasets were generated using the scripts provided with scDD package in the recently published study by Korthauer et al. (2016). More in details, 10,000 genes were simulated for two conditions with sample size of 100 cells each. 8,000 genes were simulated as not differentially expressed using the same distribution (unimodal for half of the genes and bimodal for the remaining) in the two conditions. Specifically, the unimodal genes were generated from the same Negative Binomial (NB) distribution, while the bimodal genes were generated from a two-component NB mixture. The remaining 2,000 genes were simulated as differentially expressed accordingly to the four types of differential expression, DE, DP, DM,

and DB, defined in section Introduction consistently with Korthauer et al. (2016). Five-Hundred DEGs for each group were generated. The datasets were obtained by running the script *simulateSet.R* and using as starting data the synthetic dataset *scDatEx* provided by the authors together with the package. All parameters for simulation were set as defaults and data were rounded to the nearest integer. The procedure was repeated 10 times in order to produce 10 independent synthetic replicates.

Methods for Differential Gene Expression Analysis

We tested four methods developed for differential expression analysis of genes between single-cell populations: MAST (version 1.0.5) (Finak et al., 2015), SCDE (version 1.99.1) (Kharchenko et al., 2014), Monocle (version 2.2.0) (Trapnell et al., 2014), and D^3E (version 1.0) (Delmans and Hemberg, 2016). In addition, we tested two widely used DE methods originally developed for bulk RNA-seq data, DESeq (version 1.26.0) (Anders and Huber, 2010) and edgeR (version 3.12.1) (Robinson et al., 2010). For all methods, raw data were provided as input and, except for what specified below, all the tools were run using the default parameters. Differential expression measures were retained significant when adjusted *p*-values were below a False Discovery Rate (FDR) cut-off of 0.05. Precision and Recall metrics were calculated as, respectively, the number of true positives among all positive calls and the number of true positives among the true number of DEGs.

MAST

MAST employs a generalized linear hurdle model to account simultaneously for stochastic dropouts and characteristic bimodal expression distributions in which expression is either strongly non-zero or non-detectable. The rate of expression Z, and the level of expression Y, are modeled for each gene g, indicating whether gene g is expressed in cell i (i.e., $z_{ig} = 0$ if $y_{ig} = 0$ and $z_{ig} = 1$ if $y_{ig} > 0$). A logistic regression model for the discrete variable Z and a Gaussian linear model for the continuous variable $(Y \mid Z = 1)$ are considered:

$$logit\left(P_r\left(Z_{ig} = 1\right)\right) = X_i\beta_g^D$$
$$P_r\left(Y_{ig} = y | Z_{ig} = 1\right) = N(X_i\beta_g^C, \sigma_g^2)$$

where X_i is the design matrix. The fraction of genes that are expressed and detectable in each cell, called cellular detection rate (CDR), can be explicitly modeled as a covariate (a column in the design matrix X_i), allowing a joint estimate of nuisance and treatment effects. In order to improve the inference for genes with sparse expression, the model parameters are fitted using an empirical Bayesian framework. Finally, differential expression is determined using the likelihood ratio test.

In our assessment, MAST with both the adjustment for CDR and the omission of this covariate (MASTNotCDR) were included.

SCDE

SCDE models the read counts computed for each gene using a mixture of a NB distribution and a Poisson distribution. The NB distribution models the transcripts that are amplified and detected, whereas the low-magnitude Poisson distribution models the unobserved or background-level signal of transcripts that are not amplified (i.e., dropout events). Although, the dropout component could be modeled as a constant zero (i.e., zero-inflated negative binomial process) the use of a low-magnitude Poisson process allows accounting for both the dropouts and some background signals that are typical of transcriptionally silent genes. A subset of robust genes (i.e., genes that are detected in multiple cross-cell comparisons) is used to fit, using an EM algorithm, the parameters of the mixture models. For the differential expression analysis, the posterior probability that the gene shows a fold expression difference between two conditions is computed using a Bayesian approach. An empirical *p*-value to test for significance of expression difference is determined by normalizing to unity the posterior distributions.

Monocle

Monocle is a tool originally designed for single-cell RNA-seq data analysis for ordering cells by progress through differentiation stages (pseudo-time). The tool is able to identify genes that change significantly over the time and that are differentially expressed across different cell types or conditions. The mean expression level of each gene is modeled with a GAM which relates one or more predictor variables to a response variable as

$$g\left(E\left(Y\right)\right) = \beta_0 + f_1\left(x_1\right) + f_2\left(x_2\right) + \ldots + f_m\left(x_m\right)$$

where Y is a specific gene expression level, and the x_i's are predictor variables. The function g is a link function, typically the log function, while the f_i's are non-parametric functions, such as cubic splines or some other smoothing functions. The observable (log-transformed) expression level Y is modeled using a Tobit model censored below a user defined expression detection threshold. Monocle's GAM is thus

$$E\left(Y\right) = s\left(\varphi_t\left(b_x, s_i\right)\right) + \varepsilon$$

where $\varphi_t\left(b_x, s_i\right)$ is the assigned pseudo-time of a cell and s is a cubic smoothing function with (by default) three degrees of freedom. The error term ε is normally distributed with a mean of zero. The tool also supports testing for differential expression between groups. In these tests, the GAM employs the class labels as predictor variables, with no smoothing. Finally, the test for differential expression is performed using an approximate χ^2 likelihood ratio test.

Since we are interested only in the comparison of genes among different conditions, the temporal ordering feature was not used in our study. When creating *newCellDataSet* at the beginning of the analysis we used the parameter *expressionFamily = negbinomial()* for each dataset. We were not able to estimate the data dispersion since the function performing the parametric fit failed both on simulated and real data and it was not possible to modify it for a local fit and/or a pooled estimation of dispersion.

D^{3}E

D^{3}E consists of two separate modules: a module for comparing expression profiles using the Cramér-von Mises, the likelihood ratio test, the Kolmogorov-Smirnov test or the Anderson-Darling test and a module for fitting the transcriptional bursting model (Peccoud and Ycart, 1995; Chubb et al., 2006; Raj et al., 2006). This latter provides biological insight into the mechanisms underlying the change in expression. Initially, the input read counts are normalized using the DESeq algorithm procedure and genes that are not expressed in any of the cells are removed. Second, the Cramér-von Mises (CvM) test (default), the Kolmogorov-Smirnov (KS) or the Anderson-Darling test can be used to detect differential expression. Alternatively, the transcriptional bursting model is fitted for each gene to the expression data in both conditions and the change in parameters between the two conditions is tested using the likelihood ratio test.

In our study, D^{3}E analyses were performed using both the Cramér-von Mises test (default option) and the Kolmogorov-Smirnov test.

DESeq

DESeq assumes that the number of reads in a bulk RNA-seq sample j that are assigned to gene i can be modeled by a negative binomial distribution with mean and variance estimated from the data. For each gene, the expectation value of the observed counts for gene i in sample j, i.e., the mean μ_{ij} of the NB distribution, is modeled as the product of the (unknown) expectation value of the true concentration of reads and a size factor s_j accounting for the sequencing depth. The variance of the NB distribution σ^2_{ij} is modeled as the sum of a *shot noise terms* (μ_{ij}) and a *raw variance term*:

$$\sigma^2_{ij} = \mu_{ij} + s^2_j v_{i,\rho(j)}$$

The raw variance term is proportional to the square of the scaling factor s_j and to the expected true concentration of reads $v_{i,\rho(j)}$. For each gene, the statistical test is performed defining, for each gene i, the total read counts for each of the two conditions (e.g., K_{iA} and K_{iB}, for conditions A and B) and computing, under the null hypothesis, the p-value as the probability of the events $K_{iA} = a$ and $K_{iB} = b$ for any pair of numbers a and b, given that $a + b$ equals the observed sum of counts.

Since DESeq is able to manage only non-zero data, in the specific cases of Grün and Islam datasets a pseudo-count of $+1$ was added to zero counts. Estimation of dispersion was performed using the "*local*" option.

edgeR

Similar to DESeq, edgeR models the computed read counts using a NB distribution. For each gene, the mean μ of the NB distribution is the product of the total number of reads and the (unknown) relative abundance of that gene in the current experimental condition. The variance σ^2 is related to the mean by $\sigma^2 = \mu + \alpha\mu^2$, requiring the estimation of the over-dispersion parameter α. The method estimates the gene-wise dispersions using a conditional maximum likelihood procedure, conditioning on the total read count of each gene (Smyth and Verbyla, 1996) and an empirical Bayes procedure to shrink the dispersions toward a consensus value. For each gene, the differential expression test is performed using the GLM likelihood ratio test (Robinson and Smyth, 2008).

In our tests, edgeR was run estimating the *Tagwise* dispersion, using the *glmFit* function to fit the data and *glmLRT* to compare the two conditions.

RESULTS

Results on Simulated Datasets

The number of selected DEGs resulting from the analysis of simulated data ranged, on average, from 1,021 to 1,741 with a number of true positives from 1,018 to 1,534 (**Table 2**). In general, all the tools underestimated the number of DEGs with an average of ∼1,378 called DEGs. D3E_CvM detected, on average, the highest number of DEGs with the highest variability among the ten different tests.

For each tool, we calculated the precision and recall values as described in Section Materials and Methods. The precision-recall (PR) curves of the different methods are shown in **Figure 2A**. The values of Area under the Recall Precision Curve (AURPC) obtained by the tools specifically designed for scRNA-seq data analysis tends to be high (**Figure 2B**), with median value equal to 0.914, 0.903, 0.902, and 0.885 for MAST, SCDE, D3E_KS, and Monocle, respectively. Bulk methods showed median AURPC equal to 0.895 and 0.899, for DESeq and edgeR, respectively.

All methods performed similarly in ranking DEGs, with the exception of Monocle (dark green line), which showed very low precision values for the first genes selected at differentially expressed and high variability between the ten different performed tests. When looking separately at precision and recall values (**Figures 2C,D**), MAST, SCDE, and DESeq reported the highest values for precision (median of, respectively 0.995, 0.998, and 0.994), which were even higher than the chosen cut-off of 0.95, but the lowest for recall (median of, respectively 0.574, 0.508, and 0.555). Contrarily, both D3E_CvM and D3E_KS together with Monocle showed lower values for precision with median, respectively of 0.866, 0.909, and 0.935, and higher recall with respect to the other tools (median between 0.70 and 0.80).

TABLE 2 | Mean number of DEGs (± standard deviation) detected by each of the assessed tools below the FDR cut-off of 0.05.

Tool	No. DEGs (mean ± sd)	No. true DEGs (mean ± sd)
MAST	1,153.00 ± 15.19	1,148.10 ± 15.72
MASTNotCDR	1,149.00 ± 15.55	1,144.10 ± 15.72
SCDE	1,021.30 ± 25.64	1,018.10 ± 24.92
Monocle	1,576.70 ± 8.47	1,471.30 ± 17.17
D^{3}E CvM	1,741.00 ± 34.28	1,507.30 ± 7.78
D^{3}E KS	1,700.70 ± 23.22	1,534.40 ± 16.70
DESeq	1,122.60 ± 16.95	1,116.20 ± 17.75
edgeR	1,564.50 ± 15.50	1,471.10 ± 16.75

The third column reports the average number of true DEGs (± standard deviation) among the total number of detected DEGs.

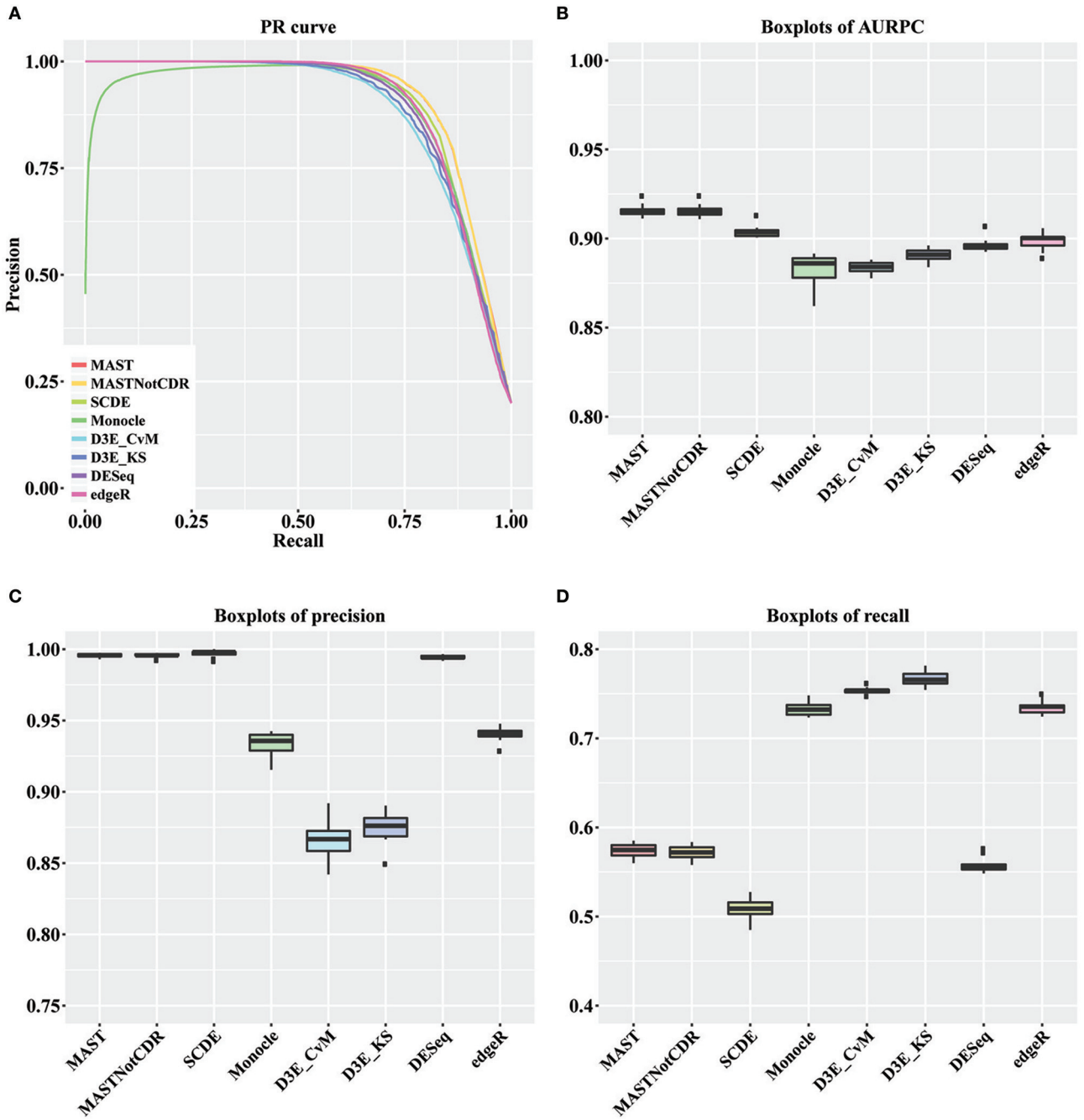


FIGURE 2 | Results of the analysis of simulated data. (A) Global PR curve for all tested tools. **(B)** Boxplots of global AURPC. **(C)** Boxplots of global Precision. **(D)** Boxplots of global Recall.

edgeR resulted in intermediate values of precision (median equal to 0.941) and recall (median equal to 0.735) with respect to all other tools.

The significant difference among tools' performance scores were assessed by a Kruskal-Wallis test (Kruskal and Wallis, 1952) followed by a paired Wilcoxon rank test (Wilcoxon, 1946). For AURPCs we obtained a Kruskal-Wallis *p*-value equal to 1.46e-12, with Wilcoxon *p*-value always lower than 3.7e-02 for the comparison of MAST and MASTNotCDR with any other method. For precision, we obtained a Kruskal-Wallis *p*-value equal to 1.22e-12, with Wilcoxon *p*-value always lower than 3.90e-03 for the comparison of MAST, MASTNotCDR, SCDE, and DESeq with any other method. For recall, we obtained a Kruskal-Wallis *p*-value equal to 1.75e-13 with Wilcoxon *p*-value always lower than 0.58e-03 for the comparison of Monocle, D^3E and edgeR with any other method.

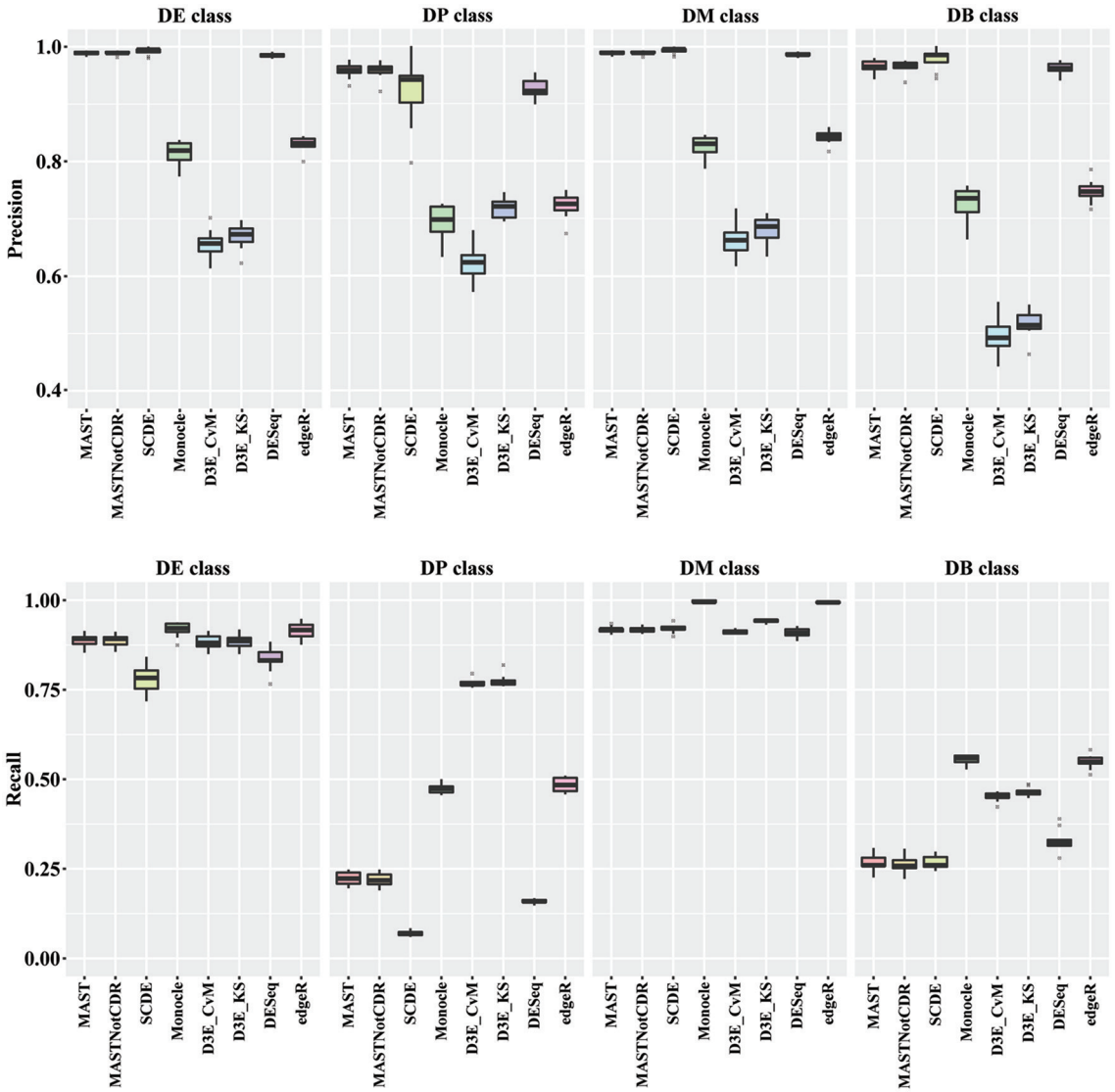


FIGURE 3 | Boxplots of Precision and Recall of simulated data for all tools, reported for the four Differential Distributions classes.

In order to understand the ability to detect DEGs in the four different scenarios DE, DP, DM, and DB, we evaluated precision and recall separately on the four classes of DEGs defined in Section Materials and Methods. In general, all tools performed better for the DE and the DM classes, which had the highest precision and recall values with respect to the other two classes (**Figure 3**). For the DE class, MAST showed the highest precision together with SCDE and DESeq; whereas the highest recall values were observed for Monocle and edgeR. For the DP class, precision resembled the results obtained for the DE and the DM classes, but MAST had a drop in recall, which was instead the highest for D^3E. Also in the case of DB class, the trend for precision was essentially the same of the other classes, but recall significantly dropped for all methods. Globally, in terms of precision, MAST and SCDE and DESeq outperformed the other tools (Kruskal-Wallis *p*-value always lower than 1e-08 for the four classes and paired Wilcoxon test *p*-value always lower than 2.70e-02 when comparing MAST, SCDE, or DESeq with any other method). edgeR and Monocle had the highest recall values for DE, DM, and DB classes (Kruskal-Wallis *p*-value equal to 7.89e-09, 8.01e-11, and 4.93e-16 followed by a paired Wilcoxon test *p*-value always lower than 5.85e-03, 5.82e-03, and 5.88e-03, for DE, DM, and DB, respectively, when comparing edgeR and Monocle with any other method), whereas D^3E performed better than other in recall for the DP class (Kruskal-Wallis *p*-value equal to 1.32e-13 followed by a paired Wilcoxon test *p*-value always lower than 5.88e-03 when comparing D^3E with any other method).

Results on Real Datasets

The analysis of Islam dataset resulted in a number of detected DEGs ranging from 271 to 8,401, depending on the tool (**Figure 4**). D^3E with CvM test (hereafter D3E_CvM) and MAST without CDR covariate (MASTNotCDR) detected the highest

number of DEGs compared to other tools. The intersection of DEGs with Moliner's reference list of the top 1,000 ranking genes accordingly to qRT-PCR (**Figures 4**, **5**), was higher for D3E_CvM (707 common genes) and MASTNotCDR (691 common genes), followed by edgeR (561), and DESeq (459). On the contrary, MAST, SCDE, and Monocle showed lower intersection. **Figure 4** also shows on the top of each red bar, the fraction of genes, within the reference list, called as significant, and, on the top of each blue bar, the ratio between the intersection with Moliner's reference list and the total number of called DEGs for each tool. This ratio can be roughly considered a true positive ratio score, although keeping in mind that, besides the validation by qRT-PCR, the number and the identity of true DEGs is not known. Notably, even having the highest intersection with Moliner reference list, tools as MASTNotCDR and D^{3}E have the lowest values of ratio due to the high numbers of called DEGs. The number of DEGs present in the Moliner's gene list and consistently called by all the compared tools was only 23 (**Figure 5**), due to the low intersection of MAST DEGs with Moliner's gene list. Indeed, when considering common genes among all tools but MAST, 214 common DEGs were obtained. The highest pair-wise intersection (135 common DEGs) was shown by D^{3}E and MASTNotCDR, which were the tools with the highest numbers of called DEGs (**Figure 4**). It is interesting to report that a small number of DEGs were called specifically by each tool with null intersection with other tools (**Figure 5**).

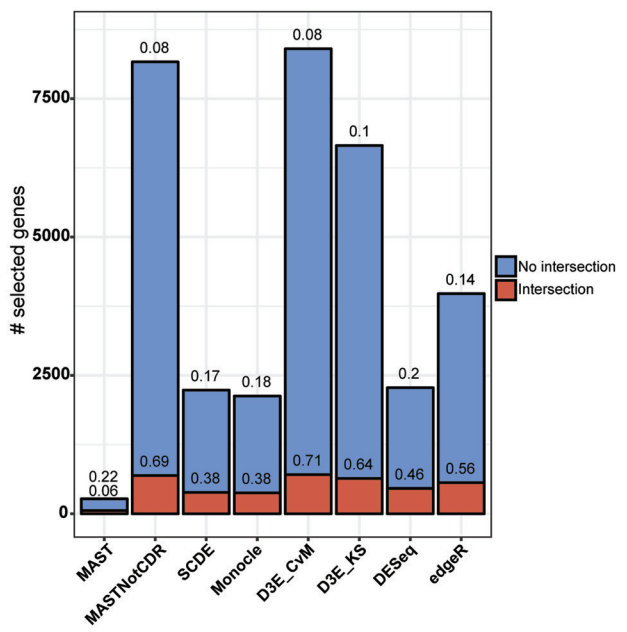


FIGURE 4 | Results of the analysis of Islam dataset using as benchmark dataset the list of top 1,000 DEGs of Moliner et al. (2008). Stacked barplots of detected DEGs are shown for all tools. The coral bar indicates the intersection with Moliner reference list. On the top of each coral bar is reported the ratio of detected Moliner genes among the total 1,000 assumed to be true positives. On the top of each blue bar is reported the ratio between the intersection with Moliner's reference list and the total number of called DEGs.

The 10 datasets derived from Grün et al. (2014) sampling the P&S samples were then used as negative control to additionally evaluate the performance of the tools, with an expectation of zero DEGs. In general, all the tools showed good performance, as they did not detect DEGs in any of the ten P&S datasets, with the exception of D3E_KS and D3E_CvM that consistently detected, in each of the 10 tests, 271 and 422 DEGs, respectively.

Running Time

We performed all the analyses on a HPC cluster consisting of 6 octa-core IBM Power7 processors, 640 Gb of RAM and running SUSE Linux Enterprise 11. All the analyses were carried out using R version 3.3.2 and, for D^{3}E, python version 2.7.6. The LoadLeveler job scheduling system version 4.1 was used designing a job for each test and assigning 8 cores to the job, when the tool supported parallel execution, as in case of MAST, SCDE and Monocle. We also used LoadLeveler to calculate the Run Time, which is defined as the difference between exiting time and starting time. Summary statistics are shown in **Table 3**, in case of both parallel (8 cores) and serial (1 core) execution. Among the tested scRNA-seq tools, MAST was the fastest to run (on average ∼4 min with 8 cores and ∼17 min with 1 core), whereas Monocle and D^{3}E were the most computationally intensive (∼7 h and ∼4 days with 1 core, respectively). Tools supporting parallel execution in general achieved a considerable speed up, especially Monocle. The remaining bulk methods were generally fast, as they did not include any heavily time-consuming steps.

DISCUSSION

Design of the Study

In this work, we evaluated the performance of six differential expression analysis methods on two published scRNA-seq datasets (Islam et al., 2011; Grün et al., 2014) and 10 simulated scRNA-seq datasets (Korthauer et al., 2016).

The scRNA-seq dataset published by Islam et al. (2011) was employed for the assessment, using a list of 1,000 top ranking DEGs obtained from a quantitative experimental validation through qRT-PCR as positive controls (Grün et al., 2014; Kharchenko et al., 2014), as previously done by others (Jaakkola et al., 2016).

Grün et al. scRNA-seq dataset (Grün et al., 2014) was instead used as "negative control" for differential expression, as it makes available P&S samples, consisting of pooled RNA from thousands of mouse Embryonic Stem Cells split into equivalent volumes. Indeed, no overall changes in gene expression are expected between any of these samples since the P&S procedure generates replicates that in principle are not expected to show any biological variability.

Since real datasets can provide only partial information in terms of positive and negative controls, we decided to use also simulated data to assess the different methods' performance.

Synthetic datasets were generated using the R scripts provided by Korthauer et al. along with their package scDD (Korthauer et al., 2016). The simulation was undertaken to allow an unbiased evaluation of precision and recall of each tool in detecting differential expression, focusing on both global results and

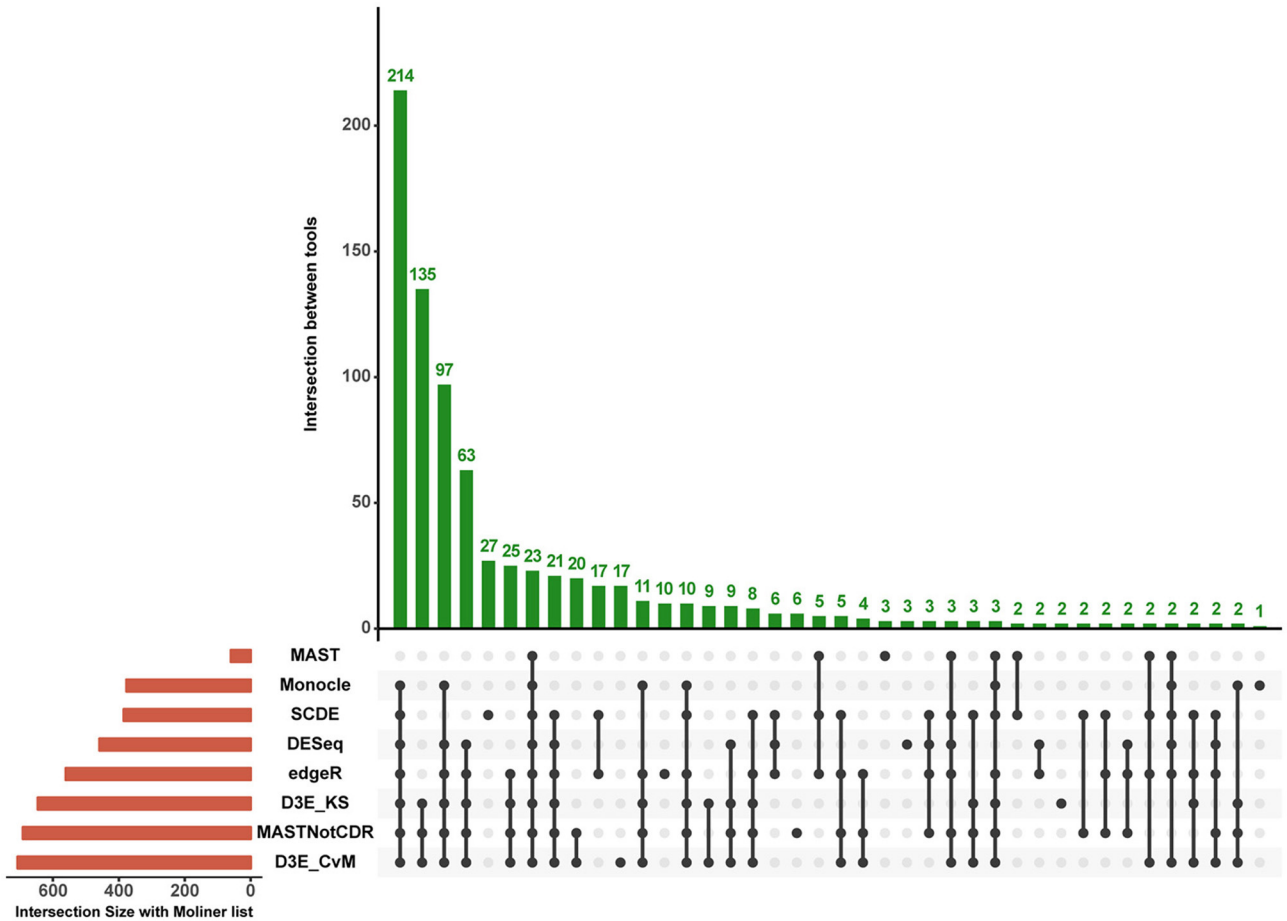


FIGURE 5 | Intersection plot of the tools under comparison. The coral-colored histogram located next to the tools' names, corresponds to the coral bar of **Figure 4**, as it reports for each tool the intersection size (i.e., number of DEGs in common) with Moliner's reference list. The green-colored histogram shows the intersection of Moliner's reference list with different combinations of tools. The "dot matrix" below the figure shows these different combinations by indicating with black dots the tools considered in the intersection and with gray dots the tools that do not contribute to the intersection.

specific gene categories, namely DE for traditional differential expression, DP for differential proportions of cells in two-components distributions, DM for differential modality with one overlapping component and DB for both differential proportions and differential modality.

Among the six assessed tools, MAST, SCDE, Monocle and D^{3}E, were recently developed for the analysis of scRNA-seq data; while the remaining two, DESeq and edgeR, are among the most popular tools used for the analysis of bulk RNA-seq data, and are currently applied also for scRNA-seq. Originally, we tested also DESeq2 (Love et al., 2014); however, since it achieved consistently lower performance than its previous version both in term of precision and recall, we decided to use DESeq.

TABLE 3 | Summary statistics of run time for all tools on simulated data.

Tool	Run time (parallel) (avg ± sd) (dd:hh:mm:ss)	Run time (serial) (avg ± sd) (dd:hh:mm:ss)
MAST	00:00:03:52 ± 00:00:00:65	00:00:16:57 ± 00:00:03:47
SCDE	00:00:19:25 ± 00:00:02:02	00:01:26:75 ± 00:00:10:08
Monocle	00:01:05:04 ± 00:00:07:08	00:07:04:44 ± 00:00:11:05
D3E_CvM	–	04:19:39:46 ± 00:01:39:35
D3E_KS	–	04:18:41:22 ± 00:01:13:33
DESeq	–	00:00:26:14 ± 00:00:02:12
edgeR	–	00:00:03:23 ± 00:00:01:10

We reported mean and standard deviations of ten tests performed.

Number of Detected DEGs

Looking globally at our results, the six analyzed tools had very different behavior in terms of the number of detected DEGs. All tools were conservative in calling DEGs on simulated datasets, as the average number of called DEGs was around 70% of the true number of simulated DEGs and the proportion was consistent across different methods.

The results were consistent across tools even when considering the number of detected DEGs in Grün datasets, with the exception of D^{3}E. Indeed, all the tools selected 0 genes as differentially expressed, whereas D^{3}E was the only one consistently detecting, across the 10 P&S datasets, of the same DEGs and in particular, 271 genes with KS test and 422 with CvM test.

On the other hand, when analyzing Islam dataset, the number of called DEGs was very different (from 271 to 8,401) across the different tools used, with MASTNotCDR and D^3E calling the highest number of DEGs.

Control of Precision and Recall

We tested the ability of each tool in detecting true DEGs or experimentally validated DEGs, in terms of precision, both on simulated and Islam real dataset (Islam et al., 2011). In case of the real dataset, the results were difficult to interpret given the fact that we cannot be sure if the 1,000 genes in the Moliner's reference list are actually true positives and if there are not any other DEGs in the dataset (Moliner et al., 2008).

Globally, the estimated percentage of true positive on simulated data ranged between 0.84 and 0.99, whereas on real data it ranged between 0.08 (for MASTNotCDR and D3E_CvM) and 0.22 (for MAST).

Among the assessed tools, SCDE outperformed the other methods in terms of precision but, consistently, had a drop in performance in terms of recall, both on real and simulated datasets. In particular, on simulated data, the average observed precision was above the 95% required as input, based on a FDR threshold of 5%, highlighting a good but slightly conservative control of false positive, with a consequent loss in recall.

MAST had a contradictory behavior on simulated with respect to real dataset. As SCDE, on simulated data the precision for MAST was above the required cut-off while the recall dropped to lower values with respect to SCDE. In case of the real dataset, the inclusion of the CDR covariate highly affected the results, with a lower number of called DEGs with respect to all the other tools when including it, and a higher number of detections when excluding this covariate. In both cases, however, the intersection size with Moliner's reference gene list (Moliner et al., 2008) was small.

Monocle showed a good trade-off between precision and recall on simulated datasets, with average precision, however, slightly lower than 95% and a number of false positive genes ranked at top differentially expressed gene positions, which contributed to the decrease of its average area under the precision-recall curve. On real datasets, however, the tool was among the best performing ones in terms of intersection size with Moliner's reference gene list (Moliner et al., 2008).

D^3E was the tool with the poorest control of false positive rates on simulated datasets, while performing best in terms of recall. This trend was consistent also when analyzing the real dataset, as it had the highest recall but the lowest precision, considering both Moliner's reference list (Moliner et al., 2008) as benchmark for true positive calls and P&S negative control datasets D^3E resulted to be the worst performing tool probably because it's not designed to account for data multimodality. Anyway, this tool includes in the computation the fit of the model of the transcriptional burst, feature that is very interesting but not tested in this study as the synthetic data did not simulate this feature of transcription.

Surprisingly, bulk methods worked well with simulated scRNA-seq data and showed good performance in handling the multimodal nature of such kind of data. Indeed, both DESeq and edgeR, reported a good trade-off between precision and recall both on real and simulated datasets.

It is worth noting that the relative performance of the methods used both in our study and in Jaakkola et al. (2016) are consistent, with SCDE outperforming DESeq and MAST, even if Islam dataset has been processed in a different way in the two studies.

Performance on Data with Different Modalities

As regards the comparison of methods performance on different type of data distributions, in general all the tools performed better on DE and DM than on DB and DP classes. DB class was the most difficult class for differentially expressed gene identification; however, it is probably a rare case scenario in real data. MAST, SCDE, and DESeq were the best tools in terms of precision in all the four classes, with recall higher than 75% for DE and DM classes, but lower than 30% for DP and DB classes.

Computational Performance

In terms of computational performance, all tools performed reasonably well but D^3E. Bulk tools had some of the shortest execution time, as they did not include any heavily time-consuming single-cell modeling step. Among the assessed scRNA-seq tools MAST, SCDE and Monocle support parallel execution, which significantly shorten the computational time needed to perform the analysis. In particular, Monocle becomes ∼7 times faster using eight cores.

Limitations of the Study and Concluding Remarks

Globally, considering our test design, none tool emerged as the best one. Some of the scRNA-seq tools (MAST and SCDE) performed best in terms of precision but had a drop in performance in terms of recall. Others (Monocle and D^3E) had an average trade-off between precision and recall but did not reach the desired cut-offs for any of these measures. All tools performed well with Grün datasets, regarding the ability in detecting true negatives, with the exception of D^3E, which reported a number of DEGs. Finally, bulk methods showed comparable performance with respect to single-cell tools, also in handling the multimodality of simulated data.

Even if our results are encouraging, they are still preliminary and there are some limitations of our approach. The analysis on synthetic datasets is limited to the two-class comparison, with differential expressed genes belonging to four differential distributions, but, for example, the dropout component was not considered in the data simulation. This could partially explain why the performance of bulk methods does not differ much from those of single-cell tools, and could be an interesting aspect to investigate more in depth. Anyway, in the tested real dataset, where the dropout phenomenon could be somehow present, the performance of the bulk methods is still comparable to that of single-cell tools. This could suggest that the modeling of the dropout component has a minor role in the accuracy of differential expression analysis.

Together with the dropout phenomenon, in future works it would be interesting to consider aspects such as different preprocessing strategies and normalization techniques, studying the effects of these steps on the accuracy of single-cell differential expression analysis.

AUTHOR CONTRIBUTIONS

AD performed acquisition and analysis of the data, interpretation and drafting of manuscript. GB performed analysis and interpretation of the data and drafting of manuscript. The conception of the study and design was performed by BD, who also performed drafting and critical revision of the manuscript.

FUNDING

This research is supported by University of Padova ex60%, CPDR150320/15 ("Systems biology approach to single cell RNA sequencing") and PRAT 2010 CPDA101217 ("Models of RNA sequencing data variability for quantitative transcriptomics") grants.

REFERENCES

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

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17, 63. doi: 10.1186/s13059-016-0927-y

Brennecke, P., Reyes A., Pinto S., Rattay K., Nguyen M., Küchler R., et al. (2015). Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat. Immunol.* 16, 933–941. doi: 10.1038/ni.3246

Chubb, J. R., Trcek, T., Shenoy, S. M., and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Curr. Biol.* 16, 1018–1025. doi: 10.1016/j.cub.2006.03.092

Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D^3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinform.* 17:110. doi: 10.1186/s12859-016-0944-6

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. doi: 10.1186/s13059-015-0844-5

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods* 11, 637–640. doi: 10.1038/nmeth.2930

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J. B., Lönnerberg, P., Linnarsson, S. et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167. doi: 10.1101/gr.110882.110

Jaakkola, M. K., Seyednasrollah, F., Mehmood, A., and Elo, L. L. (2016). Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* doi: 10.1093/bib/bbw057. [Epub ahead of print].

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005

Korthauer, K. D., Chu, L-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., et al. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* 17, 222. doi: 10.1186/s13059-016-1077-y

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583–621. doi: 10.1080/01621459.1952.10483441

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

Moliner, A., Enfors, P., Ibáñez, C. F., and Andäng, M. (2008). Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. *Stem Cells Dev.* 17, 233–243. doi: 10.1089/scd.2007.0211

Peccoud, J., and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* 48, 222–234. doi: 10.1006/tpbi.1995.1027

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4:e309. doi: 10.1371/journal.pbio.0040309

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

Robinson, M. D., and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332. doi: 10.1093/biostatistics/kxm030

Seyednasrollah, F., Rantanen, K., Jaakkola, P., and Elo, L. L. (2015). ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.* 44:e1. doi: 10.1093/nar/gkv806

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme J. T., Raychowdhury, R., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498, 236–240. doi: 10.1038/nature12172

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3:3. doi: 10.2202/1544-6115.1027

Smyth, G. K., and Verbyla, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Stat. Soc. Ser. B* 58, 565–572.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19, 335–346. doi: 10.1038/nn.4216

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.* 17:70. doi: 10.1186/s13059-016-0930-3

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

Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *J. Econ. Entomol.* 39:269. doi: 10.1093/jee/39.2.269

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 Genetics

ORIGINAL RESEARCH
published: 12 May 2017
doi: 10.3389/fgene.2017.00057

Non-coding RNAs in the Ovarian Follicle

Rosalia Battaglia[1], Maria E. Vento[2], Placido Borzì[2], Marco Ragusa[1], Davide Barbagallo[1], Desirée Arena[1], Michele Purrello[1] and Cinzia Di Pietro[1]*

[1] Section of Biology and Genetics G. Sichel, Department of Biomedical and Biotechnological Sciences, University of Catania, Catania, Italy, [2] IVF Unit, Cannizzaro Hospital, Catania, Italy

The mammalian ovarian follicle is the complex reproductive unit comprising germ cell, somatic cells (Cumulus and Granulosa cells), and follicular fluid (FF): paracrine communication among the different cell types through FF ensures the development of a mature oocyte ready for fertilization. This paper is focused on non-coding RNAs in ovarian follicles and their predicted role in the pathways involved in oocyte growth and maturation. We determined the expression profiles of microRNAs in human oocytes and FF by high-throughput analysis and identified 267 microRNAs in FF and 176 in oocytes. Most of these were FF microRNAs, while 9 were oocyte specific. By bioinformatic analysis, independently performed on FF and oocyte microRNAs, we identified the most significant Biological Processes and the pathways regulated by their validated targets. We found many pathways shared between the two compartments and some specific for oocyte microRNAs. Moreover, we found 41 long non-coding RNAs able to interact with oocyte microRNAs and potentially involved in the regulation of folliculogenesis. These data are important in basic reproductive research and could also be useful for clinical applications. In fact, the characterization of non-coding RNAs in ovarian follicles could improve reproductive disease diagnosis, provide biomarkers of oocyte quality in Assisted Reproductive Treatment, and allow the development of therapies for infertility disorders.

Keywords: human oocyte, ovarian follicle, follicular fluid, microRNAs, lncRNAs

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Yang Dai,
University of Illinois at Chicago, USA
Francesco Russo,
University of Copenhagen, Denmark

****Correspondence:***
Rosalia Battaglia
rosaliabattaglia04@gmail.com

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

Received: *17 February 2017*
Accepted: *26 April 2017*
Published: *12 May 2017*

Citation:
Battaglia R, Vento ME, Borzì P, Ragusa M, Barbagallo D, Arena D, Purrello M and Di Pietro C (2017) Non-coding RNAs in the Ovarian Follicle. Front. Genet. 8:57. doi: 10.3389/fgene.2017.00057

INTRODUCTION

Infertility, the inability to conceive and have children, is estimated to affect as many as 186 million people worldwide, and, consequently, represents an important social and medical problem (Inhorn and Patrizio, 2015). It can be considered a complex disease: first of all it is related to two different genomes (oocyte and sperm quality represent major determining factors in reproductive success), depends on endometrium receptivity and is influenced by several environmental factors (Koot and Macklon, 2013; Hart, 2016). In the same way as cancer, cardiovascular and neurodegenerative diseases, discovering a contributing factor and characterizing its involvement, is a difficult undertaking because the effect of any single factor may be obscured or confounded by other contributing factors (Manolio et al., 2009). To add to this complexity, male or female gametogenesis, fertilization and implantation, are regulated by different complex biological pathways and involve many molecules as mRNAs, non-coding RNAs, and proteins. In female gametogenesis, oocyte competence develops through protracted and complex processes beginning during embryonic life and ending at the moment that the MII oocyte is ovulated

(Hutt and Albertini, 2007). During embryonic life, primordial germ cells (PGCs) migrate to the genital ridge, proliferate by mitosis transforming into primary oocytes. Primary oocytes enter in meiosis and become arrested in prophase I at the diplotene stage (dyctiate). At this time the oocyte is enclosed in a specialized lineage of ovarian somatic cells, pre-granulosa cells, to form a primordial follicle (Zuccotti et al., 2011). The primordial follicle pool, produced during embryonic life, represents the woman's ovarian reserve (Reddy et al., 2010). After puberty, some primary follicles are cyclically recruited and develop through primary, secondary and antral stages. Inside antral follicles, in response to hormonal signaling, the oocytes are stimulated to resume meiosis. Most of the antral follicles undergo apoptosis, whereas only one, the dominant Graafian follicle, ovulates to release the mature egg, ready for fertilization (Sun et al., 2009; Zuccotti et al., 2011). At this stage, the follicles consist of the germ cell in the metaphase of the second meiotic division (MII oocytes), in a fluid-filled cavity called the antrum and different layers of somatic cells, the cumulus cells (CC) surrounding the oocytes and the granulosa cells (GC) as walls of follicle (Russell and Robker, 2007; Rodgers and Irving-Rodgers, 2010). Follicle development and oocyte maturation are strictly associated, in fact, the proliferation and the differentiation of somatic follicular cells occur in synchrony with the maturing oocyte, mediated by a constant exchange of signals between somatic cells and the germ cell (Russell and Robker, 2007). The cross-talk between the oocyte and somatic follicular cells occurs by gap-junctions established between the oocyte and CCs and through the Follicular Fluid (FF) accumulated inside the antrum (Rodgers and Irving-Rodgers, 2010). FF consists of a complex mixture of nucleic acids, proteins, metabolites, and ions, which are secreted by the oocytes and somatic cells (Revelli et al., 2009). Recently, it has been demonstrated that in human FF, microRNAs (miRNAs), carried by extracellular vesicles (EVs) such as microvesicles and exosomes are present (Santonocito et al., 2014). The discovery of mechanisms of autocrine and paracrine communication mediated by miRNAs, inside the ovarian follicle, has revealed that these non-coding RNAs represent important regulators inside the pathways involved in folliculogenesis and oocyte maturation (Santonocito et al., 2014; Di Pietro, 2016). Consequently, their characterization could improve our knowledge about female gametogenesis, and could pinpoint new molecules involved in reproductive disorders allowing the formulation of new therapeutic approaches.

The aim of this paper was to identify the miRNAs in the follicular microenvironment and position them within the different compartments of the ovarian follicle identifying the pathways regulated by their targets and the long non-coding RNAs (lncRNAs) possibly involved in oocyte growth and maturation.

MATERIALS AND METHODS

Ethics Statement

The patients, included in IVF programs, signed an informed consent (in accordance with the Declaration of Helsinki) to participate in the research project, which comprised the use of collected FF and surplus MII oocytes. The study on human MII oocytes was approved by the Institutional Ethical Committee Catania 1.

Sample Collection

Human oocytes and FF samples were collected from healthy women (without any ovarian pathology), ≤38 years old, undergoing intracytoplasmic sperm injections (ICSI) (Santonocito et al., 2014). FF of individual follicles was kept separated until decumulation of the oocytes to collect only the FF in which nuclear mature oocytes (metaphase II) had been identified. A total of 12 mature MII oocytes, two oocytes per woman, showing normal morphology, and 6 pools of FF were collected from individual follicles of 6 and 15 healthy women respectively, and used for miRNA expression profile analysis. The six couples of MII oocytes retrieved were separately placed in six independent Eppendorf tubes and rinsed in RNase-free water several times to remove any trace of cell culture medium. The oocytes were transferred to PCR tubes in 2 μl water and stored at −80°C before RNA extraction. Human FF samples were centrifuged for 20′ at 2,800 rpm at 4°C to remove follicular cell residue and any traces of blood; the supernatant was immediately transferred into a new Eppendorf tube and stored at −80°C for further analysis.

RNA Isolation, Reverse Transcription and miRNA Profiling by Taqman Low Density Array

For fluids, miRNA isolation was performed by using Qiagen miRNeasy Mini Kit (Qiagen GmbH), according to the Qiagen Supplementary Protocol for the purification of small RNAs from serum and plasma and finally eluted in a 30 μl volume of RNAse-free water. Oocytes were incubated, after adding water (2 μl), for 1′ at 100°C according to previously published protocols with minor modifications (El Mouatassim et al., 1999; Di Pietro et al., 2008; Battaglia et al., 2016) in order to release nucleic acids. Samples (3 μl of total RNA from hFF and human MII oocytes), were retrotranscribed and preamplified. Amplified products were loaded onto microfluidic cards of the TaqMan Human MicroRNA Array A v2.0 (Applied Biosystems). To prepare the real time PCR reaction mix, 9 μl of undiluted pre-amplification product was added to 450 μl of 2X TaqMan Universal PCR Master mix, no AmpErase UNG (Applied Biosystems) and nuclease free water was added to a final volume of 900 μl. 100 μl of PCR reaction mix was loaded onto 384-well TaqMan Low Density Human MicroRNA array cards (TLDA). The qRT-PCR reaction was carried out according to the manufacturer's instructions in a 7900HT Fast Real Time PCR System (Applied Biosystems).

Analysis of miRNA Expression Data

miRNA expression profiles were analyzed using real-time RQ Manager software v1.2 (Applied Biosystems). For relative quantification of miRNAs we filtered miRNAs having Ct values below 37 and detected in all biological replicates. Statistically

significant miRNA differences were identified by Significance of Microarrays Analysis (SAM)[1], applying a two-class paired test among ΔCt of FF and oocytes samples by using a *p*-value based on 100 permutations; imputation engine: K-nearest neighbors, 10 neighbors; false discovery rate < 0.15 (Ragusa et al., 2014; Battaglia et al., 2016; Fendler et al., 2017). miRNA expression changes were calculated by applying the $2^{-\Delta\Delta CT}$ method and using the average Ct of each plate as endogenous controls. We accepted only DE miRNAs common to least two SAM tests as reliable. Expression data in the Result section are shown as natural logarithms of relative quantity (RQ) values and the error was estimated by evaluating the $2^{-\Delta\Delta Ct}$ equation using ΔΔCt plus SD and ΔΔCt minus SD (Livak and Schmittgen, 2001). The expression data have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE98103.[2]

miRNA Target Prediction, Go and Pathway Analysis

In order to gain insights into biological processes (BP) regulated by FF and oocyte miRNAs we retrieved validated targets by using miRTarBase v6.0[3] to select targets experimentally verified in humans with strong validation methods (Reporter assay, Western blot and RT-qPCR). Afterward, we explored miRNA target expression through the comparison with protein coding genes expressed in ovarian follicles in humans, available from OKdb[4], and then we analyzed their Gene Ontologies (GO) and Pathways by the Panther classification system v10.0.[5] The statistical overrepresentation test was executed and the Bonferroni correction for multiple testing was used to correct the *P*-value. GOs and molecular pathways with a *P*-value < 0.05 were chosen. Target genes of miRNAs commonly expressed in hFF and oocytes were subsequently used in a signaling pathway enrichment analysis in Diana-miRPath v3.0.[6] The list of genes expressed in the human ovarian follicle was imported into Diana-miRPath to carry out pathway analysis of experimentally validated miRNA gene targets, according to the Kyoto Encyclopedia of Genes and Genomes (KEGG). The FDR method was implemented to select the biological pathways with a threshold of significance defined by $P < 0.05$ and a microT threshold of 0.8.

Prediction of lncRNAs Implicated in miRNA Regulation

In order to explore whether lncRNAs might have a regulatory function in oocyte maturation and early embryo development, we searched for experimentally verified miRNA-lncRNA interactions on LncBase v.2[7] (Paraskevopoulou et al., 2016). The identification of miRNAs that are interacting with lncRNAs was performed by selecting those experimentally verified in homo sapiens (by Northern Blot, luciferase reporter assay and qPCR), and found expressed in the ovary, embryo and ESCs, with a prediction score above the 75th percentile. To have a better understanding of lncRNA functions in human oocytes we drew interactions among candidate molecules using information from literature data and NPInter v3.0[8].

RESULTS

Profiling of microRNAs in Ovarian Follicle Compartments

Using TaqMan Low Density Array (TLDA) technology, we determined the expression profile of 384 miRNAs on 6 pools of FF and 6 pools of MII oocytes from healthy women. We identified 267 miRNAs in FF and 176 in oocytes. In order to characterize the miRNA content in the respective follicle components, we compared the sets of identified miRNAs and found 118 miRNAs (Common-miRNAs), including small nuclear RNA U6, in both components (**Figure 1A**). Moreover, by qualitative analysis, we detected a subsets of miRNAs specifically expressed in FF that were undetected in oocytes, and *vice versa*. In particular, we identified 158 miRNAs only in the FF compartment (FF-miRNAs), whereas we found 9 miRNAs exclusively expressed in human MII oocytes (O-miRNAs) (**Figure 1A**). In order to explore their role in ovarian follicle maturation, we compared the validated targets of the identified miRNAs with protein coding genes expressed in the human ovary. Of the 2,067 target genes ∼39% were known to be expressed within different follicle cell types and involved in different aspects of follicle development: oocyte maturation (29.71%), ovulation (8.52%) and antral follicle growth (9.74%) (**Figure 1B**). Moreover, we identified 277 miRNAs (FF-miRNAs and Common-miRNAs) with an overlap of 82% with miRNAs annotated on web-based resource ExoCarta. On the other hand, miR-515-5p, miR-519c-3p, miR-520d-5p miR-548a-3p, and miR-548c-3p, specifically expressed in oocyte, are not found incorporated in exosome vesicles.

Gene Ontology and Pathway Analysis of FF-miRNAs and O-miRNAs

Gene ontologies and Pathway analysis were independently performed on FF-miRNA and O-miRNA validated target genes. According to the number of miRNAs identified, the number of mRNA targets is quite different in the two compartments (1,099 FF-miRNA targets and 19 O-miRNA targets): for this reason the data cannot be considered all together. The most significant BP involving the targets of both FF-miRNAs and O-miRNAs are related to cellular response to stimulus, development and the regulation of cellular processes (**Figures 2A,C**). FF-miRNAs were statistically more represented in developmental processes, cell differentiation,

[1]http://www.tm4.org

[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98103

[3]http://mirtarbase.mbc.nctu.edu.tw/

[4]http://okdb.appliedbioinfo.net/

[5]http://pantherdb.org

[6]http://snf-515788.vm.okeanos.grnet.gr/

[7]http://www.microrna.gr/LncBase

[8]http://www.bioinfo.org/NPInter/

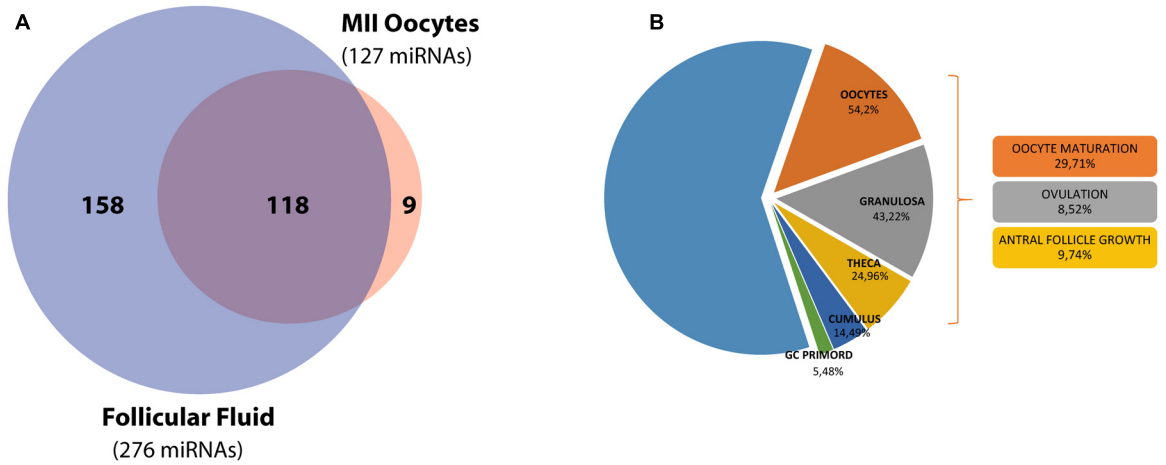


FIGURE 1 | miRNA distribution in the ovarian follicle and functional analysis of the target genes. (A) Venn diagram shows the overlap between miRNA sets in hFF and the mature MII oocyte. **(B)** A second diagram shows miRNA target expression and function inside the ovarian follicle.

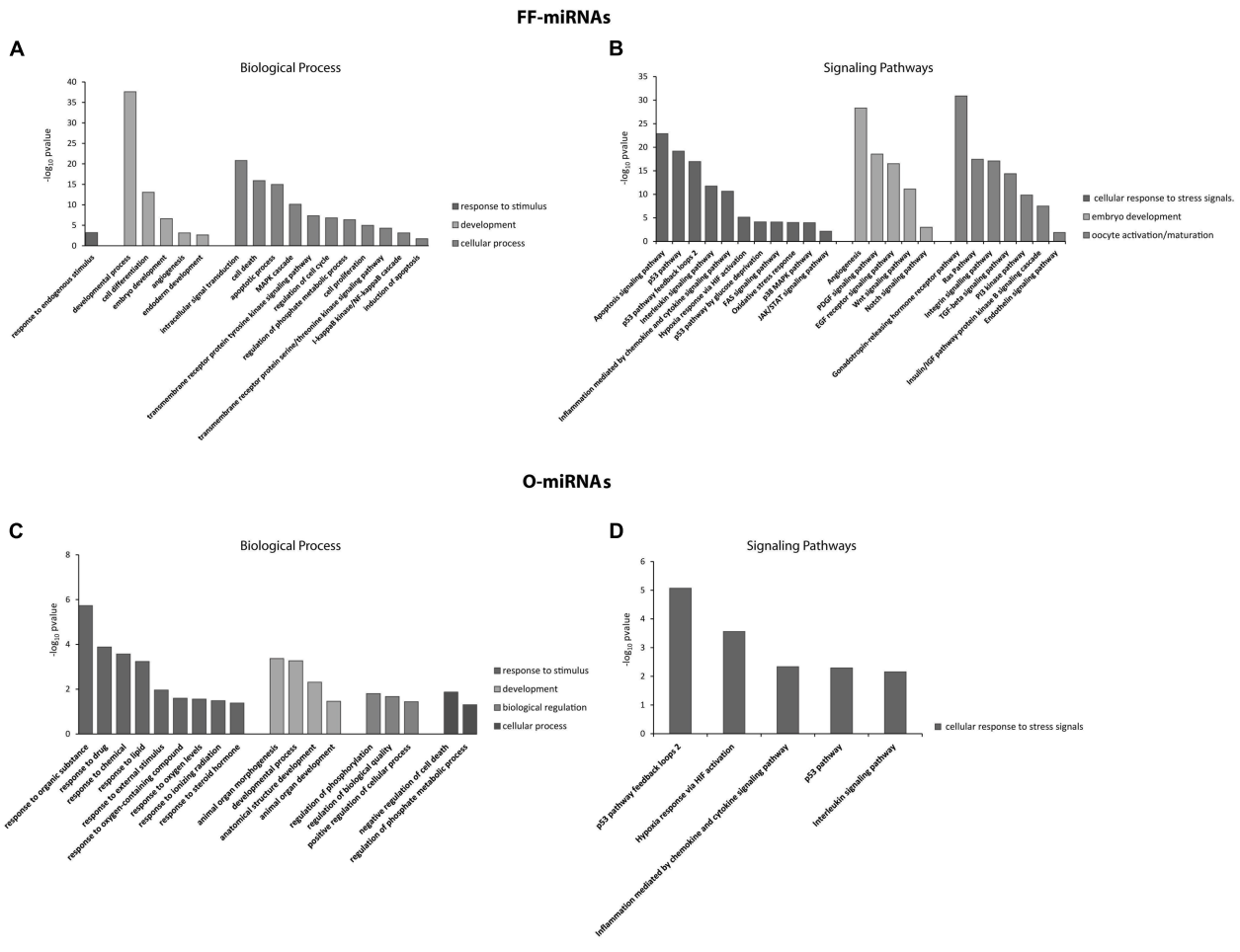


FIGURE 2 | Significant overrepresentation of GOs, in terms of Biological Processes, and signaling pathways for miRNAs identified in hFF (FF-miRNAs) and the MII oocytes (O-miRNAs) are shown in (A,B) and **(C,D)**, respectively. The significance values are reported as -log10 (*P*-value).

intracellular signal transduction, cell death, apoptotic processes and the response to endogenous stimuli than O-miRNAs (**Figures 2A,C**). Conversely, O-miRNAs showed a significant enrichment of target genes in GO terms related to the response to organic substances, drugs, chemicals, lipids, external stimuli, oxygen containing compounds and other processes such as organ morphogenesis, developmental processes, regulation of biological quality and negative regulation of cell death (**Figures 2A,C**). The most significant pathways are regulated by both FF-miRNAs and O-miRNAs and are associated with the cellular response to stress signals, such as p53 feedback loops 2, p53, Hypoxia response via HIF activation, inflammation mediated by chemokines and cytokines, and interleukin signaling pathway (**Figures 2B,D**).

Quantification and Pathway Analysis of Common-miRNAs

Finally, because it is not possible to establish the specific cell type from which the Common-miRNAs originate, we compared the expression profiles of 118 miRNAs, co-expressed in FF and oocytes. The heat map diagram in **Figure 3A**

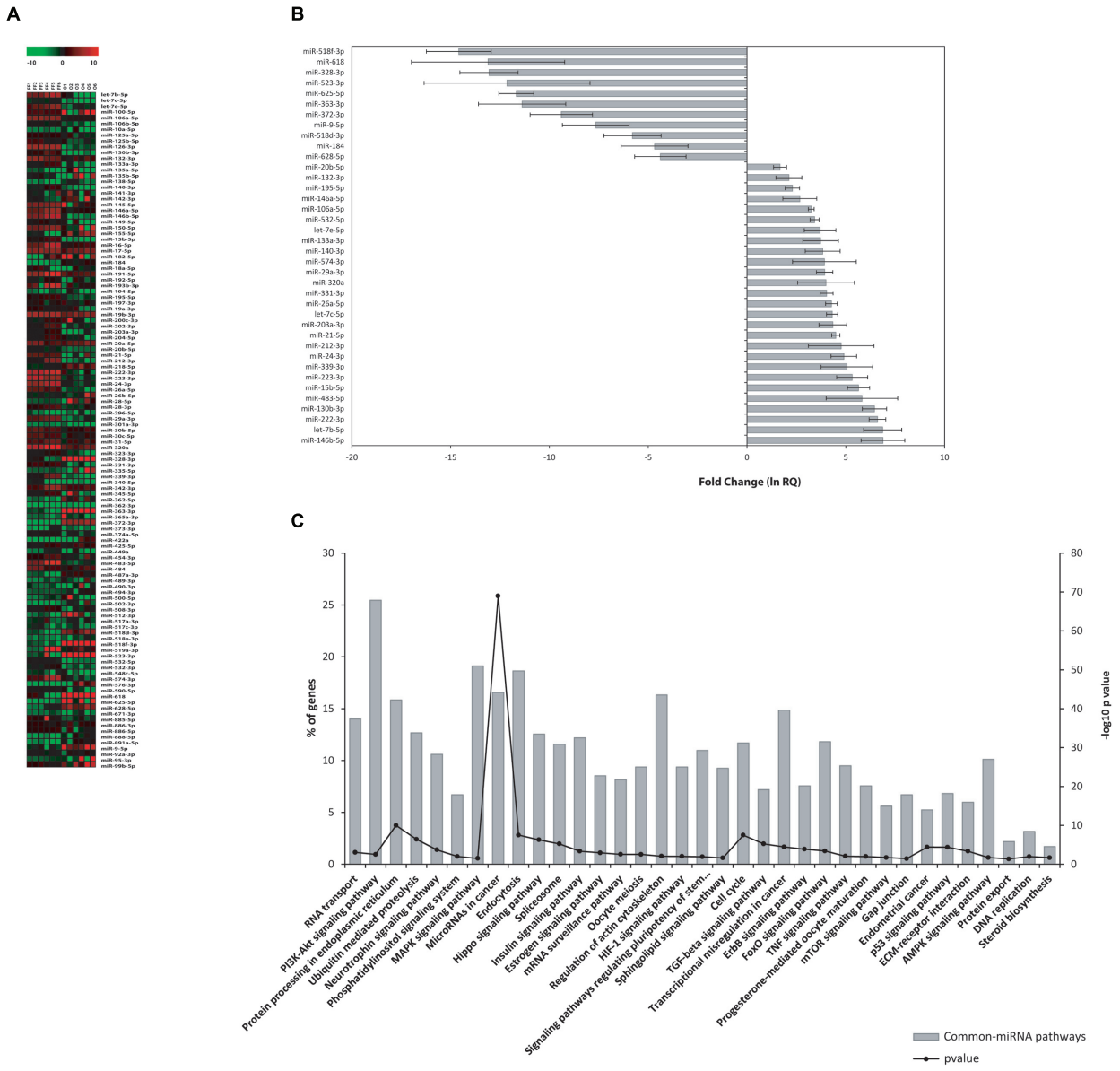


FIGURE 3 | miRNA expression in hFF and human oocytes. (A) Heat map of normalized miRNA expression data (–DCT values) of 118 Common-miRNAs for hFF and oocyte samples. The red and green colors represent up and down regulated miRNA expression levels, respectively. Equally expressed miRNAs are indicated in black. **(B)** Relative expression levels of 27 and 11 miRNAs that were differentially expressed between FF and oocytes. **(C)** Signaling pathway enrichment analysis for common miRNAs with KEGG against listed target genes. The probability values are reported as –log10 (*P*-value).

shows miRNA normalized expression levels across the different sample types (**Figure 3A**). SAM analysis revealed 38 miRNAs displaying statistical significant differences between human FF and MII oocytes (**Figure 3B** and **Supplementary Table S1**). Approximately 16% of the Common-miRNAs showed no significant variation between the two ovarian follicle components. On the contrary, we found 27 miRNAs significantly up-regulated in human FF (with respect to oocytes). Interestingly, let-7b-5p, miR-15b-5p, miR-24-3p, miR-130b-3p, miR-146b-5p, miR-212-3p, miR-222-3p, miR-223-3p, miR-339-3p and miR-483-5p showed expression fold changes higher than 100-fold (ln RQ > 4.7) compared to oocytes (**Figure 3B** and **Supplementary Table S1**). In contrast, 11 miRNAs (miR-9-5p, miR-184, miR-328-3p, miR-363-3p, miR-372-3p, miR-518d-3p, miR-518f-3p miR-523-3p, miR-618, miR-625-5p, and miR-628-5p) were significantly up-regulated in human oocytes (with respect to FF) (**Figure 3B** and **Supplementary Table S1**). Finally, Common-miRNAs, analyzed by KEGG analysis, showed their involvement in regulating 48.9% of genes expressed in the human ovarian follicle. Most of the significant pathways are shared with FF-miRNAs but we also found many pathways involved in oocyte maturation, regulation of pluripotency in stem cells and miRNAs in cancer. Moreover, protein processing in the endoplasmic reticulum, endocytosis, gap junction and protein export are also well represented (**Figure 3C**).

Long Non-coding RNAs in Oocytes

Prediction of experimentally verified interactions of lncRNAs with the 16 O-miRNAs was implemented by DIANA-LncBase database. A total of 41 lncRNAs were significantly associated with 9 O-miRNAs (**Table 1**). Each of them is expressed in embryonic cells, in the ovary or in the placenta and are involved in several human pathologies. We found that 17 lncRNAs were annotated as long intergenic RNAs (lincRNAs), 9 as antisense transcripts of genes, 6 as transcript isomorphs, 5 as retained introns, 4 as sense transcripts and 11 as circular RNAs (circRNAs) in the database circBase[9] (**Table 1**). Network analysis showed the interactions among lncRNAs and miRNAs involved in stemness, RNA maturation and epigenetics (**Figure 4**).

DISCUSSION

Regulatory non-coding RNAs (ncRNAs) control different points of gene expression including chromatin architecture, epigenetics, transcription, RNA splicing, editing, translation and turnover. They are involved in every physiological process and consequently their sequences or expression alterations cause or contribute to different human diseases (Bartel, 2004; Croce, 2009; Fu, 2014). RNA regulatory networks include miRNAs, other classes of small regulatory RNAs and thousands of longer transcripts, named long non-coding RNAs that can be categorized in sense, antisense, bidirectional, intronic, and intergenic transcripts. Surely, the miRNA world is becoming increasingly well known to researchers, but for lncRNAs most of their functions are unknown, even if different papers have demonstrated their role in cell physiology and human pathologies. Their presence in biological fluids, as well as their expression profiles associated with specific human phenotypes, open up the possibility of using them, especially miRNAs, as molecular markers of human diseases (Weber et al., 2010; Cortez et al., 2011). Moreover, their presence in EVs, identifying miRNAs as molecular tools of communication among different cells, gives the possibility to plan specific and individualized therapies (Cortez et al., 2011; Kosaka et al., 2013; Matsui and Corey, 2016).

In reproductive biology, throughout the last decade, the role of miRNAs emerged in an important way and different studies attempted to associate specific miRNA expression profiles to oocyte quality, in granulose and cumulus cells and in FF (Li et al., 2015; McGinnis et al., 2015). Hence, great efforts were made to find promising molecular markers, in order to select the best oocytes to use in In Vitro Fertilization protocols and provide possible therapies to improve oocytes quality (Li et al., 2015; McGinnis et al., 2015).

Unfortunately, the ovarian follicle constitutes, to date, a difficult model to study for different reasons. It is made up of different cell types and represents the functional unit that ensures proper oocyte maturation by processes that begin during the embryonic stage and continue during a woman's life until ovulation (Hutt and Albertini, 2007). The mature ovarian follicle and ovulation represent the results of different molecular processes prolonged in time and influenced by genetic background, environment and a woman's life style. To analyze the single components of this complex unit, at the final stage, (MII oocyte, granulosa or cumulus cells, FF) does not provide all the information we need to fully understand the oogenesis and associate specific markers to specific phenotypes. Moreover, in humans, it is not possible to perform functional studies because of ethical limits.

The aim of this paper was to investigate the role of miRNAs and lncRNAs in human ovarian follicles, trying to establish, as far as possible, their potential role within the different components of the follicle.

Firstly, we identified 285 miRNAs inside human ovarian follicles (**Figure 1**). Interestingly, about 39% of their validated targets are expressed in different cellular components of the ovarian follicle; they are abundant in oocytes (54.2%) and predominantly involved in oocyte maturation (29.71%). 118 miRNAs (more than 40%) were shared by FF and the mature MII oocyte, while FF-miRNAs fraction was larger than the miRNAs exclusively found in oocytes. We propose that the 158 miRNAs absent in MII oocytes and exclusively present in FF have been transcribed by somatic follicular cells. Subsequently secreted in FF, these miRNAs could act as paracrine factors for the different somatic cells and regulate follicular growth. As expected, 82% of FF-miRNAs have been described in exosomes, in fact, according to previous studies, specific miRNAs are preferentially sorted into vesicles (Batagov et al., 2011). On the other hand the 9 miRNAs, absent in FF and exclusively present in the oocyte (O-miRNAs), could represent maternal RNAs, that the germ cells

[9]http://www.circbase.org

TABLE 1 | LncRNAs in human oocytes.

miRNA	lncRNA	Prediction score	Biotype	Chromosome
miR-9-5p	CTB-89H12.4	0.994	Retained_intron*	5
	TUG1	0.985	Antisense*	22
	RP11-793H13.8	0.962	Retained_intron	12
	SNHG14	0.934	Antisense*	15
	RP11-314B1.2	0.930	lincRNA	2
	RP11-383J24.6	0.868	lincRNA	8
	RP11-436K8.1	0.858	lincRNA	1
	TMEM256-PLSCR3	0.857	Sense	17
	CTD-2368P22.1	0.808	Retained_intron	19
	AC007246.3	0.791	Antisense	2
	XLOC_000918	0.789	Transcript isomorph	1
	RNF144A-AS1	0.780	Antisense	2
	RP4-717I23.3	0.779	lincRNA*	1
	IPO11-LRRC70	0.777	Sense	5
	XLOC_008152	0.771	Transcript isomorph	X
	AC093323.3	0.759	lincRNA	4
	DPP10-AS1	0.755	Antisense	2
miR-136-5p	MALAT-1	0.937	lincRNA*	11
	GAS5	0.903	Retained_intron*	1
	RP11-383J24.6	0.802	lincRNA	8
	RP3-468B3.2	0.781	lincRNA	6
miR-363-3p	XIST	0.919	lincRNA*	X
	XLOC_008152	0.914	Transcript isomorph	X
	RP11-67L2.2	0.797	lincRNA	3
	OIP5-AS1	0.773	Antisense*	15
	CASC7	0.769	lincRNA	8
	CTD-3099C6.9	0.751	Sense_intronic	19
miR-519c-3p	CTB-89H12.4	0.990	Retained_intron*	5
	LOC388692	0.871	lincRNA	1
	GABPB1-AS1	0.792	Antisense	15
miR-520d-5p	CTB-89H12.4	0.974	Retained_intron*	5
	CASC7	0.935	lincRNA	8
	NORAD	0.891	lincRNA	20
miR-548a-3p	CASC7	0.960	lincRNA	8
	CTB-89H12.4	0.890	Retained_intron*	5
	LINC01355	0.944	lincRNA	1
	MALAT-1	0.865	lincRNA*	11
	NEAT1	0.950	lincRNA	11
	RP6-24A23.7	0.761	Sense-overlapping	X
	XIST	0.933	lincRNA*	X
	XLOC_006828	0.776	Transcript isomorph	8
	XLOC_011568	0.770	Transcript isomorph	15
	ZFAS1	0.761	Antisense*	20
miR-618	NORAD	0.857	lincRNA	20
	OIP5-AS1	0.884	Antisense*	15
	SNHG1	0.753	Retained_intron *	11
	SNHG14	0.770	Antisense*	15
	SNORD116-20	0.894	lincRNA *	15
miR-625-5p	CASC7	0.900	lincRNA	8
	KMT2E-AS1	0.834	lincRNA	7
	SRRM2-AS1	0.794	Antisense	16
	CTD-2619J13.14	0.769	lincRNA	19
	XLOC_012097	0.752	Transcript isomorph	17
miR-628-5p	OIP5-AS1	0.769	Antisense*	15

*Diana-LncBase prediction of miRNA-lncRNA interactions. *Indicates lncRNAs annotated as circRNA in the database circBase (www.circbase.org).*

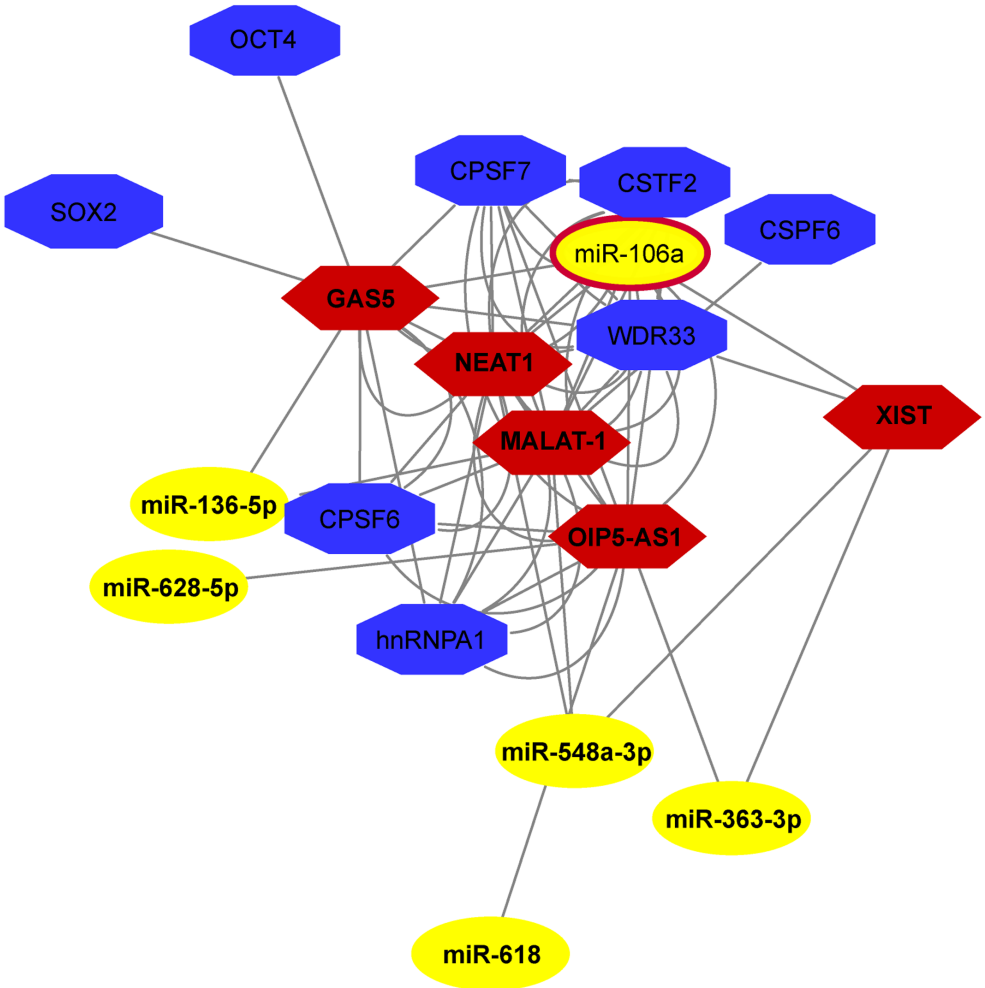


FIGURE 4 | ncRNA network in human MII oocytes. The Network, drawn by NPInter v3.0, shows the interaction among lncRNAs and miRNAs involved in stemness, RNA maturation and epigenetics.

accumulate during their differentiation. In fact, during its growth the oocyte transcribes and stores mRNAs, miRNAs and probably long non-coding RNAs to use them during the first phase of development, before the activation of the embryo genome (Schier, 2007). Interestingly, we found miR-515-5p, miR-519c-3p, miR-520d-5p, miR-548a-3p, and miR-548c-3p specifically expressed in oocytes, and according to ExoCarta, these have not been found incorporated in exosome vesicles. Moreover, miR-515-5p, miR-519c-3p and miR-520d-5p, members of C19MC, which is a primate-specific cluster, seem to have a role in early embryo development during maternal-zygotic transition, when zygotic transcription starts and maternal mRNAs are degraded (Donker et al., 2012; Battaglia et al., 2016). Go analysis showed that development processes, regulation of cell cycle, signal transduction, and cell death represent BPs shared by oocytes and somatic follicular cells. In the same way, the p53 pathway is significant in both cell types. This confirmed our knowledge about the processes involved in oocyte maturation and carried out by both compartments (Hutt and Albertini, 2007). Apoptosis has been amply described in the mammalian ovary (Tilly, 2001). Cell death by apoptosis affects about 99% of primordial follicles present at birth in mammalian ovaries (Tiwari et al., 2015). The production of the mature oocyte ready for fertilization is a highly selective process: only a few follicles in the human ovary survive to complete their growth, only fully competent oocytes will be ovulated and only an embryo without major genetic alterations will be capable of uterine implantation. Apoptosis of granulosa cells reduces the number of recruited follicles and apoptotic machinery in MII oocytes selects viable embryos (Tilly, 2001; Santonocito et al., 2014). Interestingly, we found different BPs involved in the response to exogenous stimuli, significant only for O-miRNAs (**Figure 2C**). Among the different cells of the ovarian follicle the germ cell is unique: it must be able to respond and react to external stimuli more efficiently than somatic cells. It has been described that the maturing oocyte and early embryo are quite sensitive to exogenous stresses. Oocytes and early embryos can undergo physiological adaptations to environmental perturbations; these adaptations could influence

the embryo genome signature involving epigenetic modification (Latham, 2015).

As concerns the 118 shared miRNAs, not being able to pinpoint the cells producing them, we can assert that these miRNAs could mediate the communication between the oocyte and somatic cells. Eleven of them, more abundant in oocytes, could have been transcribed by germ cells and used as signaling for cumulus and granulose cells. This hypothesis is supported by the finding that miR-518d-3p, miR-518f-3p, and miR-523-3p, up-regulated in oocytes, are members of the C19MC cluster, in the same way of some O-miRNAs. These miRNAs could be transcribed together, some of them stored in oocytes, others packaged in vesicles and secreted in FF. A similar consideration can be made for miR-372 (up-regulated in oocytes) and miR-371 (O-miRNA). On the contrary, the miRNAs up-regulated in FF, especially let-7b-5p, miR-15b-5p, miR-24-3p, miR-130b-3p, miR-146b-5p, miR-212-3p, miR-222-3p, miR-223-3p, miR-339-3p and miR-483-5p, with fold change values higher than 100-fold, could be transcribed in somatic follicular cells and move to oocytes by means of exosomes. Enrichment analysis of validated target genes of Common-miRNAs showed a strong correlation with the maintenance of the primordial follicle quiescent stage, oocyte maturation, oocyte meiosis, development, cancer and stem cell related pathways. Moreover, a high representation of apoptosis signaling pathway, hypoxia response via HIF activation and oxidative stress response have been detected (**Figure 3C**).

To understand miRNA-mediated gene regulation, we investigated if other classes of non-coding RNAs play a role inside the ovarian follicle. LncRNAs can act as miRNA sponges, reducing their regulatory effect on mRNAs introducing an extra layer of complexity in the miRNA-target interaction network (Paraskevopoulou and Hatzigeorgiou, 2016). Moreover, the important role of lncRNAs in chromatin remodeling is well known, as well as the importance of this process in oocyte maturation, when, before meiosis resumption, transcriptional silencing is mediated, over all, by mechanisms involved in large-scale chromatin structure changes. The cellular and molecular pathways involved in these processes are today poorly understood and lncRNAs could play a major role (De La Fuente et al., 2004). Recently, lncRNAs were identified in granulosa and cumulus cells, oocytes and early embryos, and their role in oocyte and early embryo development has been suggested (Yan et al., 2013; Yerushalmi et al., 2014; Hamazaki et al., 2015; Xu et al., 2015). A comprehensive review on lncRNA functions in mammalian and in different species has been recently published (Taylor et al., 2015).

By using bioinformatic prediction we found 41 lncRNAs significantly correlated with 9 oocyte miRNAs (**Table 1**). Most of them are expressed in embryonic cells, in the ovary or in the placenta and are involved in several human pathologies. NEAT1 was found exclusively localized in paraspeckles and is a core component of these nuclear bodies involved in nuclear retention of mRNAs (Bond and Fox, 2009). Moreover, NEAT1 seems to be essential for the formation of the corpus luteum and the establishment of pregnancy in mice, although its precise molecular mechanism remains to be investigated (Nakagawa et al., 2014). Recently, increased levels of NEAT1 was found associated with placental dysfunction in Idiopathic Intrauterine Growth Restriction (IUGR) fetuses (Gremlich et al., 2014). Another lincRNA that was found to reside predominantly in the nucleus is the Metastasis associated lung adenocarcinoma transcript-1 (MALAT-1) that localizes to nuclear bodies known as nuclear *speckles* involved in pre-mRNA splicing (Lennox and Behlke, 2016). MALAT-1 down-regulation was described to impair proliferation, cell cycle, apoptosis, and migration of trophoblast cells involved in the preeclampsia (Chen et al., 2015). Another human lincRNA, the non-coding RNA activated by DNA damage (NORAD) is induced after DNA damage in a p53-dependent manner and plays a crucial role in maintaining genomic stability by sequestering PUMILIO proteins, which repress the stability and translation of mRNAs to which they bind (Lee et al., 2016). In view of the location and the function of these 3 lncRNAs in regulation of mRNA stability and translation, we suppose that in oocytes these could stabilize maternal RNAs, allowing their storage and use during early development, before the activation of embryo genome, as it has been described for cytoplasmic polyadenylation (Reyes and Ross, 2016). Not surprisingly, we found the X-inactive specific transcript (XIST) and the growth arrest specific 5 (GAS5), two lncRNAs involved in embryogenesis. It has been demonstrated that many aspects of embryogenesis seem to be controlled by ncRNAs, including the maternal–zygotic transition, the maintenance of pluripotency, the patterning of the body axes, the specification and differentiation of cell types and the morphogenesis of organs (Pauli et al., 2011). XIST, responsible for the mammalian X chromosome inactivation, is the first lncRNA expressed starting at the 4-cell stage of human preimplantation embryos, consistent with embryonic genome activation, and iPSC reprogramming (Briggs et al., 2015).

Encoded within introns GAS5 increases OCT4, NANOG and SOX2 by Nodal regulation and is directly regulated by these stemness factors in hESCs forming a circuit that promotes pluripotency (Xu et al., 2016). OIP5-AS1 is an antisense transcript of the Opa interacting protein 5 (OIP5) gene. The protein encoded by this gene localizes to centromeres, where it is essential for recruitment of CENP-A and it is required for centromeric heterochromatin organization. Expression of this gene is upregulated in several cancers, making it a putative therapeutic target.

Recently, a new regulatory circuitry in which RNAs can crosstalk with each other and modulate the biological function of miRNAs has been proposed (Cesana et al., 2011). LncRNAs that localize primarily in the nucleus (e.g., XIST, NEAT1, MALAT-1) have been described to physically interact with mature miRNAs (Leucci et al., 2013; Gernapudi et al., 2015; Yu et al., 2017). Several observations have shown the presence of mature miRNAs in the nucleus (Hwang et al., 2007; Rasko and Wong, 2017). In facts, miRNAs can be transported from the cytoplasm to the nucleus and act in an unconventional manner to regulate the biogenesis and functions of ncRNAs (Liang et al., 2013).

Among the identified lncRNAs in oocytes, NEAT1, MALAT-1, GAS5, XIST and OIP5-AS1 have been predicted as components of the same network (**Figure 4**). Even if the mechanisms of most of these lncRNAs remain unknown, and it remains to be seen

whether they can function within human ovarian follicle, these putative interactions lead us to hypothesize a possible role inside the female human germ cell.

CONCLUSION

Understanding the regulation of gene expression inside the ovarian follicle is important in basic reproductive research and could also be useful for clinical applications. In fact, the characterization of non-coding RNAs in ovarian follicles could improve reproductive disease diagnosis, provide biomarkers of oocyte quality in Assisted Reproductive Treatment, and develop therapies for infertility disorders.

AUTHOR CONTRIBUTIONS

CDP conceived and designed the study. RB performed the experiments. CDP and RB analyzed, interpreted the data and wrote the manuscript. DA contributed to experiments and the bioinformatics analysis. MV and PB participated in sample collection. MR and DB contributed to the analysis of data. MP contributed to the critical revision of the manuscript.

FUNDING

This work was partially supported by Bio-Nanotech Research and Innovation Tower grant BRIT PONa3_00136, University of Catania.

ACKNOWLEDGMENTS

The authors thank the Scientific Bureau of the University of Catania for language support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00057/full#supplementary-material

TABLE S1 | Differentially expressed (DE) miRNAs from qRT-PCR comparative analysis between human follicular fluid (hFF) and human mature MII oocytes.

REFERENCES

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Batagov, A. O., Kuznetsov, V. A., and Kurochkin, I. V. (2011). Identification of nucleotide patterns enriched in secreted RNAs as putative cis-acting elements targeting them to exosome nano-vesicles. *BMC Genomics* 12(Suppl. 3):S18. doi: 10.1186/1471-2164-12-S3-S18

Battaglia, R., Vento, M. E., Ragusa, M., Barbagallo, D., La Ferlita, A., Di Emidio, G., et al. (2016). MicroRNAs are stored in human MII oocyte and their expression profile changes in reproductive aging. *Biol. Reprod.* 95, 1–13. doi: 10.1095/biolreprod.116.142711

Bond, C. S., and Fox, A. H. (2009). Paraspeckles: nuclear bodies built on long noncoding RNA. *J. Cell Biol.* 186, 637–644. doi: 10.1083/jcb.200906113

Briggs, S. F., Dominguez, A. A., Chavez, S. L., and Reijo Pera, R. A. (2015). Single-cell XIST expression in human preimplantation embryos and newly reprogrammed female induced pluripotent stem cells. *Stem Cells* 33, 1771–1781. doi: 10.1002/stem.1992

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., et al. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147, 358–369. doi: 10.1016/j.cell.2011.09.028

Chen, H., Meng, T., Liu, X., Sun, M., Tong, C., Liu, J., et al. (2015). Long non-coding RNA MALAT-1 is downregulated in preeclampsia and regulates proliferation, apoptosis, migration and invasion of JEG-3 trophoblast cells. *Int. J. Clin. Exp. Pathol.* 8, 12718–12727.

Cortez, M. A., Bueso-Ramos, C., Ferdin, J., Lopez-Berestein, G., Sood, A. K., and Calin, G. A. (2011). MicroRNAs in body fluids-the mix of hormones and biomarkers. *Nat. Rev. Clin. Oncol.* 8, 467–477. doi: 10.1038/nrclinonc.2011.76

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

De La Fuente, R., Viveiros, M. M., Burns, K. H., Adashi, E. Y., Matzuk, M. M., and Eppig, J. J. (2004). Major chromatin remodeling in the germinal vesicle (GV) of mammalian oocytes is dispensable for global transcriptional silencing but required for centromeric heterochromatin function. *Dev. Biol.* 275, 447–458.

Di Pietro, C. (2016). Exosome-mediated communication in the ovarian follicle. *J. Assist. Reprod. Genet.* 33, 303–311. doi: 10.1007/s10815-016-0657-9

Di Pietro, C., Vento, M., Ragusa, M., Barbagallo, D., Guglielmino, M. R., Maniscalchi, T., et al. (2008). Expression analysis of TFIID in single human oocytes: new potential molecular markers of oocyte quality. *Reprod. Biomed. Online* 17, 338–349.

Donker, R. B., Mouillet, J. F., Chu, T., Hubel, C. A., Stolz, D. B., Morelli, A. E., et al. (2012). The expression profile of C19MC microRNAs in primary human trophoblast cells and exosomes. *Mol. Hum. Reprod.* 18, 417–424. doi: 10.1093/molehr/gas013

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210.

El Mouatassim, S., Guérin, P., and Ménézo, Y. (1999). Expression of genes encoding antioxidant enzymes in human and mouse oocytes during the final stages of maturation. *Mol. Hum. Reprod.* 5, 720–725.

Fendler, W., Malachowska, B., Meghani, K., Konstantinopoulos, P. A., Guha, C., Singh, V. K., et al. (2017). Evolutionarily conserved serum microRNAs predict radiation-induced fatality in nonhuman primates. *Sci. Transl. Med.* 9:eaal2408. doi: 10.1126/scitranslmed.aal2408

Fu, X. D. (2014). Non-coding RNA: a new frontier in regulatory biology. *Natl. Sci. Rev.* 1, 190–204.

Gernapudi, R., Wolfson, B., Zhang, Y., Yao, Y., Yang, P., Asahara, H., et al. (2015). MicroRNA 140 promotes expression of long noncoding RNA NEAT1 in adipogenesis. *Mol. Cell. Biol.* 36, 30–38. doi: 10.1128/MCB.00702-15

Gremlich, S., Damnon, F., Reymondin, D., Braissant, O., Schittny, J. C., Baud, D., et al. (2014). The long non-coding RNA NEAT1 is increased in IUGR placentas, leading to potential new hypotheses of IUGR origin/development. *Placenta* 35, 44–49. doi: 10.1016/j.placenta.2013.11.003

Hamazaki, N., Uesaka, M., Nakashima, K., Agata, K., and Imamura, T. (2015). Gene activation-associated long noncoding RNAs function in mouse preimplantation development. *Development* 142, 910–920. doi: 10.1242/dev.116996

Hart, R. J. (2016). Physiological aspects of female fertility: role of the environment, modern lifestyle, and genetics. *Physiol. Rev.* 96, 873–909. doi: 10.1152/physrev.00023.2015

Hutt, K. J., and Albertini, D. F. (2007). An oocentric view of folliculogenesis and embryogenesis. *Reprod. Biomed. Online* 14, 758–764.

Hwang, H. W., Wentzel, E. A., and Mendell, J. T. (2007). A hexanucleotide element directs microRNA nuclear import. *Science* 315, 97–100. doi: 10.1126/science.1136235

Inhorn, M. C., and Patrizio, P. (2015). Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. *Hum. Reprod. Update* 21, 411–426. doi: 10.1093/humupd/dmv016

Koot, Y. E., and Macklon, N. S. (2013). Embryo implantation: biology, evaluation, and enhancement. *Curr. Opin. Obstet. Gynecol.* 25, 274–279. doi: 10.1097/GCO.0b013e3283630d94

Kosaka, N., Yoshioka, Y., Hagiwara, K., Tominaga, N., Katsuda, T., and Ochiya, T. (2013). Trash or treasure: extracellular microRNAs and cell-to cell communication. *Front. Genet.* 4:173. doi: 10.3389/fgene.2013.00173

Latham, K. E. (2015). Endoplasmic reticulum stress signaling in mammalian oocytes and embryos: life in balance. *Int. Rev. Cell Mol. Biol.* 316, 227–265. doi: 10.1016/bs.ircmb.2015.01.005

Lee, S., Kopp, F., Chang, T. C., Sataluri, A., Chen, B., Sivakumar, S., et al. (2016). Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell* 164, 69–80. doi: 10.1016/j.cell.2015.12.017

Lennox, K. A., and Behlke, M. A. (2016). Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. *Nucleic Acids Res.* 44, 863–877. doi: 10.1093/nar/gkv1206

Leucci, E., Patella, F., Waage, J., Holmstrøm, K., Lindow, M., Porse, B., et al. (2013). microRNA-9 targets the long non-coding RNA MALAT1 for degradation in the nucleus. *Sci. Rep.* 3:2535. doi: 10.1038/srep02535

Li, Y., Fang, Y., Liu, Y., and Yang, X. (2015). MicroRNAs in ovarian function and disorders. *J. Ovarian Res.* 8, 51. doi: 10.1186/s13048-015-0162-2

Liang, H., Zhang, J., Zen, K., Zhang, C. Y., and Chen, X. (2013). Nuclear microRNAs and their unconventional role in regulating non-coding RNAs. *Protein Cell* 4, 325–330. doi: 10.1007/s13238-013-3001-5

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Matsui, M., and Corey, D. R. (2016). Non-coding RNAs as drug targets. *Nat. Rev. Drug Discov.* doi: 10.1038/nrd.2016.117 [Epub ahead of print].

McGinnis, L. K., Luense, L. J., and Christenson, L. K. (2015). MicroRNA in ovarian biology and disease. *Cold Spring Harb. Perspect. Med.* 5:a022962. doi: 10.1101/cshperspect.a022962

Nakagawa, S., Shimada, M., Yanaka, K., Mito, M., Arai, T., Takahashi, E., et al. (2014). The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a subpopulation of mice. *Development* 141, 4618–4627. doi: 10.1242/dev.110544

Paraskevopoulou, M. D., and Hatzigeorgiou, A. G. (2016). Analyzing MiRNA-LncRNA interactions. *Methods Mol. Biol.* 1402, 271–286. doi: 10.1007/978-1-4939-3378-5_21

Paraskevopoulou, M. D., Vlachos, I. S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., et al. (2016). DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.* 44, D231–D238. doi: 10.1093/nar/gkv1270

Pauli, A., Rinn, J. L., and Schier, A. F. (2011). Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.* 12, 136–149. doi: 10.1038/nrg2904

Ragusa, M., Statello, L., Maugeri, M., Barbagallo, C., Passanisi, R., Alhamdani, M. S., et al. (2014). Highly skewed distribution of miRNAs and proteins between colorectal cancer cells and their exosomes following Cetuximab treatment: biomolecular, genetic and translational implications. *Oncoscience* 1, 132–157. doi: 10.18632/oncoscience.19

Rasko, J. E., and Wong, J. J. (2017). Nuclear microRNAs in normal hemopoiesis and cancer. *J. Hematol. Oncol.* 10, 8. doi: 10.1186/s13045-016-0375-x

Reddy, P., Zheng, W., and Liu, K. (2010). Mechanisms maintaining the dormancy and survival of mammalian primordial follicles. *Trends Endocrinol. Metab.* 21, 96–103. doi: 10.1016/j.tem.2009.10.001

Revelli, A., Delle Piane, L., Casano, S., Molinari, E., Massobrio, M., and Rinaudo, P. (2009). Follicular fluid content and oocyte quality: from single biochemical markers to metabolomics. *Reprod. Biol. Endocrinol.* 7:40. doi: 10.1186/1477-7827-7-40

Reyes, J. M., and Ross, P. J. (2016). Cytoplasmic polyadenylation in mammalian oocyte maturation. *Wiley Interdiscip. Rev. RNA* 7, 71–89. doi: 10.1002/wrna.1316

Rodgers, R. J., and Irving-Rodgers, H. F. (2010). Formation of the ovarian follicular antrum and follicular fluid. *Biol. Reprod.* 82, 1021–1029. doi: 10.1095/biolreprod.109.082941

Russell, D. L., and Robker, R. L. (2007). Molecular mechanisms of ovulation: coordination through the cumulus complex. *Hum. Reprod. Update* 13, 289–312.

Santonocito, M., Vento, M., Guglielmino, M. R., Battaglia, R., Wahlgren, J., Ragusa, M., et al. (2014). Molecular characterization of exosomes and their microRNA cargo in human follicular fluid: bioinformatic analysis reveals that exosomal microRNAs control pathways involved in follicular maturation. *Fertil. Steril.* 102, 1751–1761. doi: 10.1016/j.fertnstert.2014.08.005

Schier, A. F. (2007). The maternal-zygotic transition: death and birth of RNAs. *Science* 316, 406–407.

Sun, Q. Y., Miao, Y. L., and Schatten, H. (2009). Towards a new understanding on the regulation of mammalian oocyte meiosis resumption. *Cell Cycle* 8, 2741–2747.

Taylor, D. H., Chu, E. T., Spektor, R., and Soloway, P. D. (2015). Long non-coding RNA regulation of reproduction and development. *Mol. Reprod. Dev.* 82, 932–956. doi: 10.1002/mrd.22581

Tilly, J. L. (2001). Commuting the death sentence: how oocytes strive to survive. *Nat. Rev. Mol. Cell Biol.* 2, 838–848.

Tiwari, M., Prasad, S., Tripathi, A., Pandey, A. N., Ali, I., Singh, A. K., et al. (2015). Apoptosis in mammalian oocytes: a review. *Apoptosis* 20, 1019–1025. doi: 10.1007/s10495-015-1136-y

Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., Huang, K. H., Lee, M. J., et al. (2010). The microRNA spectrum in 12 body fluids. *Clin. Chem.* 56, 1733–1741. doi: 10.1373/clinchem.2010.147405

Xu, C., Zhang, Y., Wang, Q., Xu, Z., Jiang, J., Gao, Y., et al. (2016). Long non-coding RNA GAS5 controls human embryonic stem cell self-renewal by maintaining NODAL signalling. *Nat. Commun.* 7:13287. doi: 10.1038/ncomms13287

Xu, X. F., Li, J., Cao, Y. X., Chen, D. W., Zhang, Z. G., He, X. J., et al. (2015). Differential expression of long noncoding RNAs in human cumulus cells related to embryo developmental potential: a microarray analysis. *Reprod. Sci.* 22, 672–678. doi: 10.1177/1933719114561562

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660

Yerushalmi, G. M., Salmon-Divon, M., Yung, Y., Maman, E., Kedem, A., Ophir, L., et al. (2014). Characterization of the human cumulus cell transcriptome during final follicular maturation and ovulation. *Mol. Hum. Reprod.* 20, 719–735. doi: 10.1093/molehr/gau031

Yu, H., Xue, Y., Wang, P., Liu, X., Ma, J., Zheng, J., et al. (2017). Knockdown of long non-coding RNA XIST increases blood-tumor barrier permeability and inhibits glioma angiogenesis by targeting miR-137. *Oncogenesis* 6, e303. doi: 10.1038/oncsis.2017.7

Zuccotti, M., Merico, V., Cecconi, S., Redi, C. A., and Garagna, S. (2011). What does it take to make a developmentally competent mammalian egg? *Hum. Reprod. Update* 17, 525–540. doi: 10.1093/humupd/dmr009

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 April 2017
doi: 10.3389/fpls.2017.00527

Parsing the Regulatory Network between Small RNAs and Target Genes in Ethylene Pathway in Tomato

Yunxiang Wang[1, 2, 3, 4], *Qing Wang*[1, 2, 3, 4], *Lipu Gao*[1, 2, 3, 4], *Benzhong Zhu*[5], *Zheng Ju*[5], *Yunbo Luo*[5] *and Jinhua Zuo*[1, 2, 3, 4]*

[1] Key Laboratory of the Vegetable Postharvest Treatment of Ministry of Agriculture, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [2] Beijing Key Laboratory of Fruits and Vegetable Storage and Processing, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [3] Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China) of Ministry of Agriculture, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [4] Key Laboratory of Urban Agriculture (North) of Ministry of Agriculture, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [5] Laboratory of Postharvest Molecular Biology of Fruits and Vegetables, Department of Food Biotechnology, College of Food Science and Nutritional Engineering, China Agricultural University, Beijing, China

Small RNAs are a class of short non-coding endogenous RNAs that play essential roles in many biological processes. Recent studies have reported that microRNAs (miRNAs) are also involved in ethylene signaling in plants. *LeERF1* is one of the ethylene response factors (ERFs) in tomato that locates in the downstream of ethylene signal transduction pathway. To elucidate the intricate regulatory roles of small RNAs in ethylene signaling pathway in tomato, the deep sequencing and bioinformatics methods were combined to decipher the small RNAs landscape in wild and sense-/antisense-*LeERF1* transgenic tomato fruits. Except for the known miRNAs, 36 putative novel miRNAs, 6 trans-acting short interfering RNAs (ta-siRNAs), and 958 natural antisense small interfering RNAs (nat-siRNAs) were also found in our results, which enriched the tomato small RNAs repository. Among these small RNAs, 9 miRNAs, and 12 nat-siRNAs were differentially expressed between the wild and transgenic tomato fruits significantly. A large amount of target genes of the small RNAs were identified and some of them were involved in ethylene pathway, including AP2 TFs, auxin response factors, F-box proteins, ERF TFs, APETALA2-like protein, and MADS-box TFs. Degradome sequencing further confirmed the targets of miRNAs and six novel targets were also discovered. Furthermore, a regulatory model which reveals the regulation relationships between the small RNAs and their targets involved in ethylene signaling was set up. This work provides basic information for further investigation of the function of small RNAs in ethylene pathway and fruit ripening.

Keywords: ethylene, microRNAs, target, high-throughput sequencing, regulatory network

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Mohamed Zouine,
INR/INPT, France
Davide Stefano Sardina,
University of Catania, Italy

***Correspondence:**
Jinhua Zuo
zuojinhua@126.com

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

Received: *11 September 2016*
Accepted: *24 March 2017*
Published: *11 April 2017*

Citation:
Wang Y, Wang Q, Gao L, Zhu B, Ju Z, Luo Y and Zuo J (2017) Parsing the Regulatory Network between Small RNAs and Target Genes in Ethylene Pathway in Tomato. Front. Plant Sci. 8:527. doi: 10.3389/fpls.2017.00527

INTRODUCTION

Small RNAs are a class of non-coding endogenous RNAs ranged from 20 to 24 nucleotides (nt) that play essential roles in plant growth and development, signal transduction, response to biotic and abiotic stresses and other biological processes (Rhoades et al., 2002; Jones-Rhoades et al., 2006; Tomato Genome Consortium, 2012). MicroRNAs (miRNAs) and small-interfering RNAs (siRNAs) are two mainly classes of small RNAs divided on the difference of their precursor structures

and biosynthetic pathways (Carthew and Sontheimer, 2009). Mature miRNAs are evolved from miRNA genes with the action of Dicer-like 1 (DCL1), Hua Enhancer 1 (HEN1), and HASTY proteins (Jones-Rhoades et al., 2006; Xie et al., 2015). SiRNAs are derived from long double-stranded RNAs (dsRNAs) and could be classed to heterochromatic siRNAs (hc-siRNAs), trans-acting short interfering RNAs (ta-siRNAs) and natural antisense siRNAs (nat-siRNAs; Chen, 2009). Recent studies showed that small RNAs can negatively regulate gene expression at the post-transcriptional level based on two possible mechanisms: transcript cleavage and translational repression (Sunkar et al., 2007; Couzigou and Combier, 2016).

As a climacteric fruit model, tomato has been widely used to study the molecular mechanisms of fruit ripening and senescence as well as ethylene biosynthesis and signal transduction. Recently, increasing studies showed that small RNAs are also involved in regulating ethylene signal transduction (Pilcher et al., 2007; Moxon et al., 2008; Zhang et al., 2011; Zuo et al., 2012). For example, Moxon et al. (2008) found that one of the target genes of miR156 was CNR, which belongs to SBP-box family transcription factors (TFs), and the target gene of miR172 was AP2. It has been reported that the expression of genes that encode miRNAs is regulated at the transcriptional level by various transcriptional factors (Yant et al., 2010; Baek et al., 2013). For example, EIN3, a key transcription factor in ethylene signaling, directly binds to the promoter region of miR164 and represses its transcription (Li et al., 2013).

ERFs were a class of TFs located in the downstream of ethylene signal transduction pathways that function in diverse plant growth and metabolism processes as well as in the biotic and abiotic stress response, such as ethylene (Wu et al., 2002; Pirrello et al., 2006), high salt (Park et al., 2001; Wang et al., 2004), drought and low temperature, and so on (Qin et al., 2004; Zhang et al., 2007). Given that the miRNAs were also involved in the ethylene signaling pathways, there may be some relationships between miRNAs and ERFs. The high-throughput sequencing technology has been widely used to explore the functions of miRNA and siRNAs due to its high throughputs and accuracy (An et al., 2011; Cao et al., 2014; Thiebaut et al., 2014). In this study, High-throughput sequencing of small RNAs and degradome sequencing were used to gain a better understanding of the relationship between ethylene and small RNAs using wild type and *LeERF1* transgenic tomato fruits. MiRNAs expression patterns were profiled and their targets were conferred; the regulatory network model between the small RNAs and ethylene was set up. This research provides more evidences for understanding the regulatory pathways of miRNAs in the network of fruit ripening.

MATERIALS AND METHODS

Sample Collection and Preparation

Wild type (*Solanum lycopersicum* cv. zhongshu4) and sense-/antisense-*LeERF1* transgenic tomato plants (Li et al., 2007) were grown in the greenhouse at standard conditions. The Fruits at breaker stage were used in the experiment (**Supplementary Figure S1**). Pooled mesocarp tissues from three groups were flash frozen in liquid nitrogen and stored at −80°C until further analysis.

Small RNA (sRNA) Quantification and Qualification

The RNA samples were extracted using Trizol. Nanodrop, Qubit 2.0, and Agilent 2100 bioanalyzer were used to detect the purity, concentration and integrity of RNA samples, respectively, to ensure the use of qualified samples for sequencing. RNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). RNA concentration was measured using Qubit® RNA Assay Kit in Qubit®2.0 Flurometer (Life Technologies, CA, USA). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

Library Preparation for Small RNA Sequencing

A total amount of 1.5 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext®Ultra™ small RNA Sample Library Prep Kit for Illumina®(NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, first of all, ligated the 3′ SR Adaptor, mixed 3′ SR Adaptor for Illumina, RNA and Nuclease-Free Water, mixture system after incubationa for 2 min at 70 degrees in a preheated thermal cycler, Tube was transferred to ice. Then, add 3′ Ligation Reaction Buffer (2X) and 3′ Ligation Enzyme Mix ligate the 3′ SR Adaptor, incubated for 1 h at 25°C in a thermal cycler. To prevent adaptor-dimer formation, the SR RT Primer hybridizes to the excess of 3′ SR Adaptor (that remains free after the 3′ ligation reaction) and transforms the single stranded DNA adaptor into a double-stranded DNA molecule. sRNAs (18–30 nucleotides in length) were separated from the total RNAs by polyacrylamide gel electrophoresis (PAGE). The small RNA molecules were then ligated with 5′ and 3′ adaptor and used for reverse transcription and subsequent PCR. The final PCR product was purified and sequenced by Illumina Cluster Station and Illumina Genome Analyzer (SanDiego, CA, USA).

Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq 2500 platform and pair-end reads were generated. The sequencing results were deposited in the Sequence Read Archive (SRA) at the NCBI database (accession number: SRP094091).

Quality Control

The quality control of raw data (raw reads) in fastq format has been performed by using in-house scripts written in Perl. Reads containing adapter and poly-N sequences and reads with low

quality from raw data were removed. Then reads were cleaned by removing the sequences smaller than 18 nt or longer than 30 nt. At the same time, Q20, Q30, GC-content, and sequence duplication level of the clean data were calculated. All the downstream analyses were based on clean data with high quality.

Bioinformatic Analysis of Sequencing Data

The Clean Reads were aligned with Silva database, GtRNAdb database, Rfam database, and Repbase database respectively to filter ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), repeat sequences, and other ncRNA using Bowtie tools. The remaining reads were used to detect known miRNAs and new miRNAs predicted by comparing with known miRNAs from miRBase. RNAFold tools were used to predict the secondary structure of all new miRNAs.

Identification of siRNA and Putative Novel miRNA

The adapter reads of the Solexa sequencing results were removed (**Supplementary Figure S2**). And reads larger than 30 nt and smaller than 18 nt were discarded. All high quality reads were considered as significant and further analyzed. Small RNA reads were mapped to tomato genome with mapping tool bowtie, all tomato genome annotation information is downloaded from ITAG2.3 include repeats and protein-coding regions (http://solgenomics.net/organism/Solanum_lycopersicum/genome). Six libraries are pooled together for miRNA prediction. The potential miRNA loci were analyzed using MIREAP software (version 0.2) with default parameters followed by additional manual check criteria that included: the miRNA sequence length should be between 18 and 26 nt; the maximal free energy allowed for a miRNA precursor (−18 kcal/mol); flank sequence length of miRNA precursor (100 nt); the predicted mature miRNA reads count should be large than 10 and reading counts ratio for miRNA*/miRNA should be small than 0.1. The unique reads left were aligned with known miRNAs from miRBase 21.0 (http://www.mirbase.org/). Phased small RNAs and nat-siRNAs were predicted as described in the previous studies (Chen et al., 2007; Zhou et al., 2009). All the reading counts were normalized to per million of total mapped reads (TPM).

Target Gene Functional Annotation

Gene function was annotated based on the following databases: Nr (NCBI non-redundant protein sequences); Nt (NCBI non-redundant nucleotide sequences); Pfam (Protein family); KOG/COG (Clusters of Orthologous Groups of proteins); Swiss-Prot (A manually annotated and reviewed protein sequence database); KO (KEGG Ortholog database); GO (GeneOntology).

Quantification of Small RNAs Expression Levels and Differential Expression Analysis

Small RNA expression levels were estimated by TPM for each sample: sRNA were mapped back onto the reference genome, and read count for each small RNA was obtained from the mapping results. For the samples with biological replicates, differential expression analysis of two conditions/groups was performed using the DESeq R package (1.10.1). DESeq provide statistical routines for determining differential expression in digital miRNA expression data using a model based on the negative binomial distribution. The resulting *P*-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. MiRNA with an adjusted $p < 0.05$ and $|\log_2(\text{fold change})| \geq 1$ were assigned as differentially expressed (Anders and Huber, 2010).

GO Enrichment Analysis

Gene Ontology (GO) enrichment analysis of the differentially expressed genes (DEGs) was implemented by the GOseq R packages based on Wallenius non-central hyper-geometric distribution (Young et al., 2010), which can adjust to gene length bias in DEGs.

KEGG Pathway Enrichment Analysis

KEGG (Kanehisa et al., 2008) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism, and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (http://www.genome.jp/kegg/). We used KOBAS (Mao et al., 2005) software to test the statistical enrichment of differential expression genes in KEGG pathways.

RESULTS

Overview of the Small RNA Libraries from Tomato Fruit

To identify small RNAs and analyze their functions in ethylene pathway, the deep sequencing technology with Illumina Hiseq 2500 platform (Biomarker Technologies, China) was performed in wild and sense-/antisense-*LeERF1* transgenic tomato fruits at breaker stage. A total of 19.31, 19.80, 17.28, 15.24, 14.61, 14.60 million raw reads in CK1 (wild 1), CK2 (wild 2), F1 (sense-*LeERF1* transgenic tomato 1), F2 (sense-*LeERF1* transgenic tomato 2), R1 (antisense-*LeERF1* transgenic tomato 1), and R2 (antisense-*LeERF1* transgenic tomato 2) were generated, respectively. After removing low quality and contaminated reads, poly A-containing sequences, sequences outside of 18–30 nt, 3′ and 5′ adaptors sequences, 15.75 (CK1), 16.88 (CK2), 13.79 (F1), 13.05 (F2), 12.63 (R1), 12.94 (R2) million clean reads were remained for further analysis. Then the small RNAs were categorized into miRNAs, ribosomal (r) RNAs, transfer (t) RNAs, small nuclear (sn) RNAs, small nucleolar (sno) RNAs, repeat regions, exon and intron RNA based on genomic location and function analysis (**Table 1**).

The size distribution is one of the distinct features of the small RNAs libraries from different plants. In our experiments, the length of small RNAs ranges from 18 to 30 nt and the most abundant group of small RNAs have length 21–24 nt in all the six libraries (**Figure 1**). There is no obvious difference of the length distribution between wild and *LeERF1* transgenic tomato.

TABLE 1 | Small RNAs profiling and classification in six tomato fruit groups.

Types	CK1	CK2	F1	F2	R1	R2
Total	15,753,031	16,876,616	13,793,259	13,050,621	12,633,453	12,938,234
miRNA	606,491	678,439	529,661	514,194	506,601	500,709
rRNA	363,895	334,157	348,969	268,843	294,359	318,280
tRNA	73,283	69,194	68,966	58,727	42,953	60,809
snRNA	4,726	5,906	5,517	5,220	4,802	5,752
snoRNA	7,876	8,100	8,275	7,264	5,938	6,951
Repeat	2,451,172	2,683,382	2,008,299	1,958,898	1,867,224	1,935,560
NAT	1,228,736	1,400,759	1,213,807	1,070,151	1,174,911	1,125,626
TAS	40,958	37,129	37,242	40,457	41,690	37,521
exon:+	677,380	573,805	537,937	587,278	517,972	491,653
exon:−	247,323	275,089	219,313	216,640	228,665	238,064
intron:+	849,088	916,400	760,009	726,920	720,107	720,660
intron:−	521,425	567,054	464,833	445,026	423,221	437,312
Other	8,680,677	9,327,202	7,590,433	7,151,003	6,805,010	7,059,338

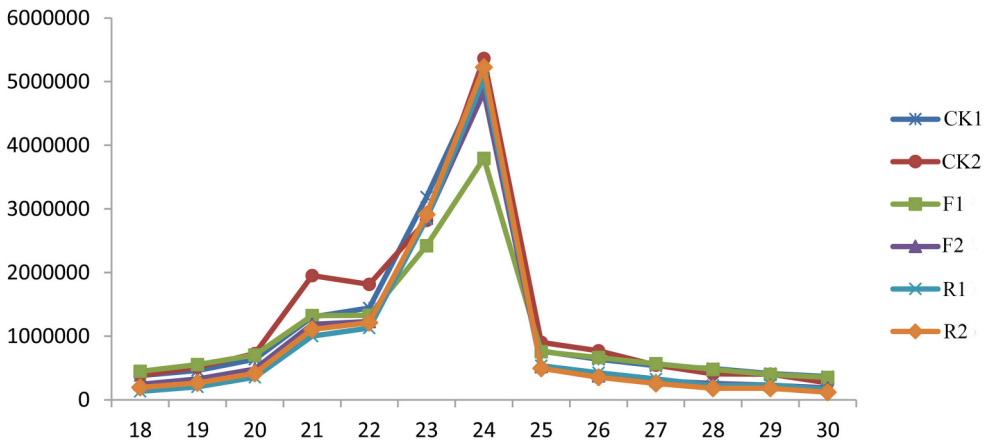


FIGURE 1 | Length distribution of small RNAs in wild (CK1, CK2), sense-*LeERF1* (F1, F2), and antisense-*LeERF1* (R1, R2) tomato fruit.

Among the 21–24 nt size small RNAs, 24-nt size class has the highest abundance, accompanied with the 23-nt sRNAs as the second largest groups, which is in accordance with that of rice (Morin et al., 2008), Arabidopsis (Rajagopalan et al., 2006), and our previous results (Zuo et al., 2012, 2013).

Identification of miRNAs and siRNAs in Tomato

To identify miRNAs and siRNAs in tomato, the clean sequences were aligned with the tomato small RNAs database (http://ted.bti.cornell.edu/cgi-bin/TFGD/sRNA/home.cgi) and the latest miRNA database (http://www.mirbase.org/, Release 21). In total, 178 known miRNAs belonging to 108 families were obtained in our libraries. Among the 108 families, 46 miRNA families (**Supplementary Table S1**) were registered in miRBase as belonging to *S. lycopersicum* and other 62 families (**Supplementary Table S2**) were less conserved and first identified in tomato (**Supplementary Figure S3**). Most of the miRNA families belonging to *S. lycopersicum* in miRBase are composed of more than one member. For instance, miR156 and miR482 were the largest ones with seven members in the families in this study. MiR171 and miR319 were the second largest family with six members. On the other hand, except for the miR548 family, other less conserved miRNA families had only one member detected in this study. Sequence length statistical results showed that the 21-nt miRNAs were the main type of the 178 known miRNAs.

In addition, 36 putative novel miRNAs with hairpin structures renamed as miRZ101 to miRZ136 were predicted and all of them were found to have star sequences (**Table 2** and **Supplementary Table S3**). The length of the putative novel miRNAs were 18–24 and 24 nt miRNAs accounted for the predominance. Most of the first nucleotide of the putative novel miRNAs were A, which was in accordance with previous study that 24 nt miRNAs used to had an A as the first nucleotide (Jain et al., 2014). The minimum folding free energies varied from −149.8 to −28.3 kcal/mol (**Supplementary Table S3**).

TABLE 2 | Putative novel miRNAs found in tomato.

MiRNA	Length	Sequence	Chromosome	Star	MEF (kcal/mol)
miRZ101	22	uaacuucgucuagcucgccuuc	10	+	−70
miRZ102	24	guagagaacucuagaaccuucuag	10	+	−84.1
miRZ103	24	aaaggacuccuagauuucucuagu	11	+	−93.9
miRZ104	24	aaagacuguucaauuacugcuuga	11	+	−28.3
miRZ105	24	uauguccuuuaacuuugagugugc	12	+	−110.4
miRZ106	24	uuaguauaguauaagugugucucu	12	+	−57.1
miRZ107	24	acacacucugcauucaauuaaauu	12	+	−63.4
miRZ108	24	acguugcucagacucuucaaaaau	12	+	−60.3
miRZ109	22	auuuauggcuaugaauuugagu	12	+	−62.9
miRZ110	24	uuaguauaguauaagugugucucu	1	+	−41.7
miRZ111	24	uuaguuuaauuaagaugugucucu	1	+	−105.2
miRZ112	21	gcacggcagauaguuauuggc	1	+	−114.6
miRZ113	24	guagagaacucuagaaccuucuag	1	+	−71.4
miRZ114	24	aagcgaugacuuuagugaccuagu	1	+	−39.9
miRZ115	22	cacggucguaccuugacaaggc	2	+	−77.8
miRZ116	22	uuguuucuguuuuuguuugagu	2	+	−149.8
miRZ117	23	guugcucggacucuucaaaaaug	2	+	−69.1
miRZ118	20	auaacacaaaucugagccuc	2	+	−56.5
miRZ119	22	agugacucgcucgaucuuucuu	3	+	−64
miRZ120	24	uuucgucuugaaaguuugcccaug	4	+	−58.3
miRZ121	24	auuuccgaucaaacuuaaacuguu	4	+	−40.8
miRZ122	23	guugcucgaacucuucaaaaaug	5	+	−62.7
miRZ123	24	augugaucgcuguaaugaccuuac	5	+	−132.9
miRZ124	24	ucgagggucuaucagaaacaacau	6	+	−50.8
miRZ125	18	accugguugauccugcga	6	+	−73.3
miRZ126	24	guugcucgaacucuucaaaaaugu	6	+	−78.9
miRZ127	24	uuuucuaucggaacuaucaugugu	6	+	−69.3
miRZ128	21	ucaacgcugcacucaaucaug	7	+	−75.2
miRZ129	24	aagacguuugaaucugaaaaagau	8	+	−57.7
miRZ130	23	uuauacuauacuaagguccuauu	8	+	−117
miRZ131	24	cgagugcucauuccacagauaagu	8	+	−64.2
miRZ132	24	auacaucgguuacuugauagacgu	8	+	−109.4
miRZ133	24	uuaguauaguauaagugugucucu	8	+	−103.1
miRZ134	24	ugaaaucgagaugugauguagagg	9	+	−59.9
miRZ135	23	uucuucugacucauuuacuuuag	9	+	−54.2
miRZ136	24	augcucuagucuuugaacgacagg	9	+	−58.7

Moreover, several conserved and species-specific endogenous siRNAs were also characterized in our libraries. Ta-siRNAs are a special class of siRNAs that generated from TAS gene transcripts and mediated by miRNA (Xie et al., 2005; Yoshikawa et al., 2005; Li et al., 2012). On the basis of the conservation of the TAS genes in plants, three TAS5 gene family members: TAS5, TAS5b, and TAS5d (TAS5b and TAS5d were found in our previous study; Zuo et al., 2016), all miR482 targets, were identified (**Table 3**). Surprisingly, one more TAS5 family member (TAS5e) and two more TAS genes (TAS11a and TAS11b), triggered by sly-miR6024, are reported in our results (**Table 4**). In addition, 19 potential phased small RNAs and 958 nat-siRNAs were also found in this study (**Supplementary Tables S4**,**S5**).

The Effect of Overexpression Sense-/Antisense-*LeERF1* on Small RNA Profiles

To evaluate the regulatory roles of *LeERF1*on miRNA expression, differential expression of miRNAs among the wild and sense-/antisense-*LeERF1* transgenic tomato were analyzed. After normalization using a RPM method, the miR399a was found to have significant different accumulation between wild type and sense-*LeERF1* transgenic tomato fruits. The expression of the miR399a was down-regulated in sense-*LeERF1* transgenic fruit (**Figure 4A**). MiR8990 and the novel miRZ118 were the two miRNAs significant differently expressed between wild type and antisense-*LeERF1* transgenic tomato fruits, and their accumulations decreased in the transgenic fruit. Totally, there

TABLE 3 | The conserved TAS5 family in tomato fruit.

Name	Chromosome	Length	Start	End	Phased abundance	related miRNA
sly-TAS5	6	539	423,570	424,108	3,954	sly-miR482d-3p
sly-TAS5b	2	644	21,186,658	21,187,301	530	sly-miR482d-3p
sly-TAS5d	8	917	58,262,775	58,263,691	4,729	sly-miR482e-3p

Three TAS5 family members were found and were located in Chromosome 6, 2, and 8 separately. "Phased abundance" means the abundance of phased sequence and the "related miRNA" related the miRNAs that mediated TAS.

TABLE 4 | The novel TAS families in tomato fruit.

Name	Chromosome	Start	End	Length	Phased abundance	related miRNA
sly-TAS5e	11	48,467,984	48,468,816	833	22,190	sly-miR482b
sly-TAS11a	5	2,500,975	2,501,555	581	389	sly-miR6024
sly-TAS11b	11	51,986,458	51,986,681	224	137	sly-miR6024

Three members belong to two TAS families (TAS5, TAS11) were found and located in Chromosome 11, 5, and 11 separately. "Phased abundance" means the abundance of phased sequence and the "related miRNA" related the miRNAs that mediated TAS.

were nine miRNAs having significant differential expression between sense-*LeERF1* and antisense-*LeERF1* transgenic fruit. Among them, miR399a and miR8263-5p were up-regulated in antisense-*LeERF1* transgenic fruits. Meanwhile, other seven miRNAs including miR7484, miR319a, miR95-5p, miR8990, miR2569-5p, and two putative novel miRNAs (miRZ118 and miRZ131) were down-regulated.

Besides, 12 nat-siRNAs were found to show differential expression patterns. Compared with wild type fruits, most of the nat-siRNAs showed lower expression in sense-*LeERF1* tomato fruits, only two of them increased (**Figure 4B**). However, among the differentially expressed nat-siRNAs, more than half of them had higher expression levels in antisense-*LeERF1* transgenic fruits.

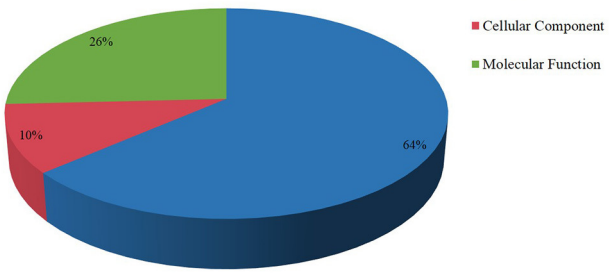


FIGURE 2 | Targets of miRNAs in tomato fruit. Most targets of the miRNAs were involved biological process.

Target Gene Identification of the miRNAs

MiRNAs regulate gene expression mainly through cleaving mRNA or inhibiting the translation process of the targets gene, so identification and analysis of the target genes were the basis to study the function of miRNAs. Bioinformatics prediction and high-throughput degradome sequencing were the two main methods to find the targets gene. Using bioinformatic prediction method, a total of 103 target genes that involved in biological process, cellular component, and molecular function were found and most of them were identified to participate in biological process (**Figure 2**). Previous studies indicated that the targets of conserved miRNAs were also conservative and most of the miRNA families had not only one target site (Jin et al., 2008; Lu et al., 2008), which was also found in our study. For example, the targets of miR166a are homeobox-leucine zipper protein Revoluta, homeobox-leucine zipper protein ATHB-14 and pentatricopeptide repeat-containing protein At5g25630. Meanwhile, one target gene was often cleaved by two or more miRNAs. For instance, AP2 is the target of miR172 and miR8737, miR319 and miR159 share the same target GAMYB. Among the identified target genes, 16 targets were found to be involved in ethylene (**Supplementary Table S6**) and most of them were AP2 TFs. Another class of targets was auxin response factors including ARF10, ARF16, ARF17 and ARF18. Two F-box proteins (F-box protein 6, F-box protein At3g07870-like), two ethylene-responsive factors (RAP2-7-like) and an APETALA2-like protein were also predicted.

High-throughput degradome sequencing is a new technology to identify miRNAs targets and is successfully applied in Arabidopsis, rice (Addo-Quaye et al., 2008; Li et al., 2010). In this study, a total of 55 cleavage sites associated with 41 miRNAs were detected and seven target genes cleaved by five miRNAs were identified to be related to ethylene synthesis and signal transduction, including five auxin response factors (ARFs), one AP2 TF, and one ERF TF (**Supplementary Table S7**). Except for the known targets, six new targets were identified. The representative target plots of new targets were shown in **Figure 3**.

In addition, 389 genes were predicted to be the targets of nat-siRNAs, and 22 of them were found to participate in ethylene pathway (**Supplementary Table S8**). Ethylene-responsive TFs, MADS-box TFs, F-box proteins were the main targets involved in fruit ripening. Moreover, 55 targets of the ta-siRNAs

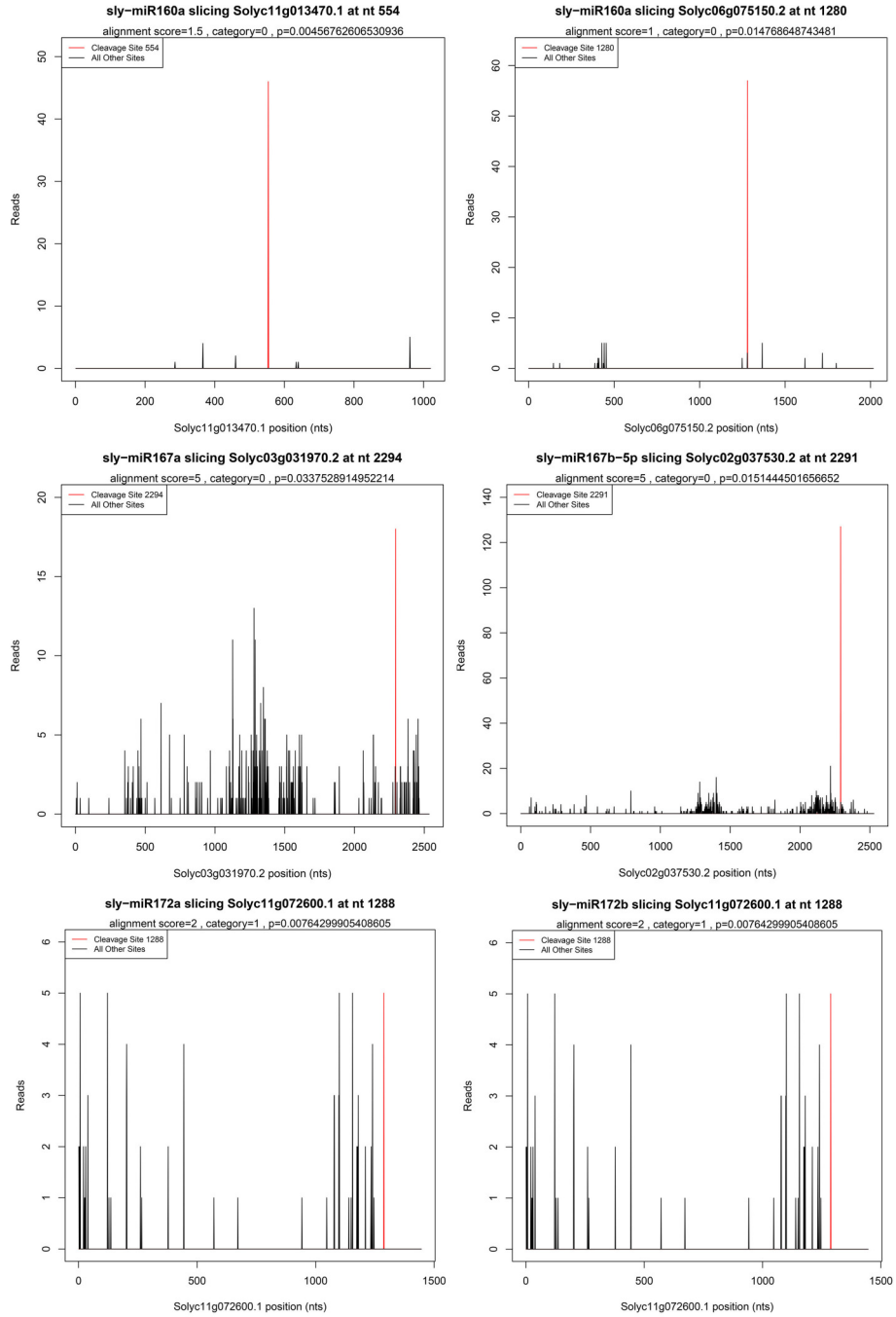


FIGURE 3 | Target plots of miRNA targets confirmed by degradome sequencing. Six new genes were found to be targets of five miRNAs.

were also predicted and two of them (Solyc02g092020.1 and Solyc02g082320.1) were related to ethylene pathway.

Target Parsing and Small RNA Regulatory Network Analysis in Tomato

To investigate the network between the miRNAs and their targets, cytoscape platform was employed. In the network, it could be clearly seen that miR6024 had 11 targets and among the targets two of them were also the targets of miR482 (**Figure 5**). In addition, miR6024 and miR6027-3p shared one common targets. Moreover, miR6022 and miR528 shared most of their targets and they also had common targets with miR6023 and miR8527.

To comprehensively understand the functions of the miRNAs, ta-siRNAs, and nat-siRNAs involved in ethylene synthesis and signal transduction, all the predicted target genes of small RNAs were screened carefully and a regulatory model including

small RNAs, the targets and their main functions was set up (**Figure 6**). As shown in **Figure 6**, it could clearly been seen that the AP2 TFs that involved in ethylene signaling were the targets of miR172a, miR172b, and miR8737. Ethylene-responsive transcription factors were the target of miR172a, miR172b and the nat-siRNAs renamed as nat-siRNA-G2009, nat-siRNA-G2010, nat-siRNA-G2011, nat-siRNA-G2012, nat-siRNA-G2013, and nat-siRNA-G2014. Meanwhile, the F-Box proteins and MADS-Box TFs that participated in ethylene signaling were the targets of miR394, miRZ131, nat-siRNA-G2015 to nat-siRNA-G2017, sly-TAS5d, phased small RNA001 and nat-siRNA-G2001 to nat-siRNA-G2005, respectively. Moreover, the auxin response factors that indirectly control ethylene signaling were the targets of miR160a and the nat-siRNA-G2006 and nat-siRNA-G2007.

DISCUSSION

Small RNAs are a class of non-coding RNAs that play vital roles in growth and development, signal transduction, biotic and abiotic stresses (Jones-Rhoades et al., 2006; Dalmay, 2010; Mohorianu et al., 2011; Zuo et al., 2012; Pashkovskiy and Ryazansky, 2013). Numerous studies have demonstrated that miRNAs were involved in the regulation of diverse physiological processes by repressing the expression of their target genes. Ethylene is an important endogenous hormone and plays important roles in fruit development and ripening. As a model plant, tomato has been widely used to study the molecular mechanisms of ethylene biosynthesis and signal transduction (Giovannoni, 2004; Osorio et al., 2011), and through the study on ripening-related mutants or transgenic plant, many advances have been achieved. ERFs were a class of TFs located in the downstream of ethylene signal transduction pathways, and as one of the members of ERF class, *LeERF1* had been showed to mediate fruit maturation and softening, enhance resistance to osmotic stress and improve plant tolerance to fungal invasion (Li et al., 2007; Lu et al., 2011; Pan et al., 2013). To better understand the relationship between *LeERF1* and small RNAs in ethylene pathway, high-throughput sequencing was employed in the sense-/antisense-*LeERF1* transgenic tomato fruits and many ethylene-related small RNAs as well as their target genes were found.

High-Throughput Sequencing of Tomato Fruit

In the past decades, miRNAs identification and their biological roles analysis were the mainly focused research fields. In tomato, 46 miRNA families were identified and registered in the miRBase database (http://www.mirbase.org/). It is well-known that many small RNAs have temporal expression patterns (Chen, 2009; Rubio-Somoza et al., 2009) and many studies had not detected all the 46 miRNA families (Candar-Cakir et al., 2016; Wu et al., 2016). In this study, the 46 families were all identified though some miRNAs did not found in all libraries, such as miR169 that only detected in wild tomatoes. This result indicated that the high-throughput sequencing had superiority in the identification of small RNA. Meanwhile, we identified 62 less conservative miRNAs that had not previous been found in tomato but documented in the miRBase for other species. For instance, miR861 was found in Arabidopsis (Fahlgren et al., 2007) and miR8010 were registered for potato (Zhang et al., 2013). MiR440, miR528, miR2922 and miR1049, miR1222, miR1063 were detected in rice and moss, respectively (Liu et al., 2005; Sunkar et al., 2005; Talmor-Neiman et al., 2006; Axtell et al., 2007; Sanan-Mishra et al., 2009).

In addition, 41 putative novel miRNAs not identified in other reports were also predicted in this study. The hairpin structures were found and the minimal folding free energies (MFEs) were −149.8 to −28.3 kcal/mol, indicated that the hairpin structures were stable. MiRNAs with detected stars were more likely to predict to be bona fide novel miRNAs (Wu et al., 2016). The renamed putative novel miRNAs in our libraries all had stars, suggested the accuracy of the novel miRNAs. Most of the putative novel miRNAs were 24 nt in length. The 24 nt small RNAs were reported to mainly match to the promoter regions of ripening-associated genes (Tomato Genome Consortium, 2012) and its high percentage in the putative novel miRNAs may

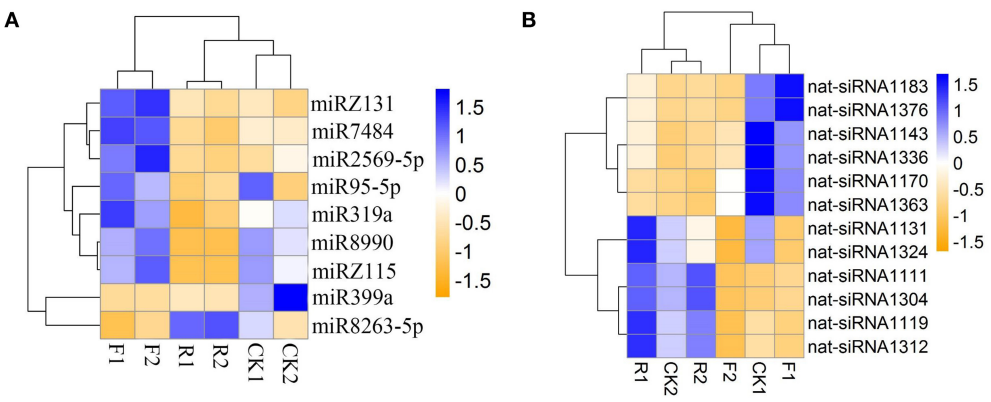


FIGURE 4 | Expression profiles of the significantly differentially expressed miRNAs (A) and nat-siRNAs **(B)** in wild (CK1, CK2), overexpression sense-*LeERF1* (F1, F2), and antisense-*LeERF1* (R1, R2) tomato fruit.

imply regulatory roles of the novel miRNAs in ethylene pathway.

Overall, the discovery of less conserved and putative novel miRNAs in tomato provided enriched insight into the plant miRNA dataset. In addition, one previously reported TAS5 gene (Li et al., 2012) and two other TAS5 genes (TAS5b, TAS5d, found in our other study) were detected in our libraries (**Table 3**). Moreover, another TAS5 (TAS5e) with the target sly-miR482b and two more TAS genes (Named TAS11a and TAS11b) triggered by sly-miR6024 were also found in our libraries (**Table 4**).

Differential Expression Profiles of the Small RNAs

It is well-known that many small RNAs have temporal expression patterns (Chen, 2009; Rubio-Somoza et al., 2009). Differential expression patterns of small RNAs can be regarded as an index for estimating the regulation contributions. We analyzed small RNAs expression in the wild and sense-/antisense-*LeERF1* transgenic tomato. It is worth to noting that 14 miRNAs had significant difference expression between the wild and transgenic tomato. Among them, four miRNA families had a significant different

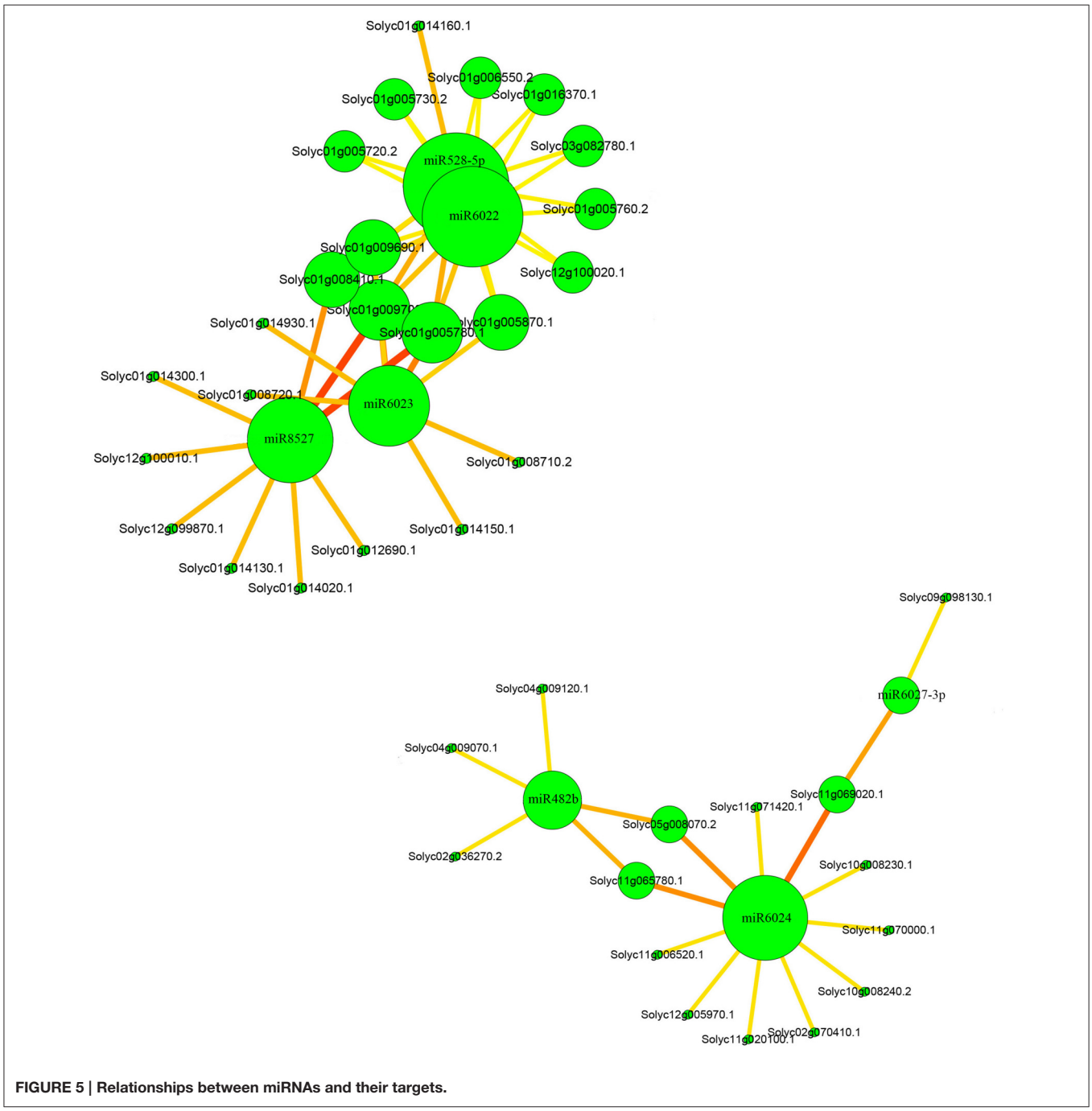


FIGURE 5 | Relationships between miRNAs and their targets.

accumulation between wild and sense-*LeERF1* transgenic tomato fruits and 10 miRNAs differentially expressed in the response to antisense-*LeERF1*, indicating their specific roles in fruit ripening (**Figure 4A**).

It has been reported that miR399 was involved in plant response to phosphate starvation (Fujii et al., 2005; Chiou et al., 2006) and its accumulation increased during fruit development in tomato (Gao et al., 2015). In this study, the miR399a is down-regulated in sense-*LeERF1* transgenic fruit, which indicated that miR399 may play an important role in ethylene signal transduction pathway. Totally, there were nine miRNAs having significant different expression between sense-*LeERF1* and antisense-*LeERF1* transgenic fruit. Among them, miR399a and miR8263-5p were up-regulated in antisense-*LeERF1* transgenic fruits, meanwhile, other nine miRNAs, including miR7484, miR319a, miR95-5p, miR8990, miR2569-5p, and two putative novel miRNAs (miRZ118 and miRZ131) were down-regulated. miR319 has been reported to control leaf development and morphogenesis through regulating transmission control protocol (TCP) transcription factors (Palatnik et al., 2003). In this study, the target of miR319 was predicted to be GAMYB, which was also related to ethylene pathway. According to our results, miR319 may also participate in ethylene signaling pathway.

Besides, 12 nat-siRNAs were found to show differential expression patterns. Compared with wild fruits, most of the nat-siRNAs showed lower expression levels in sense-*LeERF1* tomatoes, and only two of them increased. However, among the nat-siRNAs, more than half of them had higher expression levels in antisense-*LeERF1* transgenic fruits (**Figure 4B**).

Small RNAs Participated in Ethylene Pathway

To study the function of small RNAs in ethylene pathway, bioinformatic prediction, and degradome sequencing were also used in wild and sense-/antisense-*LeERF1* transgenic tomato. Results showed that most of the targets were identified to participate in various biological processes (**Figure 2**). AP2 transcription factors, AP2-like ethylene-responsive transcription factors, ethylene-responsive transcription factor were TFs belong to AP2/EREBP transcription factors family involved in ethylene signaling pathway and they were the main targets of miR172 family, which was also reported in Arabidopsis and tomato (Wu et al., 2009; Cheng et al., 2016). The AP2/EREBP transcription factors were also the target of miR5658 (Cheng et al., 2016). However, in this study, the miR5658 was not detected and miR8737 were predicted to target AP2/ERF TFs. F-Box proteins were reported to regulate ethylene signaling in Arabidopsis (Wang et al., 2009). It was also been reported that the F-Box proteins were the targets of miR393 in Arabidopsis (Liu et al., 2008). However, in this study, miR393 was not found and F-Box proteins were predicted to be the target of miR394. In addition, auxin response factors genes cleaved by miR160 were also found in our results. In Arabidopsis, ARF6 and ARF8, ARF16, and ARF17 were reported to be the targets of miR167 and miR160,

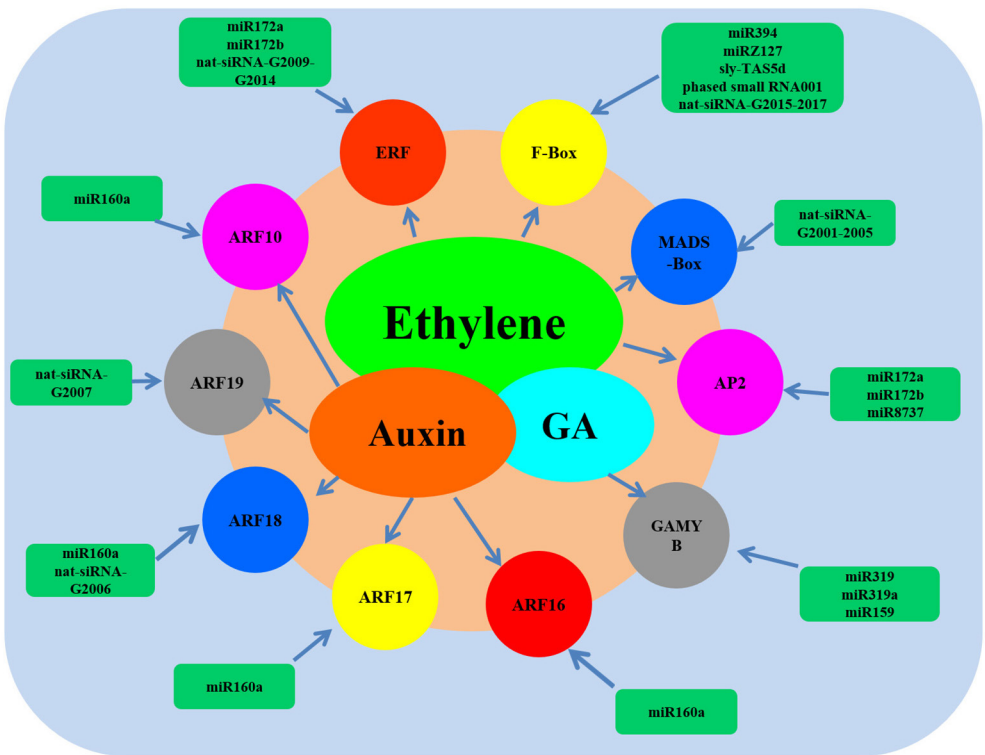


FIGURE 6 | Network model of the small RNAs and their target genes involved in ethylene.

respectively. In this study, ARF8 cleaved by miR167 was found via degradome sequencing. However, ARF6 was not identified for miR167, perhaps because the abundance of cleaved products was too low to be detected. Unexpectedly, ARF10 and ARF18 were identified to be the targets of miR160. Surprisingly, six new targets were found (**Figure 3**).

Compared with the miRNAs, most targets of the ta-siRNAs and nat-siRNAs in tomato were not verified yet. The distribution of the targets of ta-siRNAs and nat-siRNAs was different from that of the miRNAs, and a great part of the targets were predicted to be involve in all kinds of metabolic processes which were consistent with the previous studies (Zhai et al., 2011; Li et al., 2012; Shivaprasad et al., 2012; Zuo et al., 2013). In this study, several important target genes participating in fruit ripening and senescence were found including Ethylene-responsive transcription factors, F-box proteins, MADS-box TFs, and MADS-box proteins.

Network Construction Revealed the Relationship of Small RNAs and Ethylene in Tomato

To illuminating the network between small RNAs and their target genes involved in ethylene, all the predicted target genes of miRNAs, ta-siRNAs, and nat-siRNAs were screened carefully and a regulatory model was set up (**Figure 6**). From the network model, it could clearly been seen that miR394, miRZ131, miR172, miR8737, miR319, miR159, miR160, and nat-siRNA-G2001 to nat-siRNA-G2017 as well as their target genes such as auxin response factors, ethylene-responsive transcription factors and GAMYB were involved in ethylene signal pathway. These results indicate that the network of miRNAs are quite complicated, and elucidation of the molecular mechanisms underlying the interplay between miRNA and their target genes involved in ethylene pathway requires further study.

CONCLUSION

In summary, ethylene biosynthesis and signal transduction related miRNAs and siRNAs were identified in tomato fruit. These informations broaden the knowledge of the relationship between small RNAs and ethylene regulation. Additionally, many target genes of miRNAs were identified by bioinformatic prediction and degradome sequencing. The result showed that the target genes were involved in various functions and ethylene related targets were also discovered. In addition, a large amount of the target genes of nat-siRNAs were found and some of them were found to participate in ethylene regulation. A regulatory model which reveals the regulation relationship between the small RNAs and their targets was set up. Moreover, 41 putative novel miRNAs were identified and many of them were also involved in ethylene pathway. These findings lay the foundation for exploring the role of small RNAs in ethylene signaling pathway in the plant.

AUTHOR CONTRIBUTIONS

JZ and LG designed the research; YW and JZ carried out the experiments; JZ, YW, QW, and ZJ analyzed the results; YW wrote the manuscript; JZ, BZ, and YL modified the manuscript; all authors have read and approved the manuscript for publication.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 31401536), the National Key Research and Development Program of China (No. 2016YFD0400901), the China Agriculture Research System Project (No. CARS-25), the Special Fund for Agro-scientific Research in the Public Interest (No. 201203095), the Young Investigator Fund of Beijing Academy of Agricultural and Forestry Sciences (No. 201404).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00527/full#supplementary-material

Supplementary Figure S1 | The expressions of *LeERF1* wild (CK1, CK2), sense-*LeERF1* (F1, F2), and antisense-*LeERF1* (R1, R2) tomato fruit.

Supplementary Figure S2 | The pipeline for the systematic identification of small RNAs in tomato.

Supplementary Figure S3 | Expression profiles of the known and putative novel miRNAs in wild (CK1, CK2), overexpression sense-*LeERF1* (F1, F2), and antisense-*LeERF1* (R1, R2) tomato fruit.

Supplementary Table S1 | Known miRNAs identified in tomato.

Supplementary Table S2 | Less conserved miRNAs identified in tomato.

Supplementary Table S3 | Putative novel miRNAs identified in tomato.

Supplementary Table S4 | Phased small RNAs found in tomato.

Supplementary Table S5 | Nat-siRNAs found in tomato.

Supplementary Table S6 | Target genes of the miRNAs involved in ethylene by bioinformatic prediction.

Supplementary Table S7 | Target genes of the miRNAs involved in ethylene by degradome sequencing.

Supplementary Table S8 | Target genes of the nat-siRNAs involved in ethylene.

REFERENCES

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr. Biol.* 18, 758–762. doi: 10.1016/j.cub.2008.04.042

An, F. M., Hsiao, S. R., and Chan, M. T. (2011). Sequencing-based approaches reveal low ambient temperature- responsive and tissue-specific microRNAs in phalaenopsis orchid. *PLoS ONE* 6:e18937. doi: 10.1371/journal.pone.0018937

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

Axtell, M. J., Snyder, J. A., and Bartell, D. P. (2007). Common functions for diverse small RNAs of land plants. *Plant Cell* 19, 1750–1769. doi: 10.1105/tpc.107.051706

Baek, D., Park, H. C., Kima, M. C., and Yun, D. J. (2013). The role of Arabidopsis MYB2 in miR399f-mediated phosphate-starvation response. *Plant Signal. Behav.* 8:e23488. doi: 10.4161/psb.23488

Candar-Cakir, B., Arican, E., and Zhang, B. H. (2016). Small RNA and degradome deep sequencing reveals drought-and tissue-specific micrornas and their important roles in drought-sensitive and drought-tolerant tomato genotypes. *Plant Biotechnol. J.* 14, 1727–1746. doi: 10.1111/pbi.12533

Cao, X., Wu, Z., Jiang, F., Zhou, R., and Yang, Z. (2014). Identification of chilling stress-responsive tomato microRNAs and their target genes by high-throughput sequencing and degradome analysis. *BMC Genomics* 15:1130. doi: 10.1186/1471-2164-15-1130

Carthew, R. W., and Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell* 136, 642–655. doi: 10.1016/j.cell.2009.01.035

Chen, H. M., Li, Y. H., and Wu, S. H. (2007). Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 3318–3323. doi: 10.1073/pnas.0611119104

Chen, X. (2009). Small RNAs and their roles in plant development. *Annu. Rev. Cell Dev. Biol.* 25, 21–44. doi: 10.1146/annurev.cellbio.042308.113417

Cheng, H. Y., Wang, Y., Tao, X., Fan, Y. F., Dai, Y., Yang, H., et al. (2016). Genomic profiling of exogenous abscisic acid-responsive microRNAs in tomato (*Solanum lycopersicum*). *BMC Genomics* 17:423. doi: 10.1186/s12864-016-2591-8

Chiou, T. J., Aung, K., Lin, S. I., Wu, C. C., Chiang, S. F., and Su, C. L. (2006). Regulation of phosphate homeostasis by microRNA in Arabidopsis. *Plant Cell* 18, 412–421. doi: 10.1105/tpc.105.038943

Couzigou, J. M., and Combier, J. P. (2016). Plant microRNAs: key regulators of root architecture and biotic interactions. *New Phytol.* 212, 22–35. doi: 10.1111/nph.14058

Dalmay, T. (2010). Short RNAs in tomato. *J. Integr. Plant Biol.* 52, 388–392. doi: 10.1111/j.1744-7909.2009.00871.x

Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., et al. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of miRNA genes. *PLoS ONE* 2:e219. doi: 10.1371/journal.pone.0000219

Fujii, H., Chiou, T. J., Lin, S. I., Aung, K., and Zhu, J. K. (2005). A miRNA involved in phosphate-starvation response in Arabidopsis. *Curr. Biol.* 15, 2038–2043. doi: 10.1016/j.cub.2005.10.016

Gao, C., Ju, Z., Cao, D., Zhai, B., Qin, G., Zhu, H., et al. (2015). MicroRNA profiling analysis throughout tomato fruit development and ripening reveals potential regulatory role of RIN on microRNAs accumulation. *Plant Biotechnol. J.* 13, 370–382. doi: 10.1111/pbi.12297

Giovannoni, J. J. (2004). Genetic regulation of fruit development and ripening. *Plant Cell* 16, S170–S180. doi: 10.1105/tpc.019158

Jain, M., Chevala, V. V., and Garg, R. (2014). Genome-wide discovery and differential regulation of conserved and novel microRNAs in chickpea via deep sequencing. *J. Exp. Bot.* 65, 5945–5958. doi: 10.1093/jxb/eru333

Jin, W., Li, N., Zhang, B., Wu, F., Li, W., Guo, A., et al. (2008). Identification and verification of microRNA in wheat (*Triticum aestivum*). *J. Plant Res.* 121, 351–355. doi: 10.1007/s10265-007-0139-3

Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* 57, 19–53. doi: 10.1146/annurev.arplant.57.032905.105218

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Li, F., Orban, R., and Baker, B. (2012). SoMART, a web server for plant miRNA, tasiRNA and target gene analysis. *Plant J.* 70, 891–901. doi: 10.1111/j.1365-313X.2012.04922.x

Li, Y., Zhu, B., Xu, W., Zhu, H., Chen, A., Xie, Y., et al. (2007). *LeERF1* positively modulated ethylene triple response on etiolated seedling, plant development and fruit ripening and softening in tomato. *Plant Cell Rep.* 26, 1999–2008. doi: 10.1007/s00299-007-0394-8

Li, Y. F., Zheng, Y., Addo-Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., et al. (2010). Transcriptome-wide identification of microRNA targets in rice. *Plant J.* 62, 742–759. doi: 10.1111/j.1365-313X.2010.04187.x

Li, Z. H., Peng, J. Y., Wen, X., and Guo, H. W. (2013). Ethylene-insensitive3 is a senescence-associated gene that accelerates age-dependent leaf senescence by directly repressing miR164 transcription in Arabidopsis. *Plant Cell* 25, 3311–3328. doi: 10.1105/tpc.113.113340

Liu, B., Li, P., Li, X., Liu, C., Cao, S., Chu, C., et al. (2005). Loss of function of OsDCL1 affects microRNA accumulation and causes developmental defects in rice. *Plant Physiol.* 139, 296–305. doi: 10.1104/pp.105.063420

Liu, H. H., Tian, X., Li, Y. J., Wu, C. A., and Zheng, C. C. (2008). Microarray-based analysis of stress-regulated microRNAs in *Arabidopsis thaliana*. *RNA* 14, 836–843. doi: 10.1261/rna.895308

Lu, C. W., Shao, Y., Li, L., Chen, A. J., Xu, W. Q., Wu, K. J., et al. (2011). Overexpression of *SlERF1* tomato gene encoding an ERF-type transcription activator enhances salt tolerance. *Rus. J. Plant Physiol.* 58, 118–125. doi: 10.1134/S1021443711010092

Lu, S., Sun, Y. H., and Chiang, V. L. (2008). Stress-responsive microRNAs in Populus. *Plant J.* 55, 131–151. doi: 10.1111/j.1365-313X.2008.03497.x

Mao, X., Cai, T., Olyarchuk, J., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

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

Morin, R. D., Aksay, G., Dolgosheina, E., Ebhardt, H. A., Magrini, V., Mardis, E. R., et al. (2008). Comparative analysis of the small RNA transcriptomes of Pinus contorta and *Oryza sativa*. *Genome Res.* 18, 571–584. doi: 10.1101/gr.6897308

Moxon, S., Jing, R., Szittya, G., Schwach, F., Rusholme Pilcher, R. L., Moulton, V., et al. (2008). Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Res.* 18, 1602–1609. doi: 10.1101/gr.080127.108

Osorio, S., Alba, R., Damasceno, C. M., Lopez-Casado, G., Lohse, M., Zanorc, M. I., et al. (2011). Systems biology of tomato fruit development: combined transcript, protein and metabolite analysis of tomato transcription factor (nor, rin) and ethylene receptor (Nr) mutants reveals novel regulatory interactions. *Plant Physiol.* 157, 405–425. doi: 10.1104/pp.111.175463

Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., et al. (2003). Control of leaf morphogenesis by microRNAs. *Nature* 425, 257–263. doi: 10.1038/nature01958

Pan, X. Q., Fu, D. Q., Zhu, B. Z., Lu, C. W., and Luo, Y. B. (2013). Overexpression of the ethylene response factor SlERF1 gene enhances resistance of tomato fruit to *Rhizopus nigricans*. *Postharvest Biol. Technol.* 75, 28–36. doi: 10.1016/j.postharvbio.2012.07.008

Park, J. M., Park, C. J., Lee, S. B., Ham, B. K., Shin, R., and Paek, K. H. (2001). Overexpression of the tobacco Tsi1 gene encoding an EREBP/AP2-type transcription factor enhances resistance against pathogen attack and osmotic stress in tobacco. *Plant Cell* 13, 1035–1046. doi: 10.1105/tpc.13.5.1035

Pashkovskiy, P. P., and Ryazansky, S. S. (2013). Biogenesis, evolution, and functions of plant microRNAs. *Biochemistry* 78, 627–637. doi: 10.1134/s0006297913060084

Pirrello, J., Jaimes-Miranda, F., Sanchez-Ballesta, M. T., Tournier, B., Khalil-Ahmad, Q., Regad, F., et al. (2006). Sl-ERF2 a tomato ethylene response factor involved in ethylene response and seed germination. *Plant Cell Physiol.* 47, 1195–1205. doi: 10.1093/pcp/pcj084

Qin, J., Zhao, J. Y., Zuo, K. J., Cao, Y. F., Ling, H., Sun, X. F., et al. (2004). Isolation and characterization of an ERF-like gene from *Gossypium barbadense*. *Plant Sci.* 167, 1383–1389. doi: 10.1016/j.plantsci.2004.07.012

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* 20, 3407–3425. doi: 10.1101/gad.1476406

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513–520. doi: 10.1016/s0092-8674(02)00863-2

Rubio-Somoza, I., Cuperus, J. T., Weigel, D., and Carrington, J. C. (2009). Regulation and functional specialization of small RNA-target nodes during plant development. *Curr. Opin. Plant Biol.* 12, 622–627. doi: 10.1016/j.pbi.2009.07.003

Pilcher, R. L., Moxon, S., Pakseresht, N., Moulton, V., Manning, K., Seymour, G., et al. (2007). Identification of novel small RNAs in tomato (*Solanum lycopersicum*). *Planta* 226, 709–717. doi: 10.1007/s00425-007-0518-y

Sanan-Mishra, N., Kumar, V., Sopory, S. K., and Mukherjee, S. K. (2009). Cloning and validation of novel miRNA from basmati rice indicates cross talk between abiotic and biotic stresses. *Mol. Genet. Genomics* 282, 463–474. doi: 10.1007/s00438-009-0478-y

Shivaprasad, P. V., Chen, H. M., Patel, K., Bond, D. M., Santos, B. A., and Baulcombe, D. C. (2012). A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell* 24, 859–874. doi: 10.1105/tpc.111.095380

Sunkar, R., Chinnusamy, V., Zhu, J., and Zhu, J. K. (2007). Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends Plant Sci.* 12, 301–309. doi: 10.1016/j.tplants.2007.05.001

Sunkar, R., Girke, T., Jain, P. K., and Zhu, J. K. (2005). Cloning and characterization of microRNAs from rice(W). *Plant Cell* 17, 1397–1411. doi: 10.1105/tpc.105.031682

Talmor-Neiman, M., Stav, R., Frank, W., Voss, B., and Arazi, T. (2006). Novel micro-RNAs and intermediates of micro-RNA biogenesis from moss. *Plant J.* 47, 25–37. doi: 10.1111/j.1365-313X.2006.02768.x

Thiebaut, F., Rojas, C. A., Grativol, C., Motta, M. R., Vieira, T., Regulski, M., et al. (2014). Genome-wide identification of microRNA and siRNA responsive to endophytic beneficial diazotrophic bacteria in maize. *BMC Genomics* 15:766. doi: 10.1186/1471-2164-15-766

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

Wang, H., Huang, Z., Chen, Q., Zhang, Z., Zhang, H., Wu, Y. M., et al. (2004). Ectopic overexpression of tomato JERF3 in tobacco activates downstream gene expression and enhances salt tolerance. *Plant Mol. Biol.* 55, 183–192. doi: 10.1007/s11103-004-0113-6

Wang, X., Kong, H., and Ma, H. (2009). F-box proteins regulate ethylene signaling and more. *Genes Dev.* 23, 391–396. doi: 10.1101/gad.1781609

Wu, G., Park, M. Y., Conway, S. R., Wang, J. W., Weigel, D., and Scott Poethig, R. (2009). The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. *Cell* 138, 750–759. doi: 10.1016/j.cell.2009.06.031

Wu, K. Q., Tian, L. N., Hollingworth, J., Brown, D. C. W., and Miki, B. (2002). Functional analysis of tomato Pti4 in Arabidopsis. *Plant Physiol.* 128, 30–37. doi: 10.1104/pp.010696

Wu, P., Wu, Y., Liu, C. C., Liu, L. W., Ma, F. F., Wu, X. Y., et al. (2016). Identification of Arbuscular Mycorrhiza (AM)-Responsive microRNAs in Tomato. *Front. Plant Sci.* 7:429. doi: 10.3389/fpls.2016.00429

Xie, F., Jones, D. C., Wang, Q., Sun, R., and Zhang, B. (2015). Small RNA sequencing identifies miRNA roles in ovule and fibre development. *Plant Biotechnol. J.* 13, 355–369. doi: 10.1111/pbi.12296

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S. A., and Carrington, J. C. (2005). Expression of Arabidopsis miRNA genes. *Plant Physiol.* 138, 2145–2154. doi: 10.1104/pp.105.062943

Yant, L., Mathieu, J., Dinh, T. T., Ott, F., Lanz, C., Wollmann, H., et al. (2010). Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2. *Plant Cell* 22, 2156–2170. doi: 10.1105/tpc.110.075606

Yoshikawa, M., Peragine, A., Park, M. Y., and Poethig, R. S. (2005). A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. *Genes Dev.* 19, 2164–2175. doi: 10.1101/gad.1352605

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:r14. doi: 10.1186/gb-2010-11-2-r14

Zhai, J. X., Jeong, D. H., De Paoli, E., Park, S., Rosen, B. D., Li, Y., et al. (2011). MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev.* 25, 2540–2553. doi: 10.1101/gad.177527.111

Zhang, J. Y., Broeckling, C. D., Sumner, L. W., and Wang, Z. Y. (2007). Heterologous expression of two *Medicago truncatula* putative ERF transcription factor genes, WXP1 and WXP2, in Arabidopsis led to increased leaf wax accumulation and improved drought tolerance, but differential response in freezing tolerance. *Plant Mol. Biol.* 64, 265–278. doi: 10.1007/s11103-007-9150-2

Zhang, R. X., Marshall, D., Bryan, G. J., and Hornyik, C. (2013). Identification and characterization of miRNA transcriptome in potato by high-throughput sequencing. *PLoS ONE* 8:e57233. doi: 10.1371/journal.pone.0057233

Zhang, X., Zou, Z., Zhang, J., Zhang, Y., Han, Q., Hu, T., et al. (2011). Over-expression of sly-miR156a in tomato results in multiple vegetative and reproductive trait alterations and partial phenocopy of the sft mutant. *FEBS Lett.* 585, 435–439. doi: 10.1016/j.febslet.2010.12.036

Zhou, X. F., Sunkar, R., Jin, H., Zhu, J. K., and Zhang, W. X. (2009). Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*. *Genome Res.* 19, 70–78. doi: 10.1101/gr.084806.108

Zuo, J., Fu, D., Zhu, Y., Qu, G., Tian, H., Zhai, B., et al. (2013). SRNAome parsing yields insights into tomato fruit ripening control. *Physiol. Plant.* 149, 540–553. doi: 10.1111/ppl.12055

Zuo, J., Zhu, B., Fu, D., Zhu, Y., Ma, Y., Chi, L., et al. (2012). Sculpting the maturation, softening and ethylene pathway: the influences of microRNAs on tomato fruits. *BMC Genomics* 13:7. doi: 10.1186/1471-2164-13-7

Zuo, J. H., Wang, Q., Han, C., Ju, Z., Cao, D., Zhu, B., et al. (2016). SRNAome and degradome sequencing analysis reveals specific regulation of sRNA in response to chilling injury in tomato fruit. *Physiol Plant.* doi: 10.1111/ppl.12509. [Epub ahead of print].

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 Genetics

ORIGINAL RESEARCH
published: 30 September 2016
doi: 10.3389/fgene.2016.00176

Retention and Molecular Evolution of Lipoxygenase Genes in Modern Rosid Plants

Zhu Chen[1], *Danmei Chen*[1], *Wenyuan Chu*[1], *Dongyue Zhu*[1], *Hanwei Yan*[1,2] *and Yan Xiang*[1,2]*

[1] *Laboratory of Modern Biotechnology, Anhui Agricultural University, Hefei, China,* [2] *Key Laboratory of Biomass Improvement and Conversion, Anhui Agriculture University, Hefei, China*

Whole-genome duplication events have occurred more than once in the genomes of some rosids and played a significant role over evolutionary time. Lipoxygenases (LOXs) are involved in many developmental and resistance processes in plants. Our study concerns the subject of the *LOX* gene family; we tracked the evolutionary process of ancestral *LOX* genes in four modern rosids. Here we show that some members of the *LOX* gene family in the *Arabidopsis* genome are likely to be lost during evolution, leading to a smaller size than that in *Populus*, *Vitis,* and *Carica*. Strong purifying selection acted as a critical role in almost all of the paralogous and orthologous genes. The structure of *LOX* genes in *Carica* and *Populus* are relatively stable, whereas *Vitis* and *Arabidopsis* have a difference. By searching conserved motifs of *LOX* genes, we found that each sub-family shared similar components. Research on intraspecies gene collinearity show that recent duplication holds an important position in *Populus* and *Arabidopsis*. Gene collinearity analysis within and between these four rosid plants revealed that all *LOX* genes in each modern rosid were the offspring from different ancestral genes. This study traces the evolution of *LOX* genes which have been differentially retained and expanded in rosid plants. Our results presented here may aid in the selection of special genes retained in the rosid plants for further analysis of biological function.

Keywords: lipoxygenase, purifying selection, gene duplication, syntenic chromosomal block, evolutionary history

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Matteo Benelli,
University of Trento, Italy
Giovanni Bacci,
University of Florence, Italy
Debjit Ray,
Sandia National Laboratories, USA

***Correspondence:**
Yan Xiang
xiangyanahau@sina.com

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

Received: *06 June 2016*
Accepted: *16 September 2016*
Published: *30 September 2016*

Citation:
Chen Z, Chen D, Chu W, Zhu D, Yan H and Xiang Y (2016) Retention and Molecular Evolution of Lipoxygenase Genes in Modern Rosid Plants. Front. Genet. 7:176. doi: 10.3389/fgene.2016.00176

INTRODUCTION

Whole-genome duplications (WGDs) bring a huge impact on genome sizes of many angiosperms and may have provided the genetic material for evolutionary novelties (Sémon and Wolfe, 2007; Jaillon et al., 2009). Duplication events are usually followed by gene loss (Bowers et al., 2003), nucleotide divergence (Bowers et al., 2003) and structural rearrangements (Hufton and Panopoulou, 2009). It has long been hypothesized that the ancient genome triplication event happened to a single common ancestor of *Arabidopsis-Populus-Vitis-Carica* and finally caused a paleohexaploid (Tang et al., 2008a).Other than that, the two recent paleopolyploidies that have affected *Arabidopsis* are β- and α- duplications. At-α was a recent event, and the At-β was an intermediate event (Barker et al., 2009). In *Populus*, there was a single genome-wide event. This duplication event was called the "salicoid" duplication event (P-duplication; Tuskan et al., 2006). *Vitis vinifera* and *Carica papaya* each have only γ-triplication event and no other polyploidies (Tang et al., 2008a).

Polyploidy has been and continues to have an extensive effect on the number or type of genes in plant evolution (Adams and Wendel, 2005). Analysis of the differential retention and expansion of ancestral genes in modern plants provide an informative and robust way to resolve relationships among many lineages (Rokas and Holland, 2000). In this study, we will take the Lipoxygenase gene family as an example and discuss the differential retention and expansion of ancestral genes in four rosids.

Lipoxygenases (LOXs) exist extensively within plants and animals (Brash, 1999). The best known function of these enzymes are to synthesize lipid mediators (Brash, 2015): as we know, leukotrienes and resolvins are in animals, jasmonates and short-chain aldehydes are in plants. LOXs catalyze polyenoic fatty acids PUFAs (Feussner and Kühn, 2000) like linoleic acid (LA), α-linolenic acid (α-LeA), or arachidonic acid, which have a (1Z, 4Z)-pentadiene moiety. According to their positional specificity of linoleic acid oxygenation, lipoxygenases have been divided into group 9-LOX and group 13-LOX (Hildebrand, 1989). LOXs contain a region rich in histidine residues, which was previously observed to be highly conserved in the primary structure of isozymes. This region contains a cluster of 5 His residues in the form of His-(X)4-His-(X)4-His-(X)17-His-(X)8-His (Shibata et al., 1987; Steczko et al., 1992; Boyington et al., 1993; Feussner and Wasternack, 2002).

Lipoxygenases involved in food-related applications during bread-making and production of the aroma are controlled by enzymes, which were found related to the formation of volatile compounds (Leenhardt et al., 2006). Studies have shown that extractable activities of enzymes are major factors that can affect the degrading efficiency of carotenoid pigments during the kneading step of bread-making in each of the three cultivated wheat species. Lipoxygenases also have a negative relationship with the color, off-flavor and antioxidant status of plant-based foods. Studies on soy-based foods have demonstrated that lipoxygenases are responsible for the off-flavor associated with biological components present in soybean (Leenhardt et al., 2006). So far, there is sufficient evidence to prove that lipoxygenase is the most crucial element in plant defense responses (Baysal and Demirdöven, 2007; Bannenberg et al., 2009). In recent years, one *LOX* gene in *Arabidopsis* (*AtLOX2*) was thought to function exclusively in jasmonates (JA) biosynthesis upon wounding (Van Loon et al., 2006). The study is backed up by recent findings in apple which showed that *MdLOX5* gene was more likely to be responsible for aphid tolerance or resistance (Vogt et al., 2013).

Lipoxygenase genes are chosen for their biological significance. In our research, taking lipoxygenases as an example, we studied the expansion of these genes in four species. Previous analysis showed that one or more paleopolyploidy events which had an impact on these four modern rosid genomes, fluctuate remarkably in size and arrangement. Our results trace the differential retention and expansion of the ancestral Lipoxygenases in *Arabidopsis thaliana*, *Populus trichocarpa*, *V. vinifera* and *C. papaya* and help facilitate the extrapolation of the evolutionary process.

MATERIALS AND METHODS

Ethics Statement

No specific permits were required for the described field studies. No specific permissions were required for these locations and activities. The location is not privately-owned or protected in any way and the field studies did not involve endangered or protected species.

Database Search and Sequence Retrieval

LOX genes were identified following the method described by (Podolyan et al., 2010; Umate, 2014; Chen et al., 2015). Protein and cDNA sequences of LOX genes in *Arabidopsis* were obtained from the *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org/, release 10.0). Protein and cDNA sequences of *P. trichocarpa*, *V. vinifera,* and *C. papaya* were downloaded from Phytozomev.11.0 database. The respective genome sequence sites are as follows: *P. trichocarpa*, *V. vinifera*, and *C. papaya* (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV11).
Local Blast searching was performed using *Arabidopsis* LOX proteins as queries for the identification of LOX genes from *Carica* and *Vitis*, and then using the resulting poplar and grapevine sequences as secondary queries. To obtain good gene models of *CpLOX* genes, all-against-all nucleotide sequence similarity searches were performed between the gene models and EST sequences using BLASTN software (Supplementary Table 1). Besides, we also worked on the multiple sequence alignment with CpLOX proteins and 70 experimantally verified gene models (Supplementary Tables 2, 3). All of the obtained genes were further manually analyzed to confirm the presence of the LOX domain (PF00305) and PLAT/LH2 (polycystin-1, lipoxygenase, α-toxin domain, or the lipoxygenase homology) domain (PF01477) in the Pfam HMM database (http://pfam.sanger.ac.uk/) (Finn et al., 2006) and InterPro (European Bioinformatics Institute) (http://www.ebi.ac.uk/interpro/scan.html) (Supplementary Table 4; Mulder et al., 2007). Redundant sequences with different identification numbers and the same chromosome locus were removed.

Phylogenetic Trees Construction

Complete protein sequences of LOX in the four plant species were aligned with the aid of ClustalW (Larkin et al., 2007). The phylogenetic tree was constructed by MEGA version 6.0 software with the minimum evolution (ME) method (Tamura et al., 2013). Bootstrap analysis with 1000 replicates was performed to calculate the reliability of the ME tree. To confirm the robustness of the ME tree, we also constructed other phylogenetic trees by using the Neighbor-Joining (NJ) method.

Exon–Intron Structural Analysis and Identification of Conserved Motifs

The exon–intron structure positions of *LOX* genes were generated online using the online program Gene Structure Display Server (GSDS; http://gsds.cbi.pku.edu.cn/; Guo et al.,

2007) by alignment of the cDNAs with their corresponding genomic DNA sequences. To identify the conserved motifs of the *LOX* genes in four rosid plants, the structural motif annotation was employed using the MEME (Multiple Em for Motif Elicitation, Version 4.11.1) program (Bailey et al., 2006) with the following parameters: the maximum number of motifs was set at 20, and the optimum motif widths were set between six and 200 residues. Structural motif annotation was provided by using the Batch search tool in Pfam program.

Identification of Paralogs and Orthologs

Paralogs and orthologs were identified by using the same procedure described in Blanc and Wolfe (2004). This method was performed by running a BLASTN (Altschul et al., 1997) for all nucleotide sequences for each species. A pair of matching sequences were defined as pairs of paralogs when the identity was more than 40% and the alignment covered over 300 bp. To identify putative orthologs between two species, for example A and B, each sequence from species A was searched against all sequences from species B using BLASTN. At the same time, each sequence from species B was searched against all sequences from species A. The two sequences were defined as orthologs whose reciprocal best hits were each within >=300 bp of the two sequences aligned.

ω and Ks Analysis

First, pairwise protein sequence alignment was performed using MUSCLE (Edgar, 2004).Then, used in conjunction with protein alignments, CDS sequence, and an in-house PERL script, the input file format of KaKs_Calculator2.0 could be got. Finally, the input file were converted into computation of Ks (synonymous substitution rate) and Ka (non-synonymous substitution rate) values using KaKs_Calculator2.0 (Wang et al., 2010). To further assess whether positive selection acts upon specific sites, the gene pairs for all the paralogs and orthologs were used to calculate the ω, where $\omega = Ka/Ks$.

Intraspecies and Interspecies Microsynteny Analysis

Microsynteny analysis within four species was detected by MicroSyn software (Cai et al., 2011). At the beginning of the generated MSY file step, three property files are needed: the gene list file, the CDS file and the gene identifier file. The microsynteny graphic file was provided by loading the three files. Then MicroSyn creates a homologous relationship among all genes. Finally, the microsynteny graphic file was provided by the software. In order to analyze duplication of *LOX* genes of the four rosid plants we researched the expansion of *LOX* genes through segmental or whole-genome duplications (S/WGD) for *LOX* genes in each species by using the plant genome duplication database (PGDD; http://chibba.agtec.uga.edu/duplication/; Lee et al., 2013). In order to determine whether the *LOX* genes of the four rosid plants arose from a large-scale duplication event (duplicated blocks derived from whole-genome or segmental duplication) or tandem duplication, genome-wide analysis was undertaken to examine whether *LOX* genes occurred within duplicated blocks. We researched the expansion of LOX genes through segmental or whole-genome duplications (S/WGD) for *LOX* genes in each species by using the plant genome duplication database (PGDD; http://chibba.agtec.uga.edu/duplication/). The *LOX* genes duplicated through S/WGDs were inferred on the basis of gene collinearity on syntenic blocks. First, from the PGDD, we download the file containing collinear block information within and between the four rosids. Then, all download blocks information were imported into MySQL, and the *LOX* gene ID were used as the query to perform a search in these species. The LOX genes duplicated through S/WGDs were identified in this way. In this analysis, the counterparts of a particular *LOX* gene on an SCB may have been retained as LOXs, subfunctionalized into non LOXs (indicated by an "N" before the first letter in the code name). LOX genes expanded through tandem duplication (TD) were inferred following the method that (1) belong to *LOX* gene family, (2) are located within 60 kb each other (data comes from Phytozome), and (3) are separated by five or fewer gene loci (non LOXs). Syntenic blocks between species were identified using the MCscanX (Wang et al., 2012) software with default parameters. To categorize the expansion of the *LOX* gene families, the positions of the *LOX* genes in the blocks were *A. thaliana*, *P. trichocarpa*, *V. vinifera,* and *C. papaya*. Circos software was used to draw the syntenic diagram (Krzywinski et al., 2009).

RESULTS

LOX Genes in Four Modern Rosids

Based on the previous studies, we obtained 6 and 20 putative LOX genes from the *Arabidopsis*, and poplar, respectively. In a recently published report, a total of 18 LOX genes were identified in *Vitis*. By removing pseudogenes, 13 LOX genes were identified in the *Vitis* genome. In this study, by removing psedogenes we further filtered five additional LOX genes in *Vitis* and changed the total member into 13. To identify LOX in *Carica*, we performed a search against the genome database with BlastP using AtLOX protein sequences as queries. Finally, 11 LOX genes were identified in *Carica*. The detail information of each LOX genes are listed in **Table 1**.

To date, four studied rosids have been suggested to possess paleohexaploidy in a common ancestor (Jaillon et al., 2007). Based on previous results, the multiplicity ratio for an ancestral gene comparison in the genomes of four species should be 4:2:1:1. But in our research results, the number of LOX genes in *Arabidopsis* is far fewer than that estimated for other plant species. Previous studies suggest that the *V. vinifera* genome is by far the closest to the ancestral arrangement such that the ancestral gene order can be deduced from this species with no difficulty. The ratio of LOX genes for the four species is 0.5: 1.5: 0.85: 1 when the number of LOX genes in *V. vinifera* is used as a benchmark. In addition to *A. thaliana*, this result is basically in line with the expected current ratios of LOXs.

LOX Paralogs and Orthologs

We detected 33.3% (2/6, *Arabidopsis*), 72.7% (8/11, *Carica*), 95% (19/20, *Populus*), and 76.9% (10/13, *Vitis*) *LOX* genes in

TABLE 1 | Detailed information about the LOX gene family in rosid plants.

Species	Gene name	Gene ID	Chr.	Location coordinates(5′–3′)	Protein length(a.a.)	ORF length(bp)
Carica papaya	*CpLOX1*	evm.TU.supercontig_8.58	supercontig_8	393,002–397,307	797	2394
	CpLOX2	evm.TU.supercontig_17.119	supercontig_17	1,496,985–1,501,648	816	2451
	CpLOX3	evm.TU.supercontig_25.128	supercontig_25	1,317,447–1,322,293	925	2778
	CpLOX4	evm.TU.supercontig_32.35	supercontig_32	482,778–486,631	854	2565
	CpLOX5	evm.TU.supercontig_32.64	supercontig_32	770,894–774,557	855	2568
	CpLOX6	evm.TU.supercontig_43.30	supercontig_43	308,246–311,725	922	2769
	CpLOX7	evm.TU.supercontig_48.63	supercontig_48	351,329–356,020	867	2604
	CpLOX8	evm.TU.supercontig_58.126	supercontig_58	1,216,903–1,220,688	849	2550
	CpLOX9	evm.TU.supercontig_58.127	supercontig_58	1,233,011–1,236,967	849	2550
	CpLOX10	evm.TU.supercontig_458.2	supercontig_458	18,363–22,072	788	2367
	CpLOX11	evm.TU.supercontig_458.4	supercontig_458	24,745–28,879	917	3754
Arabidopsis thaliana	*AtLOX1*	AT1G55020	1	20,525,708–20,530,273	859	2580
	AtLOX2	AT3G45140	3	16,525,410–16,529,352	896	2691
	AtLOX3	AT1G17420	1	5,977,411–5,981,480	919	2760
	AtLOX4	AT1G67560	1	25,319,899–25,324,264	917	2754
	AtLOX5	AT3G22400	3	7,926,879–7,931,351	886	2661
	AtLOX6	AT1G72520	1	27,308,515–27,312,754	926	2781
Vitis vinifera	*VvLOX1*	GSVIVT01010359001	1	19,772,666–19,777,638	920	2763
	VvLOX2	GSVIVT01017943001	5	4,934,967–4,939,395	751	2256
	VvLOX3	GSVIVT01025342001	6	1,774,659–1,781,744	817	2454
	VvLOX4	GSVIVT01025340001	6	1,853,936–1,868,694	872	2619
	VvLOX5	GSVIVT01025339001	6	1,875,239–1,882,842	901	2706
	VvLOX6	GSVIVT01025328001	6	1,988,366–1,989,826	335	1008
	VvLOX7	GSVIVT01005730001	7	13,887,191–13,893,238	641	1926
	VvLOX8	GSVIVT01016738001	9	811,736–816,741	927	2784
	VvLOX9	GSVIVT01032029001	13	23,366,475–23,371,929	866	2601
	VvLOX10	GSVIVT01000083001	14	3,311,501–3,315,829	738	2217
	VvLOX11	GSVIVT01000084001	14	3,315,947–3,324,623	900	2703
	VvLOX12	GSVIVT01003798001	chr7_random	201,678–209,816	619	1860
	VvLOX13	GSVIVT01005215001	Un	19,276,130–19,281,690	533	1602
Populus trichocarpa	*PtLOX1*	Potri.001G015300	1	1,076,313–1,081,197	898	2697
	PtLOX2	Potri.001G015400	1	1,090,420–1,098,069	902	2709
	PtLOX3	Potri.001G015500	1	1,105,670–1,110,895	898	2697
	PtLOX4	Potri.001G015600	1	1,118,168–1,123,930	898	2697
	PtLOX5	Potri.001G167700	1	14,106,872–14,112,847	923	2772
	PtLOX6	Potri.003G067600	3	9,576,888–9,583,048	925	2778
	PtLOX7	Potri.005G032400	5	2,425,802–2,431,106	866	2601
	PtLOX8	Potri.005G032600	5	2,435,033–2,439,658	796	2391
	PtLOX9	Potri.005G032700	5	2,451,619–2,456,194	866	2601
	PtLOX10	Potri.005G032800	5	2,462,946–2,469,256	863	2592
	PtLOX11	Potri.008G151500	8	10,276,751–10,281,394	880	2643
	PtLOX12	Potri.008G178000	8	12,146,645–12,151,320	927	2784
	PtLOX13	Potri.009G022400	9	3,421,114–3,425,183	901	2706
	PtLOX14	Potri.010G057100	10	8,651,258–8,655,916	926	2781
	PtLOX15	Potri.010G089500	10	11,305,668–11,310,367	881	2646
	PtLOX16	Potri.013G022000	13	1,454,479–1,459,136	871	2616
	PtLOX17	Potri.013G022100	13	1,461,474–1,466,287	862	2589
	PtLOX18	Potri.014G018200	14	1,725,218–1,731,715	860	2583
	PtLOX19	Potri.014G177200	14	14,542,431–14,547,953	860	2583
	PtLOX20	Potri.017G046200	17	3,854,572–3,860,007	898	2697

each species involved in paralogous duplication (Supplementary Table 5). Thus, over half of the *LOX*s were closely bound up with intra-specific duplication in *Carica*, *Vitis,* and *Populus*. By contrast, there was just one pair of *LOX* paralogs in *Arabidopsis*, although this species has been expanded by three rounds of whole-genome duplication. The higher ratio in *Populus* reflects the preferential gene retention after multiple rounds of WGD. Our results show that *Populus* and *Vitis* shared the most orthologous pairs, of up to 26 pairs of orthologous *LOX*s. We only got one pair of orthologous *LOX*s between *Arabidopsis* and *Vitis*. Two pairs of orthologous *LOX*s were detected between *Arabidopsis* and *Populus*. After comparing *Carica* and other three species we found no orthologous *LOX*s between them.

In order to better understand the evolutionary constraints acting among the four rosid species, we measured the Ka/Ks ratios for these pairs of *LOX* paralogs and orthologs. The ratio of non-synonymous substitutions per non-synonymous site (Ka) vs. the synonymous substitutions per synonymous site (Ks) is an indicator of the history of selection (Yang and Bielawski, 2000). If Ka/Ks < 1, it suggests that the gene is undergoing purifying selection. When Ka/Ks > 1, it means there is accelerated devolution with positive selection, and Ka/Ks $= 1$ suggests neutral selection. A summary of Ka/Ks for *LOX* paralogous and orthologous pairs is shown in Supplementary Table 5. The resulting pairwise comparison data showed the Ka/Ks values of only one Vv paralogous pair larger than 1. The relatively higher Ka/Ks ratio of *VvLOX6*/11 suggests that they may have experienced relatively rapid evolution following duplication. There were two *Vitis* pairs (*VvLOX6/10* and *VvLOX7/12*) and one *Populus* pair *(PtLOX18/19)* that were larger than 0.5 but less than 1, while all of the remaining Ka/Ks ratios were less than 0.5, suggesting that the LOX family has mainly undergone strong purifying selection and these *LOX* genes are slowly evolving at the protein level. Our calculations shows that all orthologous *LOX*s between species were less than 1.

Expansion and Structural Characteristics of the *LOX* Genes in Four Rosid Plants

To investigate the extent of the expansion of the *LOX* genes in rosid plants, we performed a joint phylogenetic analysis with MEGA using the ME method (**Figure 1**) and the NJ method (Supplementary Figure 1). The ME and NJ trees show identical topologies. As mentioned above, plant lipoxygenases are clustered into two groups (9-LOX and 13-LOX). In our study, a total of 50 genes formed two distinct clades and are in agreement with the previously studied results (Brash, 1999; **Figure 1**). As shown in **Figure 1**, 9-LOX consisted of 20 *LOX* genes from four modern rosids; two from *Arabidopsis*, eight from *Populus*, six from *Carica*, and four from *Vitis*. This clade is composed of four sub-clades, one of which includes purely six *Populus LOX* genes. Yet there is another sub-clade simply containing four *Carica LOX* genes. The rest includes *LOX* genes from two or more species. Paralogous groups in this clade are *CpLOX2/4/5/7* and *CpLOX8/9* from *Carica*; *PtLOX7/8/9/10/16/17* and PtLOX11/15 from *Populus*; *VvLOX6/10/11* from *Vitis*. Beyond that there are three orthologous pairs shared by the four modern rosids. The remaining *LOX*s are placed in the 13-LOXs group. Paralogous groups in this clade were A3-A6, from *Arabidopsis*; PtLOX1/2, *PtLOX3/4/20, PtLOX17/26, and PtLOX10/12* from *Populus*; *VvLOX3/4/5, VvLOX4/9 and VvLOX7/12* from *Vitis*. In addition, this clade contained only one paralogous pair from *Carica, CpLOX10/11*. In this group, there are seven orthologous pairs shared by the four modern rosids. Besides, the genetic distances among the two LOX sub-families were studied and the result showed that the genetic distance of 9-LOX genes was smaller than 13-LOX genes, indicating that 9-LOX genes are more closely related to each other.

For a better understanding of the structural diversity of *LOX* genes, using the structures of *LOX* genes we generated the exon-intron architecture of each *LOX* gene in four rosid plants (**Figure 2**). Overall, the structures of *LOX* genes in *Carica* and *Populus* were conserved. But some changes take place in the *AtLOX* and *VvLOX* genes. The detailed structural analysis of the exon/intron are presented in **Figure 3**. Of the four species surveyed, *Carica* and *Populus LOX* genes are in a similar position with eight or nine exons, and the number of exons in *Arabidopsis* range from six to nine. *VvLOX* genes are much more dramatic, *VvLOX9/10/11* have the highest number of exons at 11, but in the same species *VvLOX6* also has the least number of exons at five. We further analyzed the exon/intron structure of the *LOX* orthologous and paralogous gene pairs discussed previously. The results showed that majority of these gene pairs have different exon numbers. Among paralogous gene pairs, the structure rationality changes obviously in *VvLOX6/10*. Simultaneously, by comparing the orthologous pairs, we found that all the differences come from *Populus* and *Vitis*.

We also studied the conserved motifs of *LOX* genes because of its particularity and the importance to the diversified functions of *LOX* genes. Therefore, we used the MEME web server to find the relatively conserved motifs which are shared with the 50 LOX proteins. In total, 20 distinct conserved motifs were found (**Figure 3**, Supplementary Table 6), and the relevant information is shown in Supplementary Table 6. Each of the putative motifs is well commented by searching in Pfam database. In detail, motifs 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 14, 15, and 19 are associated with the Lipoxygenase domain; motif 9 is found to encode the PLAT domain; motif 4 is thought to be involved in foring the PLAT and Lipoxygenase domain. However, the other motifs have no functional annotation. As illustrated in **Figure 3**, most LOX members belong to the same sub-family and are alike in motif compositions, suggesting that a lot of similarity may have many overlapping parts from a functional perspective. Motif 5 is widely presented in all fifty LOX proteins. Motif 15 and motif 20 are unique to the proteins in the 9-LOX clade. The former is considered for all of the components of the Lipoxygenase domain. Even though the function of motif 20 is still unknown, we still think that these motifs might be important to the functions of unique LOX proteins due to their specificity. To some extent, these specific motifs may help us to understand the functional divergence of *LOX* genes during evolutionary history.

Expansion Manners of *LOX* Genes within Four Rosid Plants

In order to probe the relationship between the genetic divergences within the *LOX* gene family and the corresponding expansion patterns, we further analyzed the gene duplication

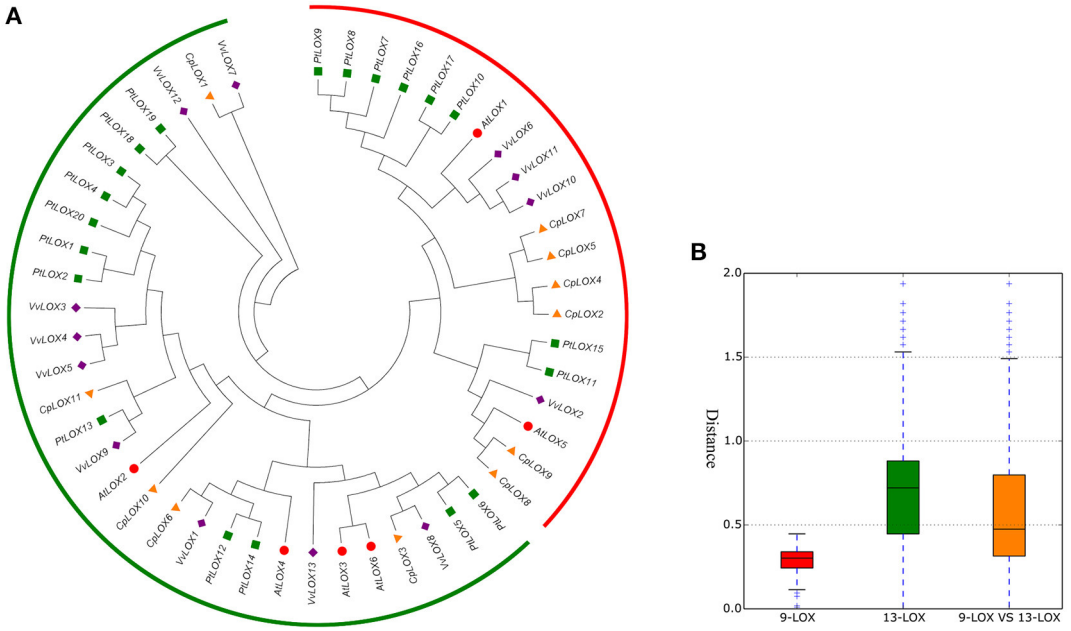


FIGURE 1 | Phylogenetic relationships among *LOX* genes in four rosid plants (A) and the genetic distance among different sub-families of *LOX* genes (B). Gene sub-families are indicated with different colors. Taxon labels are depicted in red for the 9-LOX clade and in green for 13-LOX clade. In **(A)**, the phylogenetic tree was constructed using Minimum-evolution (ME) using MEGA6. The LOXs of *Arabidopsis* are indicated by red circles, *Carica* are indicated by orange triangles, *Populus* are indicated by green squares and *Vitis* are indicated by purple rhombus symbols.

events within each species. As previously mentioned, rosid plants have experienced at least one polyploidy event. These events may have lasting implications for the evolution of *LOX* gene families. We used the MicroSyn software to investigate this possibility. If two members of the same gene family are homologous pairs, and three or more of the 50 upstream and downstream neighboring genes are also considered to be homologous pairs, we defined these two regions as those resulting from a duplication event. The number of *LOX* genes that arose from duplication events varied among the four rosids. Our survey results showed that 10 collinear gene pairs occurred in the *Populus* genome and a total of four collinear gene pairs occurred in the *Vitis* genome; however, there was only one collinear gene pair in both *Arabidopsis* and *Carica* genomes (**Figure 4**).

In *Arabidopsis*, one gene pair (*AtLOX3/AtLOX6*) was found to have conserved neighboring regions and no syntenic relationships were detected within the other four *AtLOX* genes. In *Carica*, the microsynteny between the *CpLOX8* and *CpLOX9* genes is extensive and this gene pair is located next to each other on the same supercontig 58, believed to have derived from tandem duplication events. In *Vitis*, one pair, *VvLOX2/VvLOX8,* shares a substantial collinear region. In addition, one gene cluster *VvLOX3/VvLOX4/VvLOX5* and one gene pair, *VvLOX10/VvLOX11* are located near each other on the same chromosomes and these gene pairs might be evolved from tandem duplication. In *Populus*, three gene pairs, *PtLOX7/PtLOX16, PtLOX11/PtLOX15,* and *PtLOX12/PtLOX14* share extraordinary conserved synteny, with less conserved collinear genes surrounding *PtLOX5/PtLOX6, PtLOX8/PtLOX16,* and *PtLOX15/PtLOX16*. Besides, one gene cluster, *PtLOX7/PtLOX8/PtLOX9/PtLOX10,* and one gene pair, *PtLOX18/PtLOX19,* appear to have evolved from tandem duplication events. Using this approach, we made a preliminary judgment on the duplication events within each species. To better examine the gene duplication events of *LOX* genes, we retrieved the syntenic chromosomal blocks (SCBs) associated with the expansion of *LOX* genes through S/WGDs from the plant genome duplication database. The results are consistent with the findings of most studies done in the MicroSyn software. The biggest difference occurs in *Vitis* and *Populus*. From the PGDD database, we found that the other two gene pairs *VvLOX2/VvLOX10* and *VvLOX3/VvLOX9* are associated with S/WGDs in *Vitis*. In *Populus*, two counterparts (*PtLOX1/PtLOXN1* and *PtLOX13/PtLOXN1*) of *LOXs* on SCBs are found to have sub-functionalized into other gene family members (indicated by the code name preceded by the letter "N"; Guo et al., 2014). Beyond that, the two gene pairs *PtLOX8/PtLOX16* and *PtLOX15/PtLOX16* achieved through MicroSyn software were not found in the PGDD database. Generally to consider the results from both ways, in current findings the number of *LOX* genes that arose from S/WGD are 10, five and two in *Populus*, *Vitis,* and *Arabidopsis*, respectively. Because the duplicated gene located on a SCB is simultaneous with another one, the median Ks value of duplicated genes in SCBs can be used to infer the dates of the large-scale duplication events. In this analysis, the duplicated gene pairs as well as the homologous genes in neighbor regions are used to date duplication events. The mean Ks values for each duplication

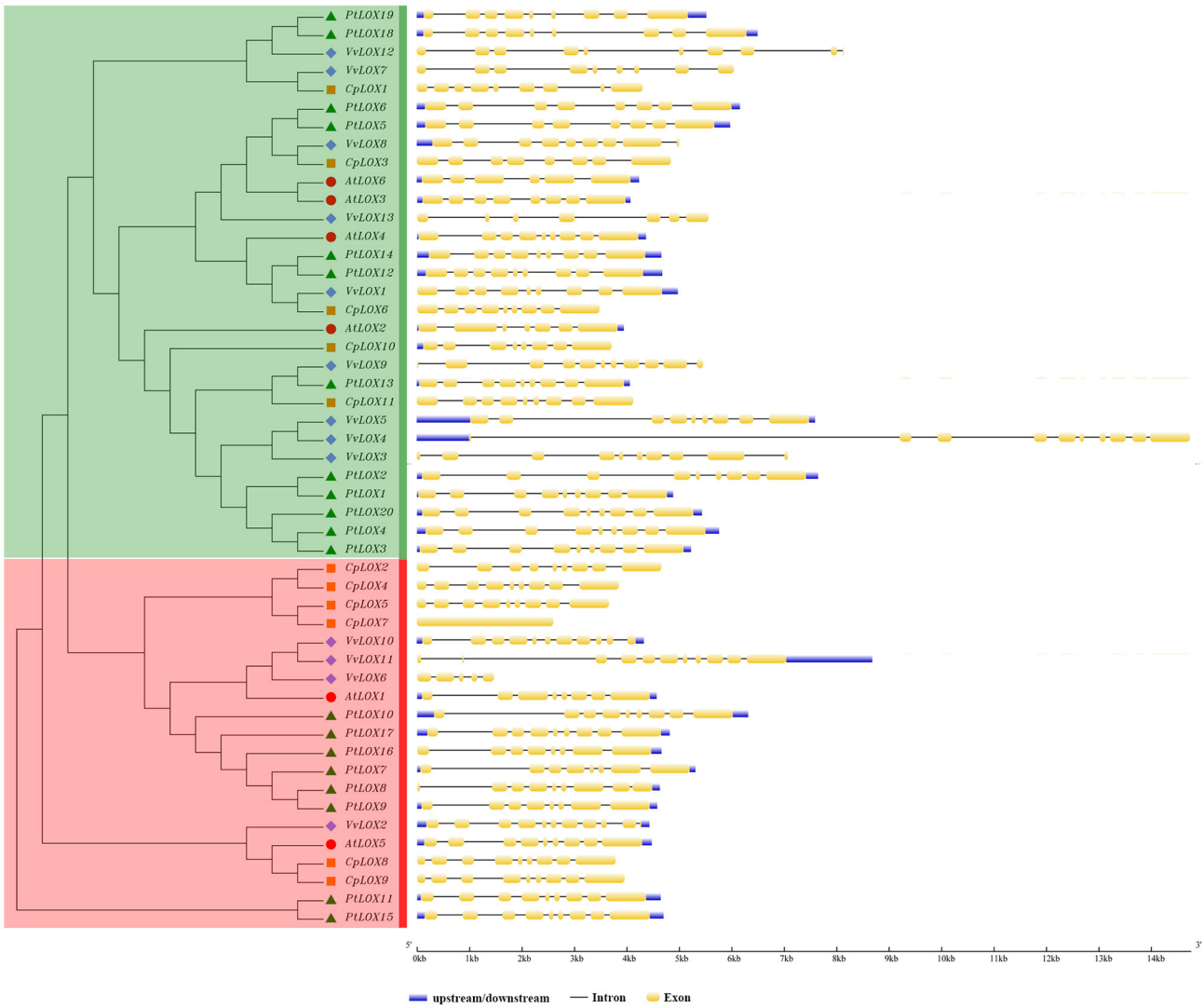


FIGURE 2 | Phylogenetic relationship of LOX proteins and the exon-intron structure of *LOX* genes from four rosid plants. Left: an unrooted phylogenetic tree constructed using MEGA 6.0 by the ME method. Different sub-families of *LOX* genes are highlighted with different colored backgrounds. Right: exon-intron structure. The exons and introns are indicated by yellow rectangles and thin lines, respectively. The untranslated regions (UTRs) are indicated by thick purple lines.

pair in the *LOX* genes are shown in **Figure 5** and **Table 2**. In *Populus*, the median Ks value of the γ triplication event is 1.54, and that related to the P-WGD is 0.27 (Tang et al., 2008b). We detected eight conserved gene pairs, which most likely resulted from SCB events. The median Ks of five in eight shows one range: 0.27–0.5. The median Ks values of the rest of the gene pairs is 2.1 and this pair is considered to associate with the most ancient γ-triplication event. In *Arabidopsis*, the median Ks values that have a relationship with β- and γ-WGDs are almost indistinguishable, and the Ks value is 2.00 (Tang et al., 2008b). To our knowledge the overall median Ks value for α-duplication in *Arabidopsis* is nearly 0.86. Therefore, the only one duplicated gene pair in *Arabidopsis* should be related with the α-duplication event. We also examined the expansion of *LOX* genes within the genomes of *Vitis*. According to previous reports, the overall median Ks value of SCBs in *Vitis* associated with the γ triplication is 1.22. In our research results, the synonymous silent substitutions per site are calculated over these three possible gene pairs. The Ks values for *VvLOX2/VvLOX10*, *VvLOX2/VvLOX8,* and *VvLOX3/VvLOX9* are 1.2, 0.83, and 1.3, respectively. Based on the predicted Ks value, *VvLOX2/VvLOX10* and *VvLOX3/VvLOX9* appear to evolve from the γ triplication, while *VvLOX2/VvLOX8* evolve from a duplication event that occurred more recently.

Based on the gene-collinearity analysis within each species, we established an idealized gene tree of the duplication groups of *LOX* genes in four rosid plants. As shown in **Figure 6**, in the *Arabidopsis LOX* genes duplicated network, one ancestor in the ancient genome duplication should have produced at least 12 *AtLOX* genes, but actually there is only one gene pair that is considered from α-duplication in our study. So we think that a possible ancient gene loss event occurred. In

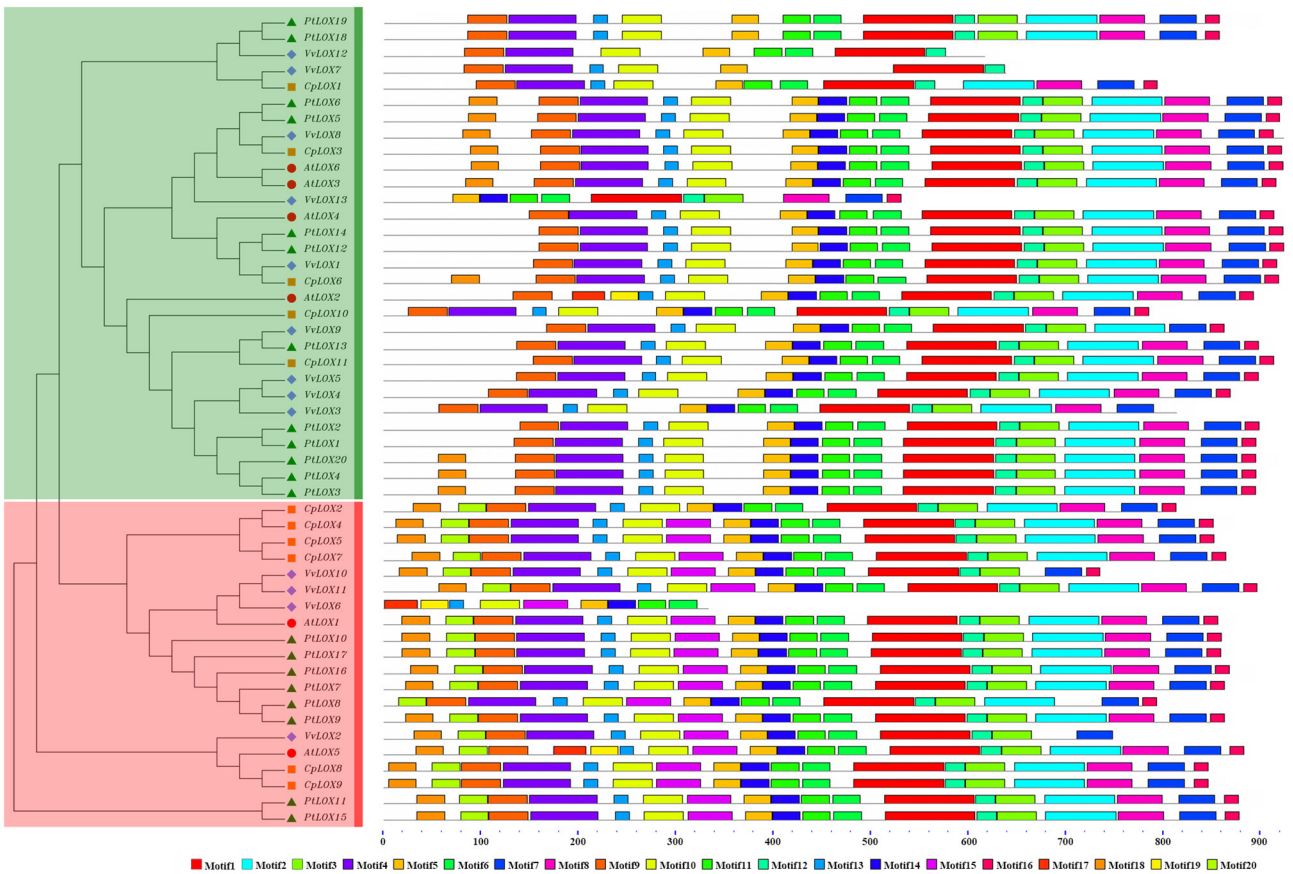


FIGURE 3 | Distribution of conserved motifs in the LOX family members. All motifs were identified by MEME using the complete amino acid sequences of 50 LOX proteins of four rosid plants. Sub-families of *LOX* genes are highlighted with different colored backgrounds, red for 9-LOX sub-family and green for 13-LOX sub-family. The different colored boxes represent different motifs and their position in each LOX sequence. For details of the motifs see Supplementary Table 6.

Populus, after two rounds of duplications, one ancestor in ancient genome duplication should have produced at least six *LOX* genes. However, two of these lines lacked the copies, which would have been obtained from p genome duplication. Moreover, *PtLOX7, PtLOX8, PtLOX11, PtLOX15* and *PtLOX16* originate from the same ancestral gene. *PtLOX1* and *PtLOXN1* evolve from a prior duplication event, while *PtLOXN1* and *PtLOX13* result from a duplication event that occurred more recently. In addition, two *LOX* gene pairs could be matched to the γ triplication in *Vitis*. In contrast, there is no *LOX*-containing segments in *Carica* being matched in any duplicated pairs. Such a huge difference existed in the expansion manners of *LOX* gene within the four rosid plants, so where did the remaining *LOX* genes in these species originate from?

Evolutionary History of *LOX* Gene Families in Four Rosid Plants

We used the *LOX* gene family members as anchor genes to further examine the orthologous relationships and evolutionary history of *LOX* genes among four rosid plants. After this interspecies microsynteny analysis, the relationships between syntenic orthologs of *LOX* genes in four rosid plants are displayed in **Figure 7**, indicating that the strongly conserved microsynteny among these regions across four species is observed significantly.

We obtained the collinear correlations of *LOX* genes in the four plant genomes by using MicroSyn. In total, 33 conserved syntenic segments were found (Supplementary Figure 2), and these syntenic segments are divided into six groups. Four of the groups contains all the four species with the *LOX* gene. *AtLOX1* in *Arabidopsis*, *PtLOX7/8/16/17* in *Populus*, *VvLOX11* in *Vitis,* and *CpLOX7* in *Carica* have conserved collinearity, and are identified as group "A". *AtLOX3/6* in *Arabidopsis*, *PtLOX5/6* in *Populus*, *VvLOX8* in *Vitis,* and *CpLOX3* in *Carica* are classified into the group "B." *AtLOX4* in *Arabidopsis*, *PtLOX12/14* in *Populus*, *VvLOX1* in *Vitis* and *CpLOX6* in *Carica* are grouped as group "C." *AtLOX5* in *Arabidopsis*, *PtLOX11/15* in *Populus*, *VvLOX2* in *Vitis* and *CpLOX8/9* in *Carica* are grouped as group "D." Group "E" consists of *LOX* genes from three species, which are *PtLOX9* in *Populus*, *VvLOX12* in *Vitis,* and *CpLOX1* in *Carica*. The *LOX* genes from two species comprise the group "F," as they are *PtLOX1* in *Populus* and *VvLOX3/4/5* in *Vitis*. The results are consistent with the findings of the phylogenetic analysis.

Subsequently, the synteny quality was calculated in four rosid plants. The quality was calculated as twice the number of

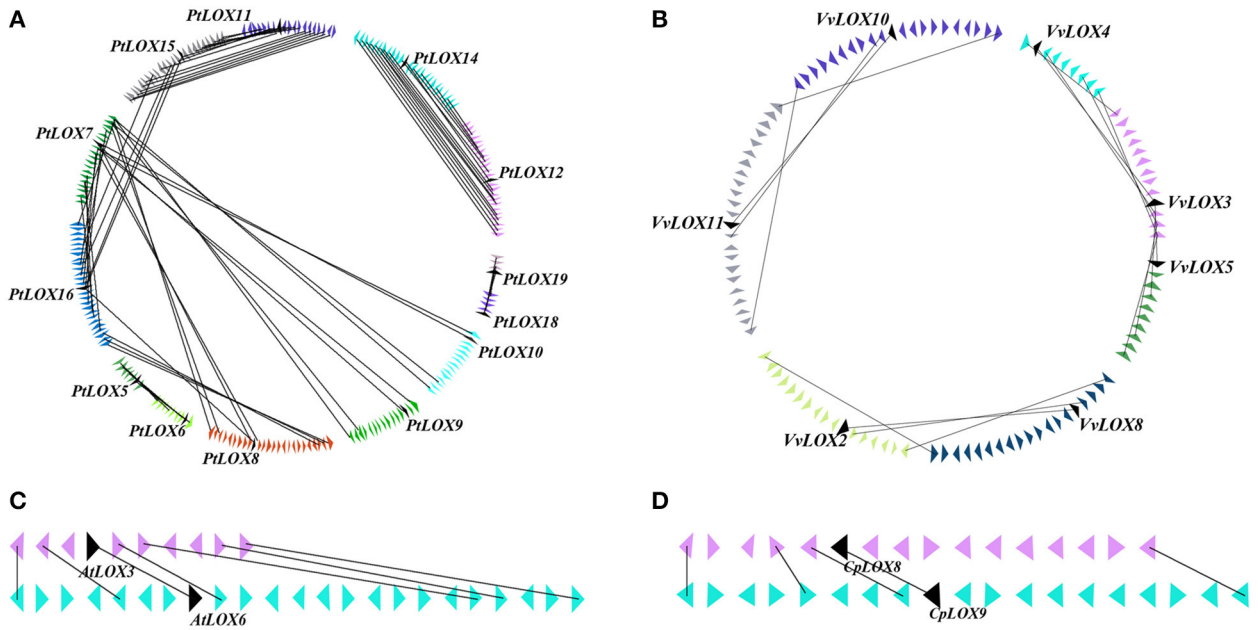


FIGURE 4 | Microsynteny related to LOX families in (A) *Populus* (B) *Vitis* (C) *Arabidopsis* (D) *Carica*. The genomic fragment is represented by a series of triangles that represent a gene in a family and its flanking genes. The genes in the same fragment show the same color except the gene in a family which is shaded by black triangle. The triangle also indicates the gene's orientation on strands. The homologous genes on two fragments are connected by a gray line.

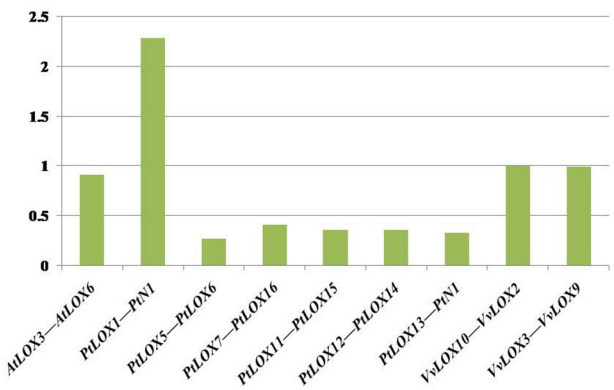


FIGURE 5 | Median Ks values of syntenic chromosomal blocks pairs associated with the expansion of LOX within the genomes of each species. The histogram shows synonymous distances (Ks) between paralogous genes (y-axis) vs. the duplicate gene pairs through segmental or whole-genome duplications (x-axis).

TABLE 2 | Median Ks values of SCB pairs associated with the expansion of LOXs within the genomes of each species.

Species	Locus_1 gene code	Locus_2 gene code	Ka	Ks	Block median Ks
Arabidopsis thaliana	*AtLOX3*	*AtLOX6*	0.10	0.98	0.91
Populus trichocarpa	*PtLOX1*	*PtN1*	0.21	2.14	2.28
	PtLOX5	*PtLOX6*	0.07	0.32	0.27
	PtLOX7	*PtLOX16*	0.05	0.26	0.41
	PtLOX11	*PtLOX15*	0.04	0.21	0.36
	PtLOX12	*PtLOX14*	0.06	0.25	0.36
	PtLOX13	*PtN1*	0.13	0.48	0.33
Vitis vinifera	*VvLOX10*	*VvLOX2*	0.29	1.21	1.00
	VvLOX3	*VvLOX9*	0.30	1.30	0.99

matches divided by the total number of genes in both segments (Cannon et al., 2006). These four species have a synteny quality of 67.77% for orthologous regions. The minimum value of synteny quality observed between *Arabidopsis* and *Vitis* was 48.97%, and the maximum value was 97.70%. The average synteny quality in the *Carica*/*Populus* syntenic regions reached over 89.24%, followed by *Carica*/*Vitis*, for which the average synteny quality was 76%. The average synteny quality in the *Arabidopsis*/*Populus* and *Arabidopsis*/*Carica* syntenic regions was 53.34 and 45.35%, respectively. Details of this comparative analysis are shown in **Table 3**.

It is thought that LOX families evolved from a process of different duplication events. However, within those LOX homologs, what kind of role is the genome-wide duplication playing? Since previous studies, the expansion of *LOX* genes within the genome of each species have been researched by using the PGDD database. Similarly, from the database, we also examined the SCBs associated with the expansion of LOXs between species. To better understand the gene-collinearity between species, a panoramic picture about the differential retention and evolution of the ancestral LOXs related to paleopolyploidy in the four rosid plants was built (**Figure 8**, **Table 2**). The study, building on previous research, has identified

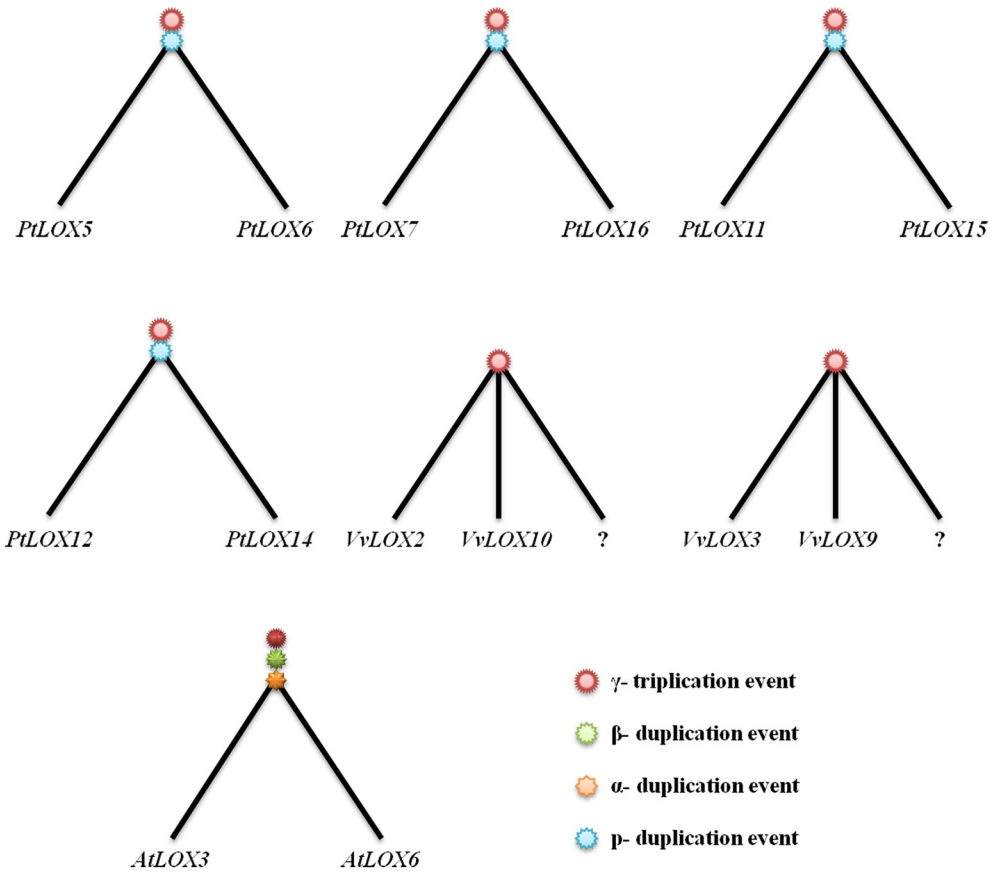


FIGURE 6 | Idealized gene trees of the duplication groups of LOX genes in *Populus, Vitis*, and *Arabidopsis*. Each tree represents a duplication group from large-scale gene duplication. As shown in the trees, the question marks indicate possible gene loss events. As shown in the trees, three paleopolyploidies affecting *Arabidopsis* (α, β, and γ duplication event). *Populus trichocarpa* has two duplication events (p and γ duplication event) and γ, which is shared by *Vitis*. The question marks indicate possible gene loss events.

differences in the duplicates of these ancestral genes through S/WGDs in each species. Five, one, eight and four *LOX* genes have been linked to paleopolyploidy in *Vitis*, *Carica*, *Populus,* and *Arabidopsis*, respectively.

In theory, it should always be possible to find out if the microsynteny were maintained among the members of four rosid plants. But this was not what we found. As shown in **Figure 8**, in the process of analyzing the gene-collinearity in individual species, we found out that one gene pair *AtLOX3/AtLOX6* in *Arabidopsis* originated from α-WGD. When we expanded to analyze the gene-collinearity between species, we found that *AtLOX3* and *AtLOX6* were all orthologous SCBs of the chromosomal block containing *PtLOX5* and *PtLOX6* in *Populus*. In addition, in *Carica*, one gene, *CpN1* on SCBs, was also found to have collinearity with *AtLOX3/AtLOX6* and *PtLOX5/PtLOX6*. This may be because *LOX* gene sub-functionalized into other gene family members over evolutionary time. A similar dynamic can be seen in the gene-collinearity analysis between *PtLOX12* and *PtLOX14*.The SCB analysis revealed that this gene pair is related to *AtLOX4* and *CpN2*. In addition, based on the results of the present study, *VvLOX2*, *VvLOX10*, *PtLOX11*, *PtLOX15,* and *AtLOX5* genes might have originated from the same ancestral gene. The orthologous *CpLOX6* genes that might have arose through S/WGDs were sub-functionalized into *VvN1* and *PtN2* in *Vitis* and *Populus*, respectively. Beyond that, the orthologous pair *CpN3* and *VvLOX12* were found to have originated from the same ancestral gene. Besides, the orthologous relationship of *VvLOX3/9* and *PtLOX1/13/N1* are no longer traceable to the *LOX* genes in the other three species. Furthermore, our opinion is that *LOX* genes which are unique to their species might represent the oldest relics of ancient LOXs differentially retained in each species.

DISCUSSION

In terms of evolution, *A. thaliana*, *P. trichocarpa*, *C. papaya*, and *V. vinifera* are believed to originate from a common paleohexaploid ancestor demonstrated by numerous studies (Ohno, 2013). Single and more recent multiple WGD events have been found in the genomes of *Populus* and *Arabidopsis*. Paleopolyploidy events provided opportunities for gene

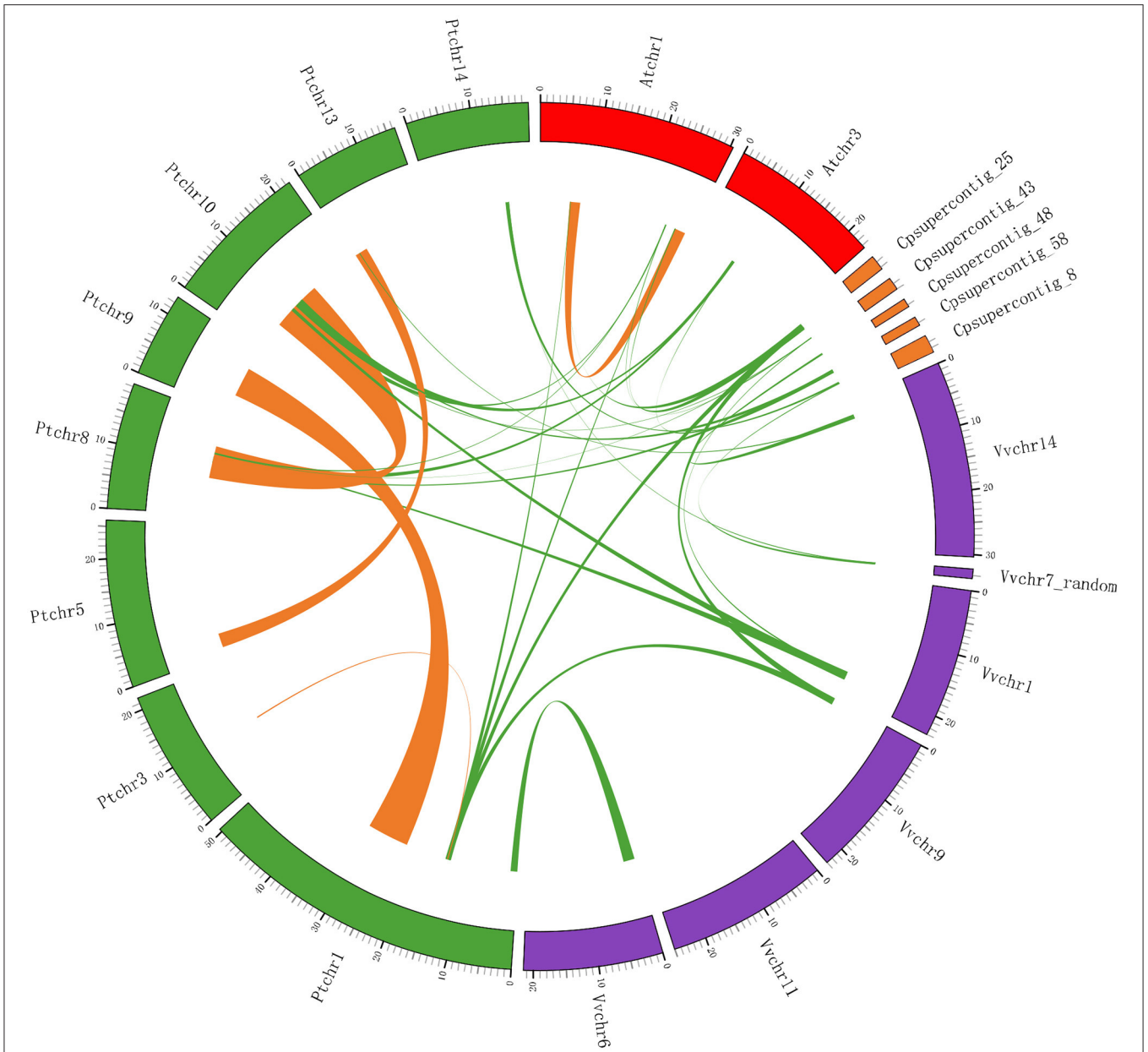


FIGURE 7 | Extensive microsynteny of LOX regions across *Populus*, *Vitis*, *Arabidopsis* and *Carica* chromosomes. The *Populus*, *Vitis*, *Arabidopsis,* and *Carica* chromosomes, shown in different colors, are labeled Pt, Vv, At, and Cp, respectively. Numbers along each chromosome box indicate sequence lengths in megabases. The whole chromosomes of these four rosid plants, harboring LOX regions, are shown in a circle. Green lines represent the syntenic relationships between LOX regions within species. Orange lines represent the syntenic relationships between LOX regions between species.

duplication, and those duplicated genes have been shown to act as an important role in evolutionary innovation (Hittinger and Carroll, 2007). Functional diversification with duplicated genes results in more complex organisms. Typically, ancient genome duplications have always been thought to be a powerful source of functional innovation and genome complexity, and is also followed by substantial gene loss. Lipoxygenase and its products are involved in the regulation of a variety of processes. In this paper, we used the model rosid plant *Arabidopsis*, as well as *Carica*, *Populus,* and *Vitis* to study the evolutionary history of this gene family.

In this study, we identified 6, 20, 13, and 11 LOXs in *Arabidopsis*, *Populus*, *Vitis,* and *Carica*, respectively. Except *Arabidopsis*, previous surveys indicated that a very high proportion of most LOX members in other three species are paralogs. In order to improve our understanding on what affects the evolutionary constraints, we measured the Ka/Ks ratios of paralogous pairs of the four species. Amidst all of the pairwise

TABLE 3 | The synteny quality of regions orthologous across four modern rosids.

	Arabidopsis	*Carica*	*Populus*	*Vitis*
ARABIDOPSIS				
Carica	45.35%			
Populus	53.34%	89.24%		
Vitis	45%	76%	97.70%	

comparison data, only one gene pair, *VvLOX6/11* exhibits a Ka/Ks ratio larger than 1, suggesting that accelerated devolution with positive selection occurred in this gene pair. Other than that, Ka/Ks ratios of all the other paralog pairs are lower than 1, indicating that the *LOX* genes at the protein level are very slow-changing and the majority of sites are often controlled by strong purifying selection.

Phylogenetic trees are quite informative for obtaining the *LOX* gene relationships with each other. In this study, the *LOX* genes are divided into two groups, 13-LOX and 9-LOX, consistent with many previous studies. The calculated genetic distances among the two LOX subfamilies were computed and the results show that the *LOX* genes of 9-LOX sub-families appear to be more closely related to each other than those *LOX* genes in 13-LOX sub-families. The number of exons in *Carica* and *Populus LOX* genes is relatively stable, whereas the exon numbers has changed dramatically in *Vitis* and *Arabidopsis*. *Vitis* have the most number of exons with 11 and the least number of exons with five. Exon-intron structural diversification has been confirmed in the evolution of many gene families, and the reason why exon-intron gain or loss occurs is because of the genetic assortment of different chromosome fragments. The MEME server identifies that each sub-family shares a similar motif, and the results could have implications for functional similarities about these LOX proteins (Paterson et al., 2006). The differential motifs in each sub-family may endow the LOX proteins with new functions or to raise their performance. Our study shows that the results meet the similarities in gene structure and motif composition of most LOX proteins from phylogenetic analysis of the *LOX* gene family. *LOX* genes differentiate into various characteristics among the different sub-families for a variety of possible reasons, and the most probable cause is that the LOX members were functionally diversified (Blanc and Wolfe, 2004).

The current tools related to investigate the relationship among genes in modern plants include sequence similarity, microsynteny analysis, and retrieve the syntenic chromosomal blocks from PGDD. As might be expected each of these approaches has advantages and shortcoming in certain situations. For example, some ancient duplicates could not be retrieved preferences from the Blast method because sequence similarity may have severely eroded in long evolutionary process. As a result, we could no longer track the paralog-ship for such duplicates based on this method. For instance, our results showed that *VvLOX3* and *VvLOX9* are duplicates resulting from γ-WGD in *Vitis* (**Figure 8**), but this gene pairs in the paralog analysis based on sequence similarity that remained undetected (Supplementary Table 5). Thanks to the plant genome duplication database, more duplicates which had arisen from segmental or whole-genome duplications could be discovered easily. But this approach had a number of drawbacks, such as we could not detect the homology pairs that proliferate via other duplication strategies. The large paralogous group unique to only one rosid plant were found abundant in our results (Supplementary Table 5). It is unscientific to infer evolutionary relationships for homologous genes among different lineages reducing only based on Microsynteny analysis. Microsynteny between two members of a gene family is calculated from their flanking genes. And we have already known that the quality of gene prediction in different genome sequencing programs is drastically different. If the flanking regions contain assembly errors, gaps or annotation errors, would cause the microsynteny that be artificial. In conclusion, to approach the evolutionary relationships for members of *LOX* gene family across four modern rosids, the above methods should be utilized compositely.

Based on the gene collinearity on syntenic chromosomal blocks within and between these four rosid plants, we set up a panoramic picture to trace the evolutionary history of *LOX* gene families (**Figure 8**). Our analysis shows that five (line 1–5 in **Figure 8**) of these ancestral LOXs are retained in more than two species. By contrast, two of the ancestral LOX genes were retained in only one of the four rosids. These observations suggests that, no matter which rosid plant is used as the model plant, the functions of a gene family inevitably get the limited amount. We could speculate genes uniquely retained in only one species may have a specific and indispensable function. As it turns out, this finding will help us dig deeper into the unique genes retained in the rosid plants and further research the function of these genes. A surprising finding of this study is that we have not found any line containing the LOX genes in all four modern rosids. Thus, we determined that all of the LOXs in each modern rosid are offspring coming from different ancestral genes.

To date, top-down analysis shows a high degree of collinearity between the four studied rosids. *Arabidopsis* (Lamesch et al., 2012), *Carica* (Ming et al., 2008), *Vitis* (Jaillon et al., 2007), and *Populus* (Tuskan et al., 2006) have been suggested to possess paleohexaploidy in a common ancestor (Jaillon et al., 2007). Previous results indicate that genome triplication (γ) occurred in a common ancestor of *Vitis*, *Arabidopsis*, *Carica,* and *Populus*. Meanwhile, the previous results show the two most recent paleopolyploidies affecting *Arabidopsis* that are often described as α and β duplication. *Populus trichocarpa* has had a unique duplication event in recent times, which is called salicoid lineage (p, following the usage in Tang et al., 2008a). Considering the paleopolyploidy events that occurred in each species, there should be 3 ancestral loci in *Carica* and *Vitis*, 6 ancestral loci in *Populus*, and 12 in *Arabidopsis*. Based on these results, the multiplicity ratio for an ancestral gene comparison in the genomes of four species should be 4:2:1:1. And in fact none of the ancestral LOXs included the extremes of each condition. However, collinear correlations of LOX genes in the four plant genomes have been obtained by using MicroSyn provides an interesting point. In our work, 33

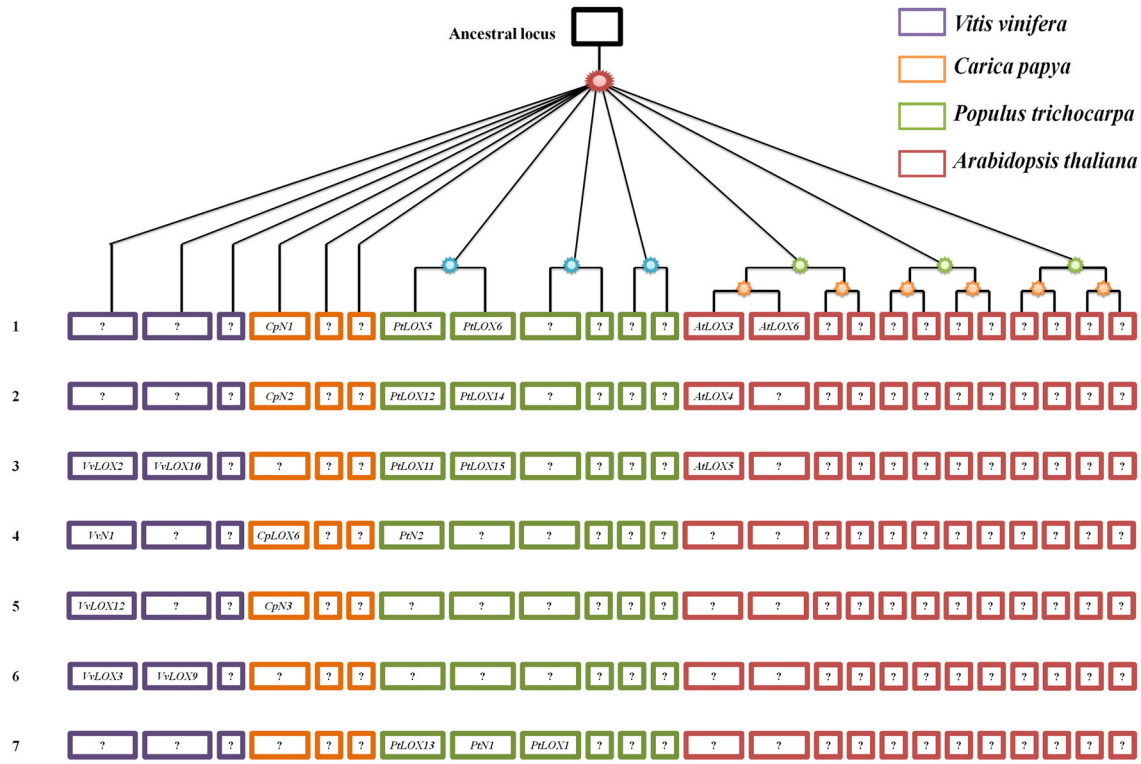


FIGURE 8 | Panoramic picture to visualize the differential retention and expansion of the ancestral *LOX*s associated with paleopolyploidy events that have occurred in four modern rosids. Square represents a SCB duplicated through paleopolyploidy events within and between species. Codes in the square correspond to associated *LOX* genes. Genes in the same line are thought to have originated from the same ancestral gene. Genes coded with an "N" between the letter and the number (e.g., CN1) represent those that have sub-functionalized into non-LOX; blank positions correspond to situations where the whole SCBs has been completely lost.

conserved syntenic segments are divided into six groups, and in almost every group, LOX genes from *Populus* are present in at least an extra copy compared with *Vitis* and *Carica*, and the two copies are paralogs. The result accords with the well-documented fact and also provides powerful evidence that *Populus* has undergone an additional whole genome duplication, which is not shared with *Vitis* and *Carica*. However, our survey results show that there are not twice as many LOX genes in *Populus* vs. *Vitis*, suggesting them could have suffered differential gene loss events could have happened to these two species. The *LOX* gene family has shrunk in the herbaceous plants and retained a large number of LOX genes in the woody plants, leading to the hypothesized that some *Arabidopsis* LOX genes might have been lost during the evolutionary process due to functional redundancy. Previous studies showed that after the paleopolyploidy events, the exponential growth in gene numbers is often tempered by massive and progressive gene death in the subsequent diploidization process (Tang et al., 2008a). Another possibility is that LOXs may have expanded faster in the other three species than in *Arabidopsis*. This expansion to more abundant LOX genes in *Populus*, *Vitis,* and *Carica* genomes suggests a great need of LOX genes to participate in more complicated physiological and biological processes in these three woody species. These results probably suggest the complex evolutionary history of the LOX family in rosid plants.

In this study, there are eight pairs of genes associated with S/WGDs, including six pairs from the PtLOX gene family, along with two pairs that have sub-functionalized into other gene family members. In contrast, there is only one collinear gene pair in both *Arabidopsis*, which is mainly caused by the rapid substitutions in *Arabidopsis*. The γ-triplication event has ever happened in the common ancestor of these four species. But all of them have different values median Ks about γ-paleologs from four modern rosids. In *Arabidopsis* the median Ks is close to 2.0, which was higher than that in *Populus* (1.54), *Carica* (1.76), and *Vitis* (1.22) (Tang et al., 2008a). Some studies have shown rapid substitutions at a rate proportional to the amount of synonymous sites (Guo et al., 2014). Over millions of years of evolution in *Arabidopsis* extensive chromosome have been actively rearranged. That might contribute to the high median Ks between γ- paleologs in *Arabidopsis* and this would destroy collinearity. The result accords with the fact that *Arabidopsis* which contains more paleopolyploidies has a smaller genome than that of *Populus*, though both originated from a common ancestor. Besides, our analyses showed that almost all *Vitis* contains many more paralog-ship LOX genes than *Carica*, although both of them were affected by the γ-WGD event. The

observations suggesting that the gene duplication impacts turn out to be small in CpLOX gene family. The dates of the large-scale duplication events have been obtained through calculating the median Ks value of duplicated genes in SCBs. In *Populus*, the number of LOX genes produced by the recent duplication event is much more than those produced from ancient duplication events. In *Arabidopsis*, the only unique gene pair is generated by α duplication event, which is also a recent duplication event. This illustrates that recent duplication host to those species which have undergone at least whole-genome duplication.

The current study provides an overview of *LOX* genes in four rosid plants, including their phylogenetic relationship, gene structure, conserved motifs, microsynteny and gene collinearity. Based on these findings, we tracked the evolutionary history of ancestral *LOX* genes among four modern rosids. The results suggest that all of the *LOX* genes in each species could have resulted from different ancestral genes. This study presented here may provide clues for exploring the unique genes retained in the rosid plants and aid in the research of the biological functions about these special genes.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: ZC, DC, and YX. Performed the experiments: ZC, DZ. Analyzed the data: ZC, WC. Wrote the paper: ZC, WC, and HY. Participated in the design of this study and revised manuscript: ZC, DC, WC.

FUNDING

Sub-project I under National Science and Technology Support Program (2015BAD07B070104). Anhui provincial Natural Science Foundation (1608085QC65). We thank the members of the Laboratory of Modern Biotechnology for their assistance in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00176

REFERENCES

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141. doi: 10.1016/j.pbi.2005.01.001

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

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi: 10.1093/nar/gkl198

Bannenberg, G., Martínez, M., Hamberg, M., and Castresana, C. (2009). Diversity of the enzymatic activity in the lipoxygenase gene family of *Arabidopsis thaliana*. *Lipids* 44, 85–95. doi: 10.1007/s11745-008-3245-7

Barker, M. S., Vogel, H., and Schranz, M. E. (2009). Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* 1, 391–399. doi: 10.1093/gbe/evp040

Baysal, T., and Demirdöven, A. (2007). Lipoxygenase in fruits and vegetables: a review. *Enzyme Microb. Technol.* 40, 491–496. doi: 10.1016/j.enzmictec.2006.11.025

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi: 10.1105/tpc.021345

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi: 10.1038/nature01521

Boyington, J. C., Gaffney, B. J., and Amzel, L. M. (1993). The three-dimensional structure of an arachidonic acid 15-lipoxygenase. *Science* 260, 1482–1486. doi: 10.1126/science.8502991

Brash, A. R. (1999). Lipoxygenases: occurrence, functions, catalysis, and acquisition of substrate. *J. Biol. Chem.* 274, 23679–23682. doi: 10.1074/jbc.274.34.23679

Brash, A. R. (2015). "Lipoxygenases: a Chronological Perspective on the Synthesis of S and R Fatty Acid Hydroperoxides," in *Bioactive Lipid Mediators*, ed T. Y. M. Murakami (Japan: Springer), 69–84.

Cai, B., Yang, X., Tuskan, G. A., and Cheng, Z.-M. (2011). MicroSyn: a user friendly tool for detection of microsynteny in a gene family. *BMC Bioinformatics* 12:1. doi: 10.1186/1471-2105-12-79

Cannon, S. B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., et al. (2006). Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14959–14964. doi: 10.1073/pnas.0603228103

Chen, Z., Chen, X., Yan, H., Li, W., Li, Y., Cai, R., et al. (2015). The Lipoxygenase gene family in Poplar: identification, classification, and expression in response to MeJA treatment. *PLoS ONE* 10:e0125526. doi: 10.1371/journal.pone.0125526

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Feussner, I., and Kühn, H. (2000). 15 Application of Lipoxygenases and related enzymes for the preparation of oxygenated lipids. *Enzymes Lipid Modification* 40:309. doi: 10.1002/3527606033.ch15

Feussner, I., and Wasternack, C. (2002). The lipoxygenase pathway. *Annu. Rev. Plant Biol.* 53, 275–297. doi: 10.1146/annurev.arplant.53.100301.135248

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., et al. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* 34, D247–D251. doi: 10.1093/nar/gkj149

Guo, A., Zhu, Q., Chen, X., and Luo, J. (2007). [GSDS: a gene structure display server]. *Yi Chuan* 29, 1023–1026. doi: 10.1360/yc-007-1023

Guo, L., Chen, Y., Ye, N., Dai, X., Yang, W., and Yin, T. (2014). Differential retention and expansion of the ancestral genes associated with the paleopolyploidies in modern rosid plants, as revealed by analysis of the extensins super-gene family. *BMC Genomics* 15:612. doi: 10.1186/1471-2164-15-612

Hildebrand, D. F. (1989). Lipoxygenases. *Physiol. Plant.* 76, 249–253. doi: 10.1111/j.1399-3054.1989.tb05641.x

Hittinger, C. T., and Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449, 677–681. doi: 10.1038/nature06151

Hufton, A. L., and Panopoulou, G. (2009). Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev.* 19, 600–606. doi: 10.1016/j.gde.2009.10.005

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Jaillon, O., Aury, J.-M., and Wincker, P. (2009). "Changing by doubling," the impact of Whole Genome Duplications in the evolution of eukaryotes. *C. R. Biol.* 332, 241–253. doi: 10.1016/j.crvi.2008.07.007

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

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210. doi: 10.1093/nar/gkr1090

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal, W., and Clustal X Version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lee, T.-H., Tang, H., Wang, X., and Paterson, A. H. (2013). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res.* 41, D1152–D1158. doi: 10.1093/nar/gks1104

Leenhardt, F., Lyan, B., Rock, E., Boussard, A., Potus, J., Chanliaud, E., et al. (2006). Genetic variability of carotenoid concentration, and lipoxygenase and peroxidase activities among cultivated wheat species and bread wheat varieties. *Eur. J. Agronomy* 25, 170–176. doi: 10.1016/j.eja.2006.04.010

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991–996. doi: 10.1038/nature06856

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2007). New developments in the InterPro database. *Nucleic Acids Res.* 35, D224–D228. doi: 10.1093/nar/gkl841

Ohno, S. (2013). *Evolution by Gene Duplication*. Berlin; Heidelberg: Springer Science & Business Media.

Paterson, A. H., Chapman, B. A., Kissinger, J. C., Bowers, J. E., Feltus, F. A., and Estill, J. C. (2006). Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends Genet.* 22, 597–602. doi: 10.1016/j.tig.2006.09.003

Podolyan, A., White, J., Jordan, B., and Winefield, C. (2010). Identification of the lipoxygenase gene family from *Vitis vinifera* and biochemical characterisation of two 13-lipoxygenases expressed in grape berries of Sauvignon Blanc. *Funct. Plant Biol.* 37, 767–784. doi: 10.1071/FP09271

Rokas, A., and Holland, P. W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* 15, 454–459. doi: 10.1016/S0169-5347(00)01967-4

Sémon, M., and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512. doi: 10.1016/j.gde.2007.09.007

Shibata, D., Steczko, J., Dixon, J., Hermodson, M., Yazdanparast, R., and Axelrod, B. (1987). Primary structure of soybean lipoxygenase-1. *J. Biol. Chem.* 262, 10080–10085.

Steczko, J., Donoho, G. P., Clemens, J. C., Dixon, J. E., and Axelrod, B. (1992). Conserved histidine residues in soybean lipoxygenase: functional consequences of their replacement. *Biochemistry* 31, 4053–4057. doi: 10.1021/bi00131a022

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008a). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954. doi: 10.1101/gr.080978.108

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Umate, P. (2014). Genome-wide analysis of lipoxygenase gene family in *Arabidopsis* and rice. *Plant Signal. Behav.* 6, 335–338. doi: 10.4161/psb.6.3.13546

Van Loon, L. C., Rep, M., and Pieterse, C. (2006). Significance of inducible defense-related proteins in infected plants. *Annu. Rev. Phytopathol.* 44, 135–162. doi: 10.1146/annurev.phyto.44.070505.143425

Vogt, J., Schiller, D., Ulrich, D., Schwab, W., and Dunemann, F. (2013). Identification of lipoxygenase (LOX) genes putatively involved in fruit flavour formation in apple (Malus× domestica). *Tree Genet. Genomes* 9, 1493–1511. doi: 10.1007/s11295-013-0653-5

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49–e49. doi: 10.1093/nar/gkr1293

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom. Bioinform.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503. doi: 10.1016/S0169-5347(00)01994-7

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 Genetics

ORIGINAL RESEARCH
published: 15 June 2017
doi: 10.3389/fgene.2017.00078

Genome-Wide Analysis Suggests the Relaxed Purifying Selection Affect the Evolution of *WOX* Genes in *Pyrus bretschneideri*, *Prunus persica*, *Prunus mume,* and *Fragaria vesca*

*Yunpeng Cao[1], Yahui Han[2], Dandan Meng[1], Guohui Li[1], Dahui Li[1], Muhammad Abdullah[2], Qing Jin[1], Yi Lin[1] and Yongping Cai[1]**

[1] School of Life Sciences, Anhui Agricultural University, Hefei, China, [2] State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China

OPEN ACCESS

Edited by:
Rosalba Giugno, University of Verona, Italy

Reviewed by:
Manoj Prasad, National Institute of Plant Genome Research, India
Rosario Muleo, Università degli Studi della Tuscia, Italy

***Correspondence:**
Yongping Cai
swkx12@ahau.edu.cn

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

Received: *10 March 2017*
Accepted: *29 May 2017*
Published: *15 June 2017*

Citation:
Cao Y, Han Y, Meng D, Li G, Li D, Abdullah M, Jin Q, Lin Y and Cai Y (2017) Genome-Wide Analysis Suggests the Relaxed Purifying Selection Affect the Evolution of WOX Genes in Pyrus bretschneideri, Prunus persica, Prunus mume, and Fragaria vesca. Front. Genet. 8:78. doi: 10.3389/fgene.2017.00078

WUSCHEL-related homeobox (WOX) family is one of the largest group of transcription factors (TFs) specifically found in plant kingdom. WOX TFs play an important role in plant development processes and evolutionary novelties. Although the roles of *WOXs* in Arabidopsis and rice have been well-studied, however, little are known about the relationships among the main clades in the molecular evolution of these genes in Rosaceae. Here, we carried out a genome-wide analysis and identified 14, 10, 10, and 9 of *WOX* genes from four Rosaceae species (*Fragaria vesca*, *Prunus persica*, *Prunus mume,* and *Pyrus bretschneideri*, respectively). According to evolutionary analysis, as well as amino acid sequences of their homodomains, these genes were divided into three clades with nine subgroups. Furthermore, due to the conserved structural patterns among these *WOX* genes, it was proposed that there should exist some highly conserved regions of microsynteny in the four Rosaceae species. Moreover, most of *WOX* gene pairs were presented with the conserved orientation among syntenic genome regions. In addition, according to substitution models analysis using PMAL software, no significant positive selection was detected, but type I functional divergence was identified among certain amino acids in WOX protein. These results revealed that the relaxed purifying selection might be the main driving force during the evolution of *WOX* genes in the tested Rosaceae species. Our result will be useful for further precise research on evolution of the *WOX* genes in family Rosaceae.

Keywords: *WOX* genes, phylogenetic analysis, microsynteny, selection, functional divergence

INTRODUCTION

The WUSCHEL-related homeobox (WOX) gene family encodes a group of plant-specific transcription factors (TFs), which belongs to the homeodomain (HD) TF superfamily (Deveaux et al., 2008; Zhang et al., 2015). There are 15 and 13 members of the *WOX* family in *Arabidopsis thaliana* and rice (*Oryza sativa*) genomes, respectively (Haecker et al., 1991; Graaff et al., 2009; Xin et al., 2010). WOX TFs have been reported to play important role in plant development,

such as regulating dynamic balance of stem cell division and differentiation, embryo development, and post-embryonic development (Palovaara et al., 2010; Yadav et al., 2010; Ueda et al., 2011). Bioinformatics analysis showed that *WOX* homology sequences were found in the genomes of *Selaginella*, *Bryophyta* and *Chlorophyta*, but not in the *Rhodophyta* genome, indicating that *WOX* family might originate from green algae (Mukherjee et al., 2009; Lian et al., 2014). According to the phylogenetic analysis among *A. thaliana* and *Petunia hybrida*, tomato (*Solanum lycopersicum*) and rice (*O. sativa*), the *WOX* family was divided into three separate clades: WUS/modern clade, the intermediate clade, and the ancient clade (Haecker et al., 1991). Research on the structural characteristics of WOX members showed that the evolutionary branch members contained a specific WUS box (T-L-X-L-F-P-X-X, where X represents an amino acid) (Haecker et al., 1991). WUS box is an essential component for WUS regulation of stem tip meristem stem cell homeostasis and floral meristem morphogenesis (Ikeda and Ohme-Takagi, 2009). Moreover, *AtWUS* and *AtWOX5* within modern/WUS clade, can redundantly maintain the apical stem cells under undifferentiated status (Sarkar et al., 2007); *AtWOX4* can influent the process of secondary growth by modulating the activity of vascular cambium (Hirakawa et al., 2010); *AtWOX1/3* can coordinate the development of paraxial and distal ends during the leaf development; the primordial initiation and development within meristem were terminated by overexpression of *AtWOX6*. Within the intermediate clade, *AtWOX9* can maintain the growth and division of meristematic cells (Wu et al., 2005); *AtWOX11* was specifically expressed in the cambium, and can promote the formation of adventitious roots (Zhao et al., 2009). Within the ancient clade, *AtWOX13* can promote the formation of embryonic placenta during fruit development (Romera-Branchat et al., 2013); *AtWOX14* and *AtWOX4* can redundantly regulate the differentiation of vascular meristem (Etchells et al., 2013). These studies suggest that the *WOX* gene family is widely involved in the regulation of plant meristem. The members of *WOX* gene family appear to be functionally diverse. Although this gene family in some model plants, such as Arabidopsis and rice, has been studied on a phylogenetic scale, a comprehensive molecular evolutionary study remains elusive in Rosaceae species. Recently, a number of researches on application of comparative genome in analysis of evolution and function of the gene family have been reported (Cao et al., 2016a). Similar to other highly conserved genes, the *WOX* gene and its flanking sequences are likely to be conserved with microsynteny, which can promote the transfer of genetic knowledge among the related species of Rosaceae. The genomes of pear (*Pyrus bretschneideri*), peach (*Prunus persica*), mei (*Prunus mume*), and strawberry (*Fragaria vesca*) were published in Shulaev et al. (2011), Zhang et al. (2012), Verde et al. (2013), and Wu et al. (2013), respectively. Therefore, the availability of whole-genome sequences for four members of three Rosaceae subfamilies (*Fragaria, Prunus,* and *Pyrus*) have enabled us to explore the selection regimes under which *WOX* genes have diversified during the radiation of the Rosaceae. The present research could lead to a better understanding of *WOX* gene family on evolutionary history and diversification in Rosaceae.

MATERIALS AND METHODS

Database Search

WOX genes were identified from the genome data representing the four Rosaceae species (*P. bretschneideri*, *P. persica*, *F. vesca,* and *P. mume*, respectively). Two different methods were used to identify *WOX* genes in the *P. bretschneideri*, *P. persica*, *F. vesca,* and *P. mume* genome: (1) BLASTP search using Arabidopsis and rice WOX protein sequences according to previous research methods (Cao et al., 2016a,d), and (2) the screening of Hidden Markov Model profile (PF00046) in four Rosaceae genome using DNAtools software with an e-value cut off of 0.001 (Gehring, 1992). All candidate WOX proteins were confirmed to have a complete WOX domain using both Pfam (Punta et al., 2011), SMART dadabases (Letunic et al., 2012) and InterProScan database (Zdobnov and Apweiler, 2001).

Phylogenetic Trees Construction

The multiple alignment of WOX proteins in five plant species (*P. bretschneideri*, *P. persica*, *F. vesca*, *P. mume,* and *A. thaliana*) was performed using CLUSTAL_X software (Thompson et al., 1997). Subsequently, we constructed NJ (neighbor-joining) tree using MEGA version 5.1 software (Tamura et al., 2011) with the following parameters: bootstrap (1000 replicate), pairwise deletion and Poisson correction. At the same time, we used ML (maximum-likelihood) and ME (Minimum-evolution) methods to generate the phylogenetic trees to validate the topologies.

Exon–Intron Structural Analysis and Identification of Conserved Motifs

The online program Gene Structure Display Server (Hu et al., 2014) was used to detect the exon–intron structure of cDNAs and genomic DNA sequences. Subsequently, the MEME (Multiple Em for Motif Elicitation, Version 4.11.1) program (Bailey et al., 2015) was used to obtain the motifs in all candidate WOX proteins, with the parameters: the maximum number of motifs at 20, and the optimum motif width between six and 200 residues. Furthermore, the Pfam database (Punta et al., 2011), SMART software (Letunic et al., 2012), and InterProScan database (Zdobnov and Apweiler, 2001) were used to annotate these structure motifs.

Microsynteny Analysis

According to the comparisons of the specific regions containing *WOX* genes, we carried out microsynteny analysis across the four Rosaceae species. Similarly, the *WOX* genes of *P. bretschneideri*, *P. persica*, *P. mume,* and *F. vesca* were categorized based on their classification in the evolutionary tree. Subsequently, all *WOX* genes in *P. bretschneideri*, *P. persica*, *F. vesca,* and *P. mume* were set as anchor sites, according to their physical location. Then the flanking protein-coding genes of the *WOX* gene in one species were compared with those in other species. The criterion for

dividing an interspecific synteny block is to locate three or more conserved homologous genes within 100 KB between genomes (BLASTP E-value $< 10^{-10}$) (Wang et al., 2012).

Selective Pressure and Functional Divergence Analysis

To future understand whether the *WOX* genes have undergone positive selection during evolution, maximum likelihood codon models (site models and branch-site models) in PAML software (Yang, 2007) were performed. Three pairs of models (M0 vs. M3, M1a vs. M2a, and M7 vs. M8) were utilized to detect positive selection sites. In the free site models, M0 (one ratio), M1a (neutral), M2a (selection), M3 (discrete), M7 (beta), and M8 (beta and ω) were evaluated by the likelihood ratio test (LRT). The LRT was used to judge which model was more suitable in the two models, and the amino acids sites with positive selection were obtained by the Bayesian method of PAML software (Yang, 2007).

Functional divergence analysis of amino acid sequence data was performed using Diverge 2.0 combined with constructed phylogenetic tree (Gu, 1999, 2006; Gu et al., 2013). The type I functional divergence led to a change of functional limitation, which was highly correlated with the evolution rate after gene duplication (Gu, 1999, 2006; Gu et al., 2013). Type II functional divergence did not result in a change in the functional limitation of the members after gene duplication, but the change of physical and chemical properties of amino acid residues (Gu, 1999, 2006; Gu et al., 2013).

cis-Acting Elements Analysis

To identify putative *cis*-elements in promoter regions of *WOX* genes, the PlantCARE database (Lescot et al., 2002) was used. 2000 bp genomic sequence upstream of the start codon (ATG) was used for *cis*-acting elements analysis.

Expression Profiles of *FvWOX* Genes

The normalized data (Fragments Per Kilobase Exon model per Million mapped fragments, FPKM) during *F. vesca* development was reported by Darwish et al. (2013), and available from SGR GBrowse. A gene was thought to be expressed if the FPKM value was greater than or equal to 0 FPKM in at least one of the 14 tissues. Subsequently, the transcriptome data of *FvWOXs* was visualized using the R software[1].

RESULTS

Identification and Chromosomal Distribution of *WOX* Genes in Rosaceae

For identification of *WOXs* gene families, the genome data of tested Rosaceae species was subjected to HMM and BLASTP searches. As a result it is revealed, presence of 9, 10, 10, and 14 *WOX* genes in *P. mume, P. bretschneideri*, *P. persica*, and *F. vesca*, respectively (Supplementary Table S1). These *WOX* genes were named according to method of Haecker et al. (1991). For this purpose, the phylogenetic tree was carried out based on multiple sequence alignments for the full-length WOX protein sequences of tested Rosaceae species and *A. thaliana*. The identified *WOX* genes of these tested four Rosaceae species were renamed according to the evolutionary relationship as shown in Supplementary Table S1. Subsequently, the distribution of these *WOX* genes on chromosomes was identified, based on genomic annotation information. As shown in **Figure 1**, it is discovered that the *WOX* genes were unevenly distributed among the chromosomes in each of tested four species. In the *P. bretschneideri* genome, four of *WOX* genes were distributed on chromosome 15, while remaining distributed on chromosomes 3, 5, 6, 12, and 17. In *F. vesca*, five and four of 14 *WOX* genes were distributed among chromosomes 3 and 5, respectively. In both *P. persica* and *P. mume*, three *WOX* genes distributed were found on one chromosome (no. 7 and no. 8, respectively), with others scattered across different chromosomes (**Figure 1**).

Evolution of *WOX* Genes in Rosaceae

To investigate the possible evolutionary history of the *WOX* genes in the tested Rosaceae species, we carried out a joint phylogenetic analysis using three methods; ME, ML, and NJ. Based on previous report that WOX13 subfamily was an ancient member in the *WOX* gene family (Deveaux et al., 2008), the WOX13 subfamily was selected as an outgroup to root phylogenetic tree. All the tree topologies generated by the three methods (ME, ML, and NJ) were largely consistent with each other, with only minor changes in internal branches (**Figure 2** and **Supplementary Figures S1**, **S2**). Therefore, only NJ phylogenetic tree was used in the following analysis. Previous studies on *WOX* genes have confirmed that motifs FYWFQNH, FYWFQNR, and YNWFQNR were representative markers for the WUS/modern clade, intermediate clade and ancient clade, respectively (Graaff et al., 2009; Ge et al., 2016). Confining the previous results (Deveaux et al., 2008; Xin et al., 2010), our evolutionary analysis exposed a total of 58 members of the *WOX* genes in *P. bretschneideri*, *P. persica*, *F. vesca*, *P. mume* along with *A. thaliana*. These 58 members of *WOX* genes were divided into three clades and nine subfamilies. The Modern clade contained a total of six subfamilies (WUS, WOX1, WOX2, WOX3, WOX4, and WOX5), and Intermediate clade included two subfamilies (WOX9 and WOX11), while Ancient clade just had a WOX13 subfamily, which were consistent with the evolutionary relationships of WOXs in other species (Xin et al., 2010; Hedman et al., 2013; Nardmann and Werr, 2013; Lian et al., 2014). Remarkably, we found that all subfamilies contained at least one *WOX* member from each of the four Rosaceae species (**Figure 2**). These results imply that rapid duplication of *WOX* genes occurred before these dicotyledonous species were diverged.

Phylogenetic analysis revealed that three pairs of paralogous genes were found among the *WOX* genes, which were consistent with the previous notion that most members of the *WOX* gene family are represented by pairs of orthologous genes. As shown in **Figure 3**, up to eight pairs of orthologous *WOX* genes were shared by *P. persica* and *P. mume*, whereas only two pairs of orthologous *WOX* genes found between *P. bretschneideri* and

[1]http://www.bioconductor.org

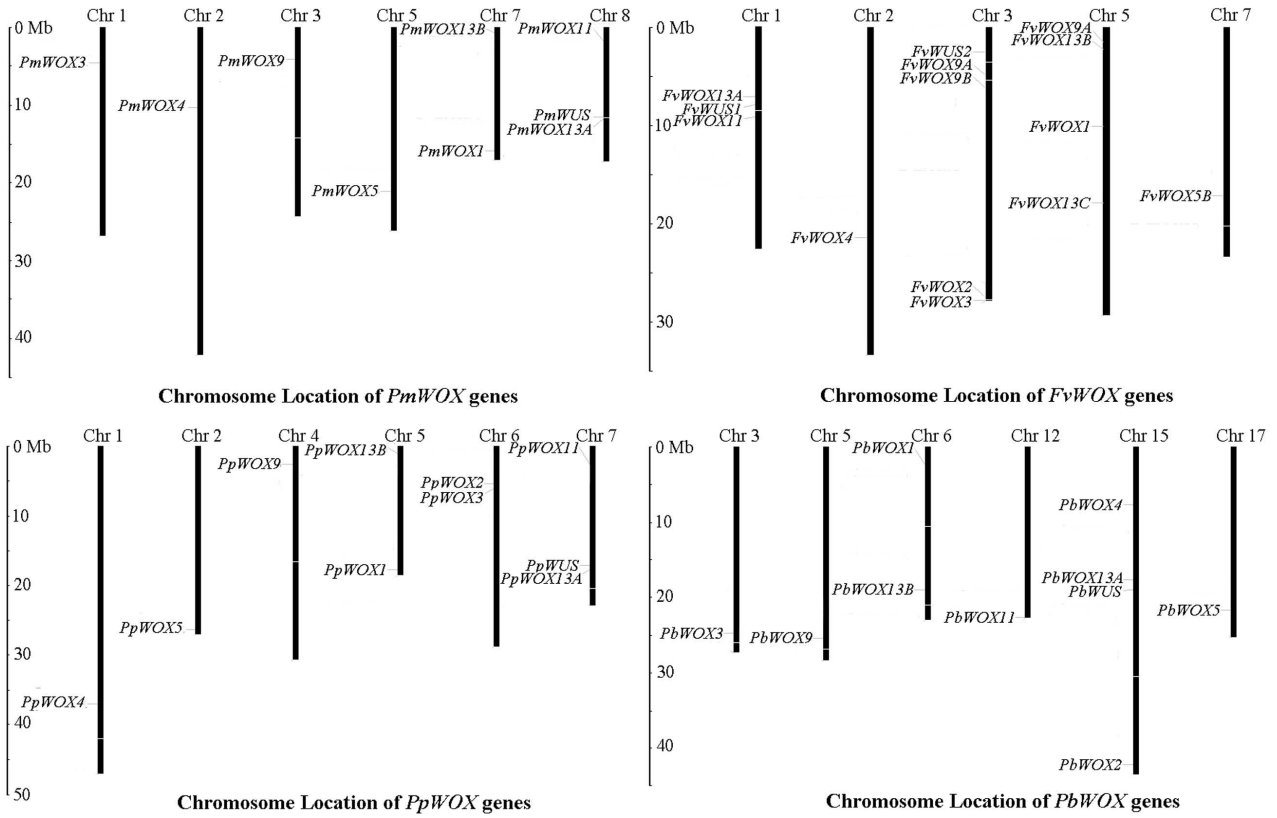


FIGURE 1 | Chromosomal location of *WOX* genes from *Pyrus bretschneideri*, *Prunus persica*, *Fragaria vesca* and *Prunus mume*. The chromosome number was represented by the top of each chromosome.

P. persica. However, no orthologous *WOX* genes were found between *F. vesca* and other species. These results were consistent with the evolutionary relationships among these four Rosaceae species (Dickinson et al., 2007; Cao et al., 2016b).

Analysis of Exon–Intron Structure and the Conserved Motifs

Previous studies have shown that gene structural diversity is an important resource for the evolution of multigene families (Liu et al., 2009; Cao et al., 2016c). To understand the structural diversity of the *WOX* genes in Rosaceae, gene structures of *PbWOXs*, *PpWOXs*, *PmWOXs,* and *FvWOXs* were deduced. It is revealed that these *WOX* genes contained different numbers of exons as shown in **Figure 4A**. For example, *FvWOX11* only contained one exon, while *FvWOX9A* contained the largest number of exons (5). Moreover, 16, 15, and 3 of *WOX* genes contained two, three and four exons, respectively. These results suggested that the functional diversity of *WOX* genes may be in consequence due to exon loss or gain during the evolution of the *WOX* gene family. Subsequently, gene structures of the *WOX* paralogous and orthologous gene pairs were further analyzed. Among these genes, we found that the exon number of seven gene pairs had changed, including *FvWOX9A*/*FvWOX9B*, *FvWOX13B*/*FvWOX13C*, *PmWOX3*/*PpWOX3*, *PbWOX4*/*PpWOX4*, *PmWUS*/*PpWUS*, *PmWOX11*/*PpWOX11,* and *PmWOX13B*/*PpWOX13B.* By comparing among these seven gene pairs, it was found that one exon was lost in *FvWOX9B*, *PmWOX3*, *PpWOX4*, *PpWUS*, *PmWOX11* and *PmWOX13B*, while one exon was obtained in *FvWOX9A*, *PpWOX3*, *PbWOX4*, *PmWOXWUS*, *PpWOX11,* and *PpWOX13B.* It may happen during the long evolutionary period. Previous studies have proposed that introns could be specifically inserted and remained in the plant genome during evolution (Rogozin et al., 2003; Carmel et al., 2007; Cao et al., 2016d). In our study, these phenomenon were observed, which might explain the functional differences and diversity of closely related *WOX* genes, such as *PbWOX3* and *FvWOX3*, *PbWOX13B,* and *FvWOX13A* (**Figure 4A**).

Furthermore, it was observed that 20 of the conserved motifs were found in the 43 WOX proteins using MEME website (Supplementary Table S2). These motifs were annotated by using Pfam and SMART. Motif 1, present in all subfamilies, was identified to encode for a conserved homodomain. In addition to the homodomain, most of the WOX members within the same clade shared the similar motif compositions as shown in **Figure 4B**. These results reinforced the classification of WOX subfamilies. However, several motifs were unique to the proteins in some clades. For example, Motif 4 was unique to Ancient clade (clade I: WOX13 subfamily), Motif 5 to WUS clade (clades A–F)

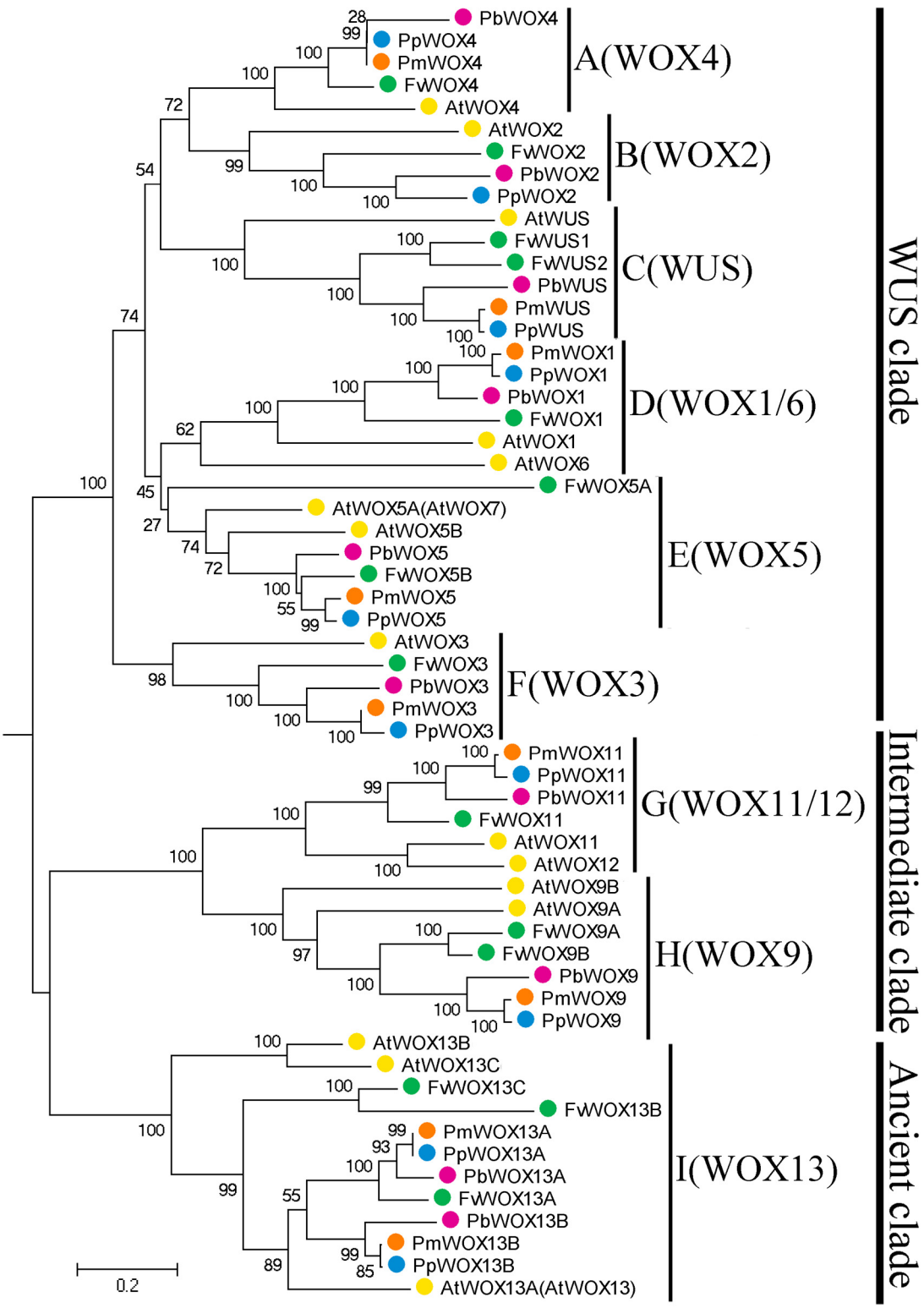


FIGURE 2 | Neighbor-Joining tree of *WOX* family members in four Rosaceae species, *F. vesca* (Fv, green), *P. mume* (Pm, orange), *P. persica* (Pp, blue), and *P. bretschneideri* (Pb, red). Numbers indicate bootstrap support for branches. The clade I *WOX* genes (and only this group) are found both in some green algae and in all land plants, and so provide a root for this tree.

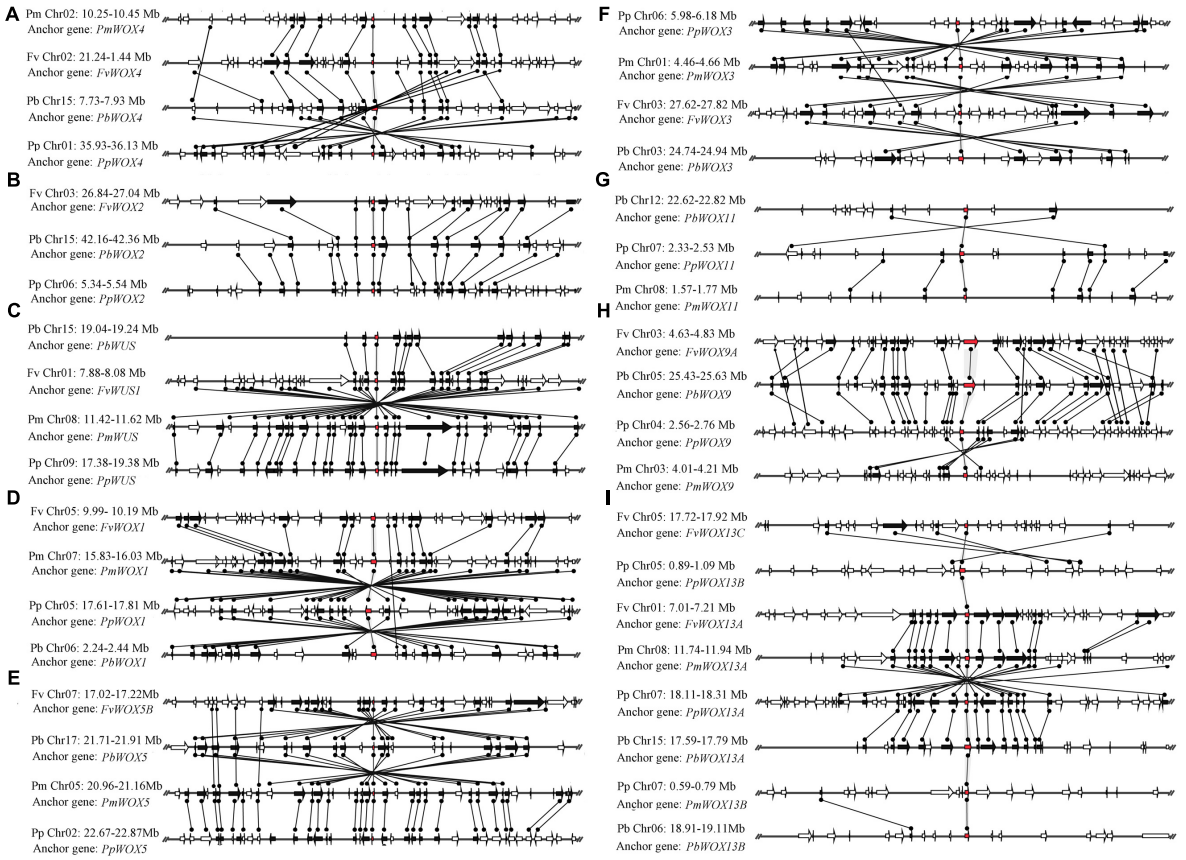


FIGURE 3 | Extensive microsynteny of *WOX* regions across *P. bretschneideri* (Pb), *P. persica* (Pp), *F. vesca* (Fv), and *P. mume* (Fm) chromosomes. The gene's orientation on strands was indicated by the triangle. Remarkably, the relative positions of all flanking protein-coding genes were defined by anchored *WOX* genes, highlighted in red. Subsequently, we used black lines to connect the homologous genes on two fragments. All genes are numbered from left to right, in order, for each segment. The **(A–I)** subfamilies in figure were consistent with those in **Figure 2**.

and Motif 15 to clade C (Supplementary Table S2 and **Figure 4B**). To some extent, these specific motifs may play an important role in the clade or subfamily, as well as contribution to the functional divergence of *WOX* genes.

Sequence Analysis of WOX Domains

Based on their amino acid sequences, the newly identified *WOX* gene family members were found to contain the conserved homeodomain by multiple sequence alignment of Cluxa2.0 with default parameters. The conserved homeodomain was selected for the visualized results by ESPript 3 (Gouet et al., 2005). The homeodomain structures of these four species were highly similar with each other. They contained a helix-loop-helix-turn-helix structure with either 65 or 66 amino acid residues. A total of 11 conserved sites (Q, L and Y in helix1; I, V, W, F, N, K, and R in helix3) of homeodomain reported previously (Gehring, 1992; Xin et al., 2010), were also conservative in the *WOX* proteins of the Rosaceae species (**Figure 5**). These findings suggested that these amino acid residues could play an important role in their functions. In addition to the previously reported conserved amino acid sites, other conserved amino acid sites have been identified in this study, such as P, L, and I in helix 2, Q and F in helix 3, as well as G in the turn region. Interestingly, an extra Y residue was found in the homeodomain of PbWUS, PmWUS, PpWUS, FvWUS and AtWUS, compared with other members of *WOX* gene family in *P. bretschneideri*, *P. persica*, *P. mume* and *F. vesca* and *A. thaliana*. Similar finding have been reported by Mayer et al. (1998) and Xin et al. (2010), that the homeodomains of *A. thaliana* WUS, *O. sativa* WUS, *Z. mays* WUS1, *Z. mays* WUS2, *S. bicolor* WUS and *P. trichocarpa* WUS were composed of 66 amino acid residues containing an extra Y residue by multiple sequence alignment, which indicates that this residue might play an important role on the function of WUS TF. Remarkably, in Arabidopsis, AtWOX5 (without Y residue between Helix1 and Loop) could replace AtWUS (containing Y residue) to maintain the dynamic balance of stem cells in the shoot apical meristem (Sarkar et al., 2007)

It was reported previously that the WUS protein contains three functional domains, including WUS-box, acidic region and EAR-like motif (Xin et al., 2010; Xiaoxu et al., 2016). These functional domains significantly contribute to its function as a TF (Xin et al., 2010; Xiaoxu et al., 2016). In present study, Motif 5 (WUS box: amino acids, TLLFP) was observed to be in

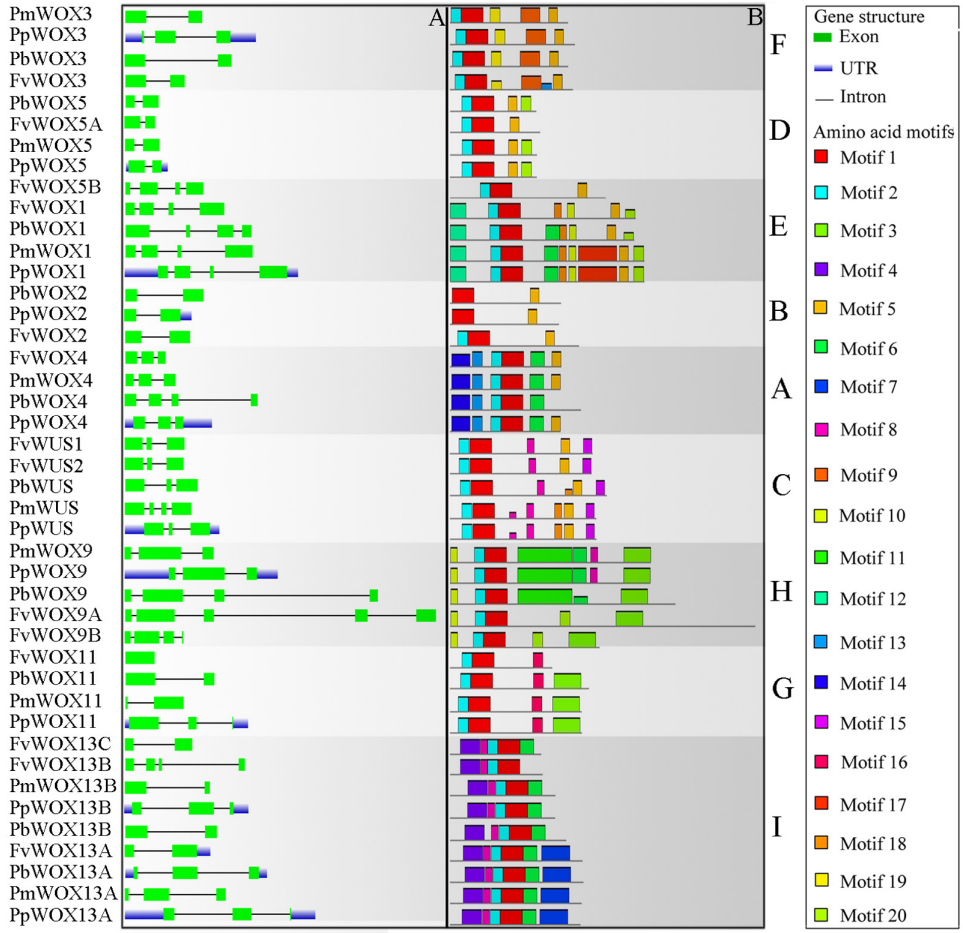


FIGURE 4 | Exon/intron structure **(A)** and motif compositions **(B)** of the Rosaceae *WOX* genes. Relative protein or gene lengths can be estimated by gray bars. Untranslated regions (UTRs), exons and introns are represented by blue lines, green boxes and gray lines, respectively. Motif sequences were shown in Supplementary Table S2. The A-I subfamilies in figure were consistent with those in **Figure 2**.

the presence of all WOX proteins in WUX clade (clades A–F) (Supplementary Table S2 and **Figure 4B**). In clade C, the Motif 15 (EAR-like motif: amino acids, SLELSL) was found in all WOX proteins. However, no acidic region was identified in all WOX proteins (Supplementary Table S2 and **Figure 4B**). These results were consistent with previous findings that acidic region may be an important function domain only in Arabidopsis *WUS* gene (Xin et al., 2010; Xiaoxu et al., 2016).

Microsynteny Analysis of *WOX* Genes

Microsynteny has been surveyed in different species to understand the position of the homologous genes (orthology or paralogy) (Cannon et al., 2003; Yan et al., 2004; Cao et al., 2016a). In this study, microsynteny analysis was carried out for identification of homologous relationships within the *WOX* genes in *P. bretschneideri*, *P. persica*, *F. vesca,* and *P. mume* (**Figure 3**). Additionally, to measure the linkages and molecular history among *WOX* genes, a stepwise gene-by-gene reciprocal comparison was performed. In general, if the flanking genes in the chromosome region of the target gene contained three or more pairs of genes that are collinear, they could be considered as the conserved microsynteny (Lin et al., 2014; Cao et al., 2016a).

Primarily, the intraspecies microsynteny was investigated among four Rosaceae species. However, it was revealed that no collinear *WOX* genes were observed (**Figure 3**). These results suggested that independent duplication events were the main expansion pattern of *WOX* gene family members. Consequently, we analyzed the relationship of the *WOX* genes within each interspecies. The results exposed that the nine clades containing 38 *WOX* genes were found, among which 10 were from *P. persica* and *F. vesca*, 9 from *P. bretschneideri* and *P. mume*, respectively. Then several higher levels of microsynteny found in subfamilies A–G. Among these microsynteny some were remarkably inverted, duplicated such as *PpWOX3*/*PmWOX3*, *PbWOX1*/*PpWOX1*, and *FvWOX13C*/*PpWOX13B* (**Figure 3**). Usually, genome segments in the same group may evolve from a single sequence, which led to species differentiation (Tripoli et al., 2005; Jing et al., 2016). However, sequence fragments from the same group are considered to be homologous genes, and

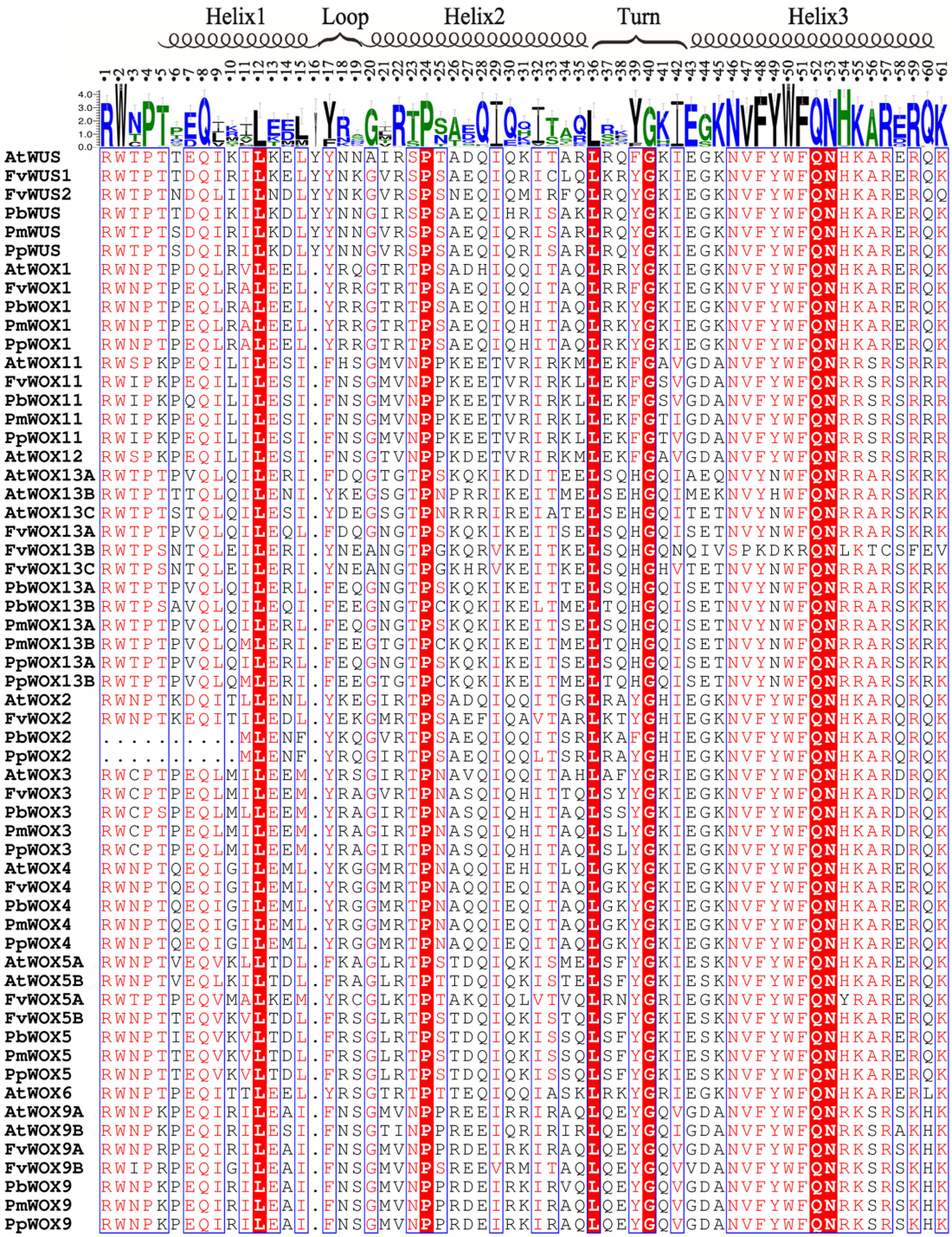


FIGURE 5 | Alignment of the WOX homodomain sequences in five plant species. Highly conserved residues of homodomains were represented by the red shaded blocks in the five tested species. The secondary structure was indicated according to Mayer et al. (1998).

their genetic evolution resulted in species segregation (Tripoli et al., 2005; Jing et al., 2016). Remarkably, with the construction of the phylogenetic tree, the conservation of microsynteny in different families gradually emerged. Furthermore, some flanking genes were not conserved in each microsyntenic group. Therefore, it was speculated that these new genes were later than this duplication event. Interestingly, several lower levels of microsynteny were also found, such as *PpWOX13B*/*FvWOX13A* and *PbWOX13A*/*PmWOX13B* in clade I, *PbWOX11*/*PpWOX11* and *PpWOX11*/*PmWOX11* in clade H (**Figure 3**). These results

strongly suggested that the ancient large-scale duplications could follow by gene rearrangement and loss.

Analysis of Selection Pressures and Functional Divergence

To investigate whether *WOX* genes have undergone strong selection pressures in the evolution of *WOX* gene family, we used site and branch-site models in the CODEML program of PAML software to detect positive selection sites (Yang, 2007). However, no positive selection was detected among these genes (Supplementary Table S3). These finding imply that relaxed purifying selection might play a major role in the evolution of *WOX* genes. Nardmann and Werr (2013) have shown that the WO*X* gene expansion was resulted from the increased complexity of plant morphology (Nardmann and Werr, 2013). These results proposed that the novel members after gene expansion were retained with partly overlapping expression domains and functions. These results are similar with findings of Nardmann and Werr (2013) who reported that with the relaxed purifying, dosage effects will lead to a selective advantage (Nardmann and Werr, 2013).

For further investigation to comprehend significant differences in selection pressures among WUS clade, intermediate clade and ancient clade, the branch-site models were performed using PAML software. It was exposed that no significant positive selection observed in the different branches of *WOX* gene family (Supplementary Table S4). This result was in contrast to the previous report that some significant positive selection sites were fixed in *WOX* genes of peanut (Wang et al., 2015). Because significant positive selection usually exerts its effects only in few sites and in a short period of evolutionary processes, it is difficult to detect positive selection. Thus, the selected signal could be diluted by the purifying selection (Zhang, 2005). However, as the *WOX* coding regions are highly conserved among members of orthologous families, the absence of strong positive selection was expected in Rosaceae species.

Due to the fact that significant positive selection could only detect a limited number of adaptive selection events, we performed a functional divergence analysis according to method used by Cao et al. (2016a). The DIVERGE software was used to calculate functional divergence of type I or II between gene clades in *WOX* genes with posterior analysis. In general, type I functional divergence usually resulted in a specific amino acid selectivity change, i.e., evolutionary rate change. The type II functional divergence only led to the change of physical and chemical properties of amino acids, which were occurred after gene duplication. In present study, to avoid the emergence of false positives, the sites with a posterior probability Q K $>$ 0.9 were set as the key amino acid sites arising the functional differences according to previous experimentation reported (Yin et al., 2013; Cao et al., 2016a). Our results showed that five key amino acid sites (144, 154, 161, 165, and 166) were identified as type I functional divergence between Ancient and Modern (**Table 1**), while just one key site (152) between Intermediate and Modern (**Table 1**). The chi-square test (x 2) found that the *P*-values of Ancient/Modern and Intermediate/Modern were less than 0.05, reaching a significant level. Interestingly, among these three clades, no specific type II functional divergence site (Q K $>$ 0.9) was detected (Supplementary Table S5), suggesting that the physicochemical properties of amino acid sequences between these Rosaceae *WOX* genes were highly identical.

cis-Acting Element Analysis of *WOX* Genes

Two thousand bp sequences of upstream from start codon (ATG) among the putative *WOX* genes, were used for analysis of *WOX* promoters by searching, against the PlantCARE website. Consequently, we detected various types of *cis*-acting elements in the promoter region of 43 *WOX* genes (Supplementary Table S6). These results indicated that the same type of *WOX* might carry out different functions. MBS and ABRE elements were found to be distributed in promoter region of most *WOX* genes, implying that *WOX* genes were transcriptionally regulated upon salt stress and dehydration. Remarkably, we found that the *cis*-elements exhibit significant differences in the promoter regions of duplicated *WOX* genes. These results indicated that the duplicated *WOX* genes may exhibit different regulation features.

Expression Profiles of *F. vesca WOX* Genes

To explore the role of the *WOX* gene family in *F. vesca* development process, the expression of the *FvWOX* genes was explored. The results showed that their expression levels were divergent from each other, indicating that they may be functionally active among all tissues except *FvWOX5B* (**Supplementary Figure S3**), which was located in Pollen with no expression. At the same time, most of *FvWOX* genes exhibited developmental stage-specificity, such as higher expression of *FvWOX13A*, *FvWOX3*, *FvWOX13B*, and *FvWOX1* in flowering, and *FvWOX4*, *FvWOX5,* and *FvWOX5A* in embryo (**Supplementary Figure S3**). Surprisingly, we found that *FvWOX13A*, *FvWOX9A,* and *FvWOX1* were highly expressed among all tissues, indicating that these genes were persistent and very important during development process of *F. vesca*.

TABLE 1 | Analysis of type I functional divergence.

Group 1	Group 2	Θ ± SE	LRT	Q K > 0.9	*P*
Ancient	Intermediate	0.277 ± 0.286	0.938	Not allowed	$P < 0.05$
Ancient	Modern	0.754 ± 0.153	24.289	144,154,161,165,166	$P < 0.05$
Intermediate	Modern	0.368 ± 0.114	10.454	152	$P < 0.05$

SE, standard error; LRT, value of likelihood ratio test.

DISCUSSION

In present study, 43 *WOX* genes from four Rosaceae species were identified. It is observed that no direct relevance between genome sizes and the number of *WOX* gene family members. For example, there was no significant variety in the genome size of *P. bretschneideri* (271.9 Mb) (Wu et al., 2013) and *F. vesca* (240 Mb) (Shulaev et al., 2011), the number of *WOX* genes have been obviously changed. On the contrary, the number of *WOX* genes of the *P. persica* (224.6 Mb) (Verde et al., 2013) and *P. mume* (201 Mb) (Zhang et al., 2012) had a corresponding relationship with their genome sizes. In addition, we also noted that *P. bretschneideri* undergoes two genome-wide duplication events compared with those from *P. persica*, *P. mume,* and *F. vesca* (Wu et al., 2013). Nevertheless, the members of the *WOX* gene family among these four species did not change significantly. These findings indicate that the recent genome-wide duplication event did not contribute to the expansion of *P. bretschneideri WOX* gene family numbers. These results were supported by microsynteny analysis (**Figure 3**).

Previous studies suggested that the *WOX* gene family was divided into three major clades; the ancient clade were mainly present in land plants and green algae, while the intermediate and modern clades were only present in ferns and seed plants (Deveaux et al., 2008; Graaff et al., 2009; Nardmann and Werr, 2012, 2013). In present study, we found that all *WOX* genes from four Rosaceae species were distributed in the three clades, and was supported by the result of exon–intron and conserved domains analysis. At the same time, we also found that each clade contained its specific conserved motifs, implying these specific conserved motifs were likely required for subfamily-specific functions, such as Motif 5 to WUS clade (clades A–F). In the *WOX* gene family, the modern/WUS clade and intermediate clade were evolved from the ancient clade. It is well-known that gene sequence divergence, recombination, and duplications were considered to be the main driving forces for the evolution of gene families (Lin et al., 2014). In our study, the selection pressure was analyzed by using PAML program (Yang, 2007). In general, values of dn/ds (ω) >1, $=1$, and <1 represents positive selection, neutral evolution and purifying selection on the target gene, respectively. In this study, we found that the ω value of *WOX* genes was 0.07304 in M0 model (Supplementary Table S3). These results implied that *WOX* genes from four Rosaceae mainly underwent purifying selection during evolution, which was consistent with the hypothesis that highly conserved genes remain in the genome due to purifying selection. For example, the conserved *WOX* clade genes were all retained from green alga to seed plants (Nei, 2007).

Hedman et al. (2013) found that most conifer *Picea abies WOX* genes expressed at high levels in all developmental stages, while a few *PaWOXs* expression were low in specific tissues (Hedman et al., 2013). Zhang et al. (2015) reported that 10 *Citrullus lanatus WOX* genes were expressed in almost all tissues (Na et al., 2015). In our study, we found that the most of the *FvWOX* genes were expressed in different tissues. Among them, *FvWOX4*, *FvWOX5*, *FvWOX5A,* and *FvWOX9B* were mainly expressed in embryo stage with a very low expression for these genes in other tissues, which implied that these genes might have the same function as the key regulation factor *AtWOX9A* and *AtWOX9B* which was involved in the maintenance of the SAM (Wu et al., 2005; Skylar et al., 2010). The high expression of *FvWOX13A, FvWOX13B,* and *FvWOX13C* (Ancient clade) in flower tissue implied it had an important role similar to *AtWOX13A* and *AtWOX13B* in floral transition (Deveaux et al., 2008).

In this work, we identified 43 *WOX* genes in four Rosaceae species. These genes were divided into three well-supported clades (ancient, modern/WUS, intermediate) with nine subgroups. We also found that *WOX* genes phylogenetic relationship was supported by the presence of gene structure and conserved motif distribution. Our study demonstrated the existence of extensive microsynteny between *WOX* genes by comparing the *WOX* genes across four Rosaceae genomic sequences. The results showed that the maintenance of gene copy number after a whole genome duplication event was the main force to shape the *WOX* family evolution, with the purifying selection and a period of possibly relaxed constraint. Functional divergence was detected among the ancient, intermediate, and modern clades, which leaded to functional constraints, especially different evolutionary rates, after gene duplication. Furthermore, the expression profile of *FvWOX* gene identified that these genes play crucial roles in the floral transition during strawberry growth and development. The comprehensive analysis of the *WOX* family genes and the preliminary results presented here will be useful in the selection of appropriate candidate genes for further research on biological functions of *WOX* genes in strawberry.

AUTHOR CONTRIBUTIONS

YuC and YH conceived and designed the experiments; YuC, QJ, and YH performed the experiments; YuC, YH, and DM analyzed the data; YuC, YH, DL, GL, MA, YL, and YoC contributed reagents/materials/analysis tools; YuC and YH wrote the paper.

ACKNOWLEDGMENTS

This study was supported by The National Natural Science Foundation of China (grant 31640068) and 2017 Graduate innovation fund of Anhui Agriculture University (2017yjs-31).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00078/full#supplementary-material

FIGURE S1 | Maximum-Likelihood tree of *WOX* family members in four Rosaceae species.

FIGURE S2 | Minimum-Evolution tree of *WOX* family members in four Rosaceae species.

FIGURE S3 | Expression patterns of *FvWOX* genes during strawberry growth and development.

REFERENCES

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43, W39–W46. doi: 10.1093/nar/gkv416

Cannon, S. B., Mccombie, W. R., Sato, S., Tabata, S., Denny, R., Palmer, L., et al. (2003). Evolution and microsynteny of the apyrase gene family in three legume genomes. *Mol. Genet. Genomics* 270, 347–361. doi: 10.1007/s00438-003-0928-x

Cao, Y., Han, Y., Jin, Q., Lin, Y., and Cai, Y. (2016a). Comparative genomic analysis of the GRF genes in Chinese pear (*Pyrus bretschneideri* Rehd), poplar (Populous), grape (*Vitis vinifera*), Arabidopsis and rice (*Oryza sativa*). *Front. Plant Sci.* 7:1750.

Cao, Y., Han, Y., Li, D., Lin, Y., and Cai, Y. (2016b). MYB transcription factors in chinese pear (*Pyrus bretschneideri* Rehd.): genome-wide identification, classification, and expression profiling during fruit development. *Front. Plant Sci.* 7:577. doi: 10.3389/fpls.2016.00577

Cao, Y., Han, Y., Li, D., Lin, Y., and Cai, Y. (2016c). Systematic analysis of the 4-coumarate:coenzyme a ligase (4CL) related genes and expression profiling during fruit development in the chinese pear. *Genes* 7:89.

Cao, Y., Han, Y., Meng, D., Li, D., Jin, Q., Lin, Y., et al. (2016d). Structural, evolutionary, and functional analysis of the class III peroxidase gene family in chinese pear (*Pyrus bretschneideri*). *Front. Plant Sci.* 7:1874.

Carmel, L., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2007). Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.* 17, 1034–1044. doi: 10.1101/gr.6438607

Darwish, O., Slovin, J. P., Kang, C., Hollender, C. A., Geretz, A., Houston, S., et al. (2013). SGR: an online genomic resource for the woodland strawberry. *BMC Plant Biol.* 13:223. doi: 10.1186/1471-2229-13-223

Deveaux, Y., Toffanonioche, C., Claisse, G., Thareau, V., Morin, H., Laufs, P., et al. (2008). Genes of the most conserved WOX clade in plants affect root and flower development in Arabidopsis. *BMC Evol. Biol.* 8:291. doi: 10.1186/1471-2148-8-291

Dickinson, T. A., Lo, E., and Talent, N. (2007). Polyploidy, reproductive biology, and Rosaceae: understanding evolution and making classifications. *Plant Syst. Evol.* 266, 59–78. doi: 10.1007/s00606-007-0541-2

Etchells, J. P., Provost, C. M., Mishra, L., and Turner, S. R. (2013). WOX4 and WOX14 act downstream of the PXY receptor kinase to regulate plant vascular proliferation independently of any role in vascular organisation. *Development* 140, 2224–2234. doi: 10.1242/dev.091314

Ge, Y., Liu, J., Zeng, M., He, J., Qin, P., Huang, H., et al. (2016). Identification of WOX family genes in *Selaginella kraussiana* for studies on stem cells and regeneration in lycophytes. *Front. Plant Sci.* 7:93. doi: 10.3389/fpls.2016.00093

Gehring, W. J. (1992). The homeobox in perspective. *Trends Biochem. Sci.* 17, 277–280. doi: 10.1016/0968-0004(92)90434-B

Gouet, P., Robert, X., and Courcelle, E. (2005). ESPript/ENDscript: sequence and 3D information from protein structures. *Acta Crystallogr.* 61, 42–43. doi: 10.1107/S0108767305098211

Graaff, E. V. D., Laux, T., and Rensing, S. A. (2009). The WUS homeobox-containing (WOX) protein family. *Genome Biol.* 10:248. doi: 10.1186/gb-2009-10-12-248

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664–1674. doi: 10.1093/oxfordjournals.molbev.a026080

Gu, X. (2006). A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol. Biol. Evol.* 23, 1937–1945. doi: 10.1093/molbev/msl056

Gu, X., Zou, Y., Su, Z., Huang, W., Zhou, Z., Arendsee, Z., et al. (2013). An update of DIVERGE software for functional divergence analysis of protein family. *Mol. Biol. Evol.* 30, 1713–1719. doi: 10.1093/molbev/mst069

Haecker, A., Grosshardt, R., Geiges, B., Sarkar, A., Breuninger, H., Herrmann, M., et al. (1991). Expression dynamics of WOX genes mark cell fate decisions during early embryonic patterning in *Arabidopsis thaliana*. *Development* 131, 657–668. doi: 10.1242/dev.00963

Hedman, H., Zhu, T., Von, A. S., and Sohlberg, J. J. (2013). Analysis of the WUSCHEL-RELATED HOMEOBOX gene family in the conifer picea abies reveals extensive conservation as well as dynamic patterns. *BMC Plant Biol.* 13:89. doi: 10.1186/1471-2229-13-89

Hirakawa, Y., Kondo, Y., and Fukuda, H. (2010). TDIF peptide signaling regulates vascular stem cell proliferation via the WOX4 homeobox gene in Arabidopsis. *Plant Cell* 22, 2618–2629. doi: 10.1105/tpc.110.076083

Hu, B., Jin, J., Guo, Y. A., Zhang, H., Luo, J., and Gao, G. (2014). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31:1296. doi: 10.1093/bioinformatics/btu817

Ikeda, M., and Ohme-Takagi, M. (2009). Arabidopsis WUSCHEL is a bifunctional transcription factor that acts as a repressor in stem cell regulation and as an activator in floral patterning. *Plant Cell* 21, 3493–3505. doi: 10.1105/tpc.109.069997

Jing, J., Kong, J., Qiu, J., Zhu, H., Peng, Y., and Jiang, H. (2016). High level of microsynteny and purifying selection affect the evolution of WRKY family in Gramineae. *Dev. Genes Evol.* 226, 15–25. doi: 10.1007/s00427-015-0523-2

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 40, D302–D305. doi: 10.1093/nar/gkr931

Lian, G., Ding, Z., Wang, Q., Zhang, D., and Xu, J. (2014). Origins and evolution of WUSCHEL-related homeobox protein family in plant kingdom. *Sci. World J.* 2014, 534140–534140. doi: 10.1155/2014/534140

Lin, Y., Cheng, Y., Jin, J., Jin, X., Jiang, H., Yan, H., et al. (2014). Genome duplication and gene loss affect the evolution of heat shock transcription factor genes in legumes. *PLoS ONE* 9:e102825. doi: 10.1371/journal.pone.0102825

Liu, S. L., Zhuang, Y., Zhang, P., and Adams, K. L. (2009). Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. *Mol. Biol. Evol.* 26, 875–891. doi: 10.1093/molbev/msp011

Mayer, K. F., Schoof, H., Haecker, A., Lenhard, M., Jürgens, G., and Laux, T. (1998). Role of WUSCHEL in regulating stem cell fate in the Arabidopsis shoot meristem. *Cell* 95, 805–815. doi: 10.1016/S0092-8674(00)81703-1

Mukherjee, K., Brocchieri, L., and Bürglin, T. R. (2009). A comprehensive classification and evolutionary analysis of plant homeobox genes. *Mol. Biol. Evol.* 26, 2775–2794. doi: 10.1093/molbev/msp201

Na, Z., Xing, H., Bao, Y., Bo, W., Liu, L., Dai, L., et al. (2015). Genome-wide identification and expression profiling of WUSCHEL-related homeobox (WOX) genes during adventitious shoot regeneration of watermelon (*Citrullus lanatus*). *Acta Physiol. Plant.* 37, 1–12.

Nardmann, J., and Werr, W. (2012). The invention of WUS-like stem cell-promoting functions in plants predates leptosporangiate ferns. *Plant Mol. Biol.* 78, 123–134. doi: 10.1007/s11103-011-9851-4

Nardmann, J., and Werr, W. (2013). Symplesiomorphies in the WUSCHEL clade suggest that the last common ancestor of seed plants contained at least four independent stem cell niches. *New Phytol.* 199, 1081–1092. doi: 10.1111/nph.12343

Nei, M. (2007). The new mutation theory of phenotypic evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12235–12242. doi: 10.1073/pnas.0703349104

Palovaara, J., Hallberg, H., Stasolla, C., and Hakman, I. (2010). Comparative expression pattern analysis of WUSCHEL-related homeobox 2 (WOX2) and WOX8/9 in developing seeds and somatic embryos of the gymnosperm *Picea abies*. *New Phytol.* 188, 122–135. doi: 10.1111/j.1469-8137.2010.03336.x

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2011). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., and Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 13, 1512–1517. doi: 10.1016/S0960-9822(03)00558-X

Romera-Branchat, M., Ripoll, J. J., Yanofsky, M. F., and Pelaz, S. (2013). The WOX13 homeobox gene promotes replum formation in the *Arabidopsis thaliana* fruit. *Plant J.* 73, 37–49. doi: 10.1111/tpj.12010

Sarkar, A. K., Luijten, M., Miyashima, S., Lenhard, M., Hashimoto, T., Nakajima, K., et al. (2007). Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* 446, 811–814. doi: 10.1038/nature05703

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Skylar, A., Hong, F., Chory, J., Weigel, D., and Wu, X. (2010). STIMPY mediates cytokinin signaling during shoot meristem establishment in *Arabidopsis* seedlings. *Development* 137, 541–549. doi: 10.1242/dev.041426

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Tripoli, G., D'elia, D., Barsanti, P., and Caggese, C. (2005). Comparison of the oxidative phosphorylation (OXPHOS) nuclear genes in the genomes of *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. *Genome Biol.* 6:R11. doi: 10.1186/gb-2005-6-2-r11

Ueda, M., Zhang, Z., and Laux, T. (2011). Transcriptional activation of Arabidopsis axis patterning genes WOX8/9 links zygote polarity to embryo development. *Dev. Cell* 20, 264–270. doi: 10.1016/j.devcel.2011.01.009

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., et al. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494. doi: 10.1038/ng.2586

Wang, P., Li, C., Li, C., Zhao, C., Xia, H., Zhao, S., et al. (2015). Identification and expression dynamics of three WUSCHEL related homeobox 13 (WOX13) genes in peanut. *Dev. Genes Evol.* 225, 221–233. doi: 10.1007/s00427-015-0506-3

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., et al. (2013). The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* 23, 396–408. doi: 10.1101/gr.144311.112

Wu, X., Dabi, T., and Weigel, D. (2005). Requirement of homeobox gene STIMPY/WOX9 for Arabidopsis meristem growth and maintenance. *Curr. Biol.* 15, 436–440. doi: 10.1016/j.cub.2004.12.079

Xiaoxu, L., Cheng, L., Wei, L., Zenglin, Z., Xiaoming, G., Hui, Z., et al. (2016). Genome-wide identification, phylogenetic analysis and expression profiling of the WOX family genes in *Solanum lycopersicum*. *Hereditas* 38, 444–460. doi: 10.16288/j.yczz.15-499

Xin, Z., Jie, Z., Liu, J., Yin, J., and Zhang, D. (2010). Genome-wide analysis of WOX Gene Family in Rice, Sorghum, Maize, Arabidopsis and Poplar. *J. Integr. Plant Biol.* 52, 1016–1026. doi: 10.1111/j.1744-7909.2010.00982.x

Yadav, R. K., Tavakkoli, M., and Reddy, G. V. (2010). WUSCHEL mediates stem cell homeostasis by regulating stem cell number and patterns of cell division and differentiation of stem cell progenitors. *Development* 137, 3581–3589. doi: 10.1242/dev.054973

Yan, H. H., Mudge, J., Kim, D. J., Shoemaker, R. C., Cook, D. R., and Young, N. D. (2004). Comparative physical mapping reveals features of microsynteny between *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. *Genome* 47, 141–155. doi: 10.1139/g03-106

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yin, G., Xu, H., Xiao, S., Qin, Y., Li, Y., Yan, Y., et al. (2013). The large soybean (*Glycine max*) WRKY TF family expanded by segmental duplication events and subsequent divergent selection among subgroups. *BMC Plant Biol.* 13:148. doi: 10.1186/1471-2229-13-148

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zhang, J. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479. doi: 10.1093/molbev/msi237

Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., et al. (2012). The genome of *Prunus mume*. *Nat. Commun.* 3:1318. doi: 10.1038/ncomms2290

Zhang, Y., Yue, J., Liu, Z., and Zhu, Y. X. (2015). ROW1 maintains quiescent centre identity by confining WOX5 expression to specific cells. *Nat. Commun.* 6:6003. doi: 10.1038/ncomms7003

Zhao, Y., Hu, Y., Dai, M., Huang, L., and Zhou, D. X. (2009). The WUSCHEL-related homeobox gene WOX11 is required to activate shoot-borne crown root development in rice. *Plant Cell* 21, 736–748. doi: 10.1105/tpc.108.061655

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 May 2017
doi: 10.3389/fpls.2017.00775

Genome Wide Identification of Orthologous *ZIP* Genes Associated with Zinc and Iron Translocation in *Setaria italica*

Ganesh Alagarasan[1], Mahima Dubey[1], Kumar S. Aswathy[2] and Girish Chandel[1]*

[1] Department of Plant Molecular Biology and Biotechnology, Indira Gandhi Agricultural University, Raipur, India, [2] Department of Agricultural Microbiology, Tamil Nadu Agricultural University, Coimbatore, India

Genes in the *ZIP* family encode transcripts to store and transport bivalent metal micronutrient, particularly iron (Fe) and or zinc (Zn). These transcripts are important for a variety of functions involved in the developmental and physiological processes in many plant species, including most, if not all, Poaceae plant species and the model species Arabidopsis. Here, we present the report of a genome wide investigation of orthologous *ZIP* genes in *Setaria italica* and the identification of 7 single copy genes. RT-PCR shows 4 of them could be used to increase the bio-availability of zinc and iron content in grains. Of 36 *ZIP* members, 25 genes have traces of signal peptide based sub-cellular localization, as compared to those of plant species studied previously, yet translocation of ions remains unclear. *In silico* analysis of gene structure and protein nature suggests that these two were preeminent in shaping the functional diversity of the *ZIP* gene family in *S. italica*. NAC, bZIP and bHLH are the predominant Fe and Zn responsive transcription factors present in *SiZIP* genes. Together, our results provide new insights into the signal peptide based/independent iron and zinc translocation in the plant system and allowed identification of *ZIP* genes that may be involved in the zinc and iron absorption from the soil, and thus transporting it to the cereal grain underlying high micronutrient accumulation.

Keywords: zinc and iron regulated transporters, signal peptide, *Setaria italica*, expression profiling, gene characterization

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Thomas Triplet,
École Polytechnique de Montréal, Canada
Vincenzo Bonnici,
University of Verona, Italy

****Correspondence:***
Ganesh Alagarasan
alagarasan.ganesh@hotmail.com

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

Received: *07 December 2016*
Accepted: *25 April 2017*
Published: *15 May 2017*

Citation:
Alagarasan G, Dubey M, Aswathy KS and Chandel G (2017) Genome Wide Identification of Orthologous ZIP Genes Associated with Zinc and Iron Translocation in Setaria italica. Front. Plant Sci. 8:775. doi: 10.3389/fpls.2017.00775

INTRODUCTION

Bio-fortification of food crops with Fe and Zn remains a priority area of research. Iron (Fe) and Zinc (Zn) are basic nutrient elements for plants, which assist metabolism and development in plant parts (Haydon and Cobbett, 2007; Samira et al., 2013; Kabir et al., 2014). Plants face challenges in maintaining homeostasis of these two metals, as they may generate highly reactive hydroxyl radicals. The hydroxyl radicals can harm most cell parts, for example, DNA, proteins, lipids and sugars. Zinc serves as an essential basic element in many proteins, including DNA-binding Zn-finger protein (Vallee and Auld, 1990; Rhodes and Klug, 1993), RING finger proteins and LIM domain- containing proteins (Vallee and Falchuk, 1993), whereas iron plays a significant part in electron transfer in photosynthesis and respiration. Thus, plants have developed a firmly controlled framework to balance the uptake and storage of these metal ions (Grotz and Guerinot, 2006;

Palmgren et al., 2008; Chandel et al., 2010). Accordingly, Fe and Zn homeostasis in plants have clearly evolved. Since a deficiency of nutrients like Zinc and Iron diminishes the growth of plants, for example influencing rice grain production, both in terms of quantity and quality, whereas over-abundance of Zn and Fe might cause significant toxicity to some biological systems (Pahlsson, 1989; Price and Hendry, 1991). Various metal transporters are available in plants, which pass the metal ions over the layer in the cytoplasm that maintains metal homeostasis (Kambe et al., 2004; Taylor et al., 2004; Colangelo and Guerinot, 2006; Barberon et al., 2014). These include the P-type ATPase (P1B) family, Zinc & Iron-regulated transporter – like Protein (ZIP) (Milner et al., 2012; Thakur et al., 2016), Normal Resistance-Related Macrophage Protein (NRAMP), and the Cation Dissemination Facilitator (CDF) family (Colangelo and Guerinot, 2006; Palmer et al., 2014). It has been reported that OsZIP4, OsZIP5 and OsZIP8 are functional zinc transporters and are localized to the plasma membrane (PM) (Ishimaru et al., 2005; Lee et al., 2010a,b). AtIRT2 is an iron transporter and is localized to the intracellular vesicles, suggesting a crucial role in preventing metal toxicity through compartmentalization and remobilizing iron stores from inner storage vesicles (Vert et al., 2009).

ZRT and the IRT-like protein (ZIP) family has been described far and wide in living beings, including archaea, bacteria, parasites, plants and has been seen with high micronutrient contributor in the endosperm of minor millets. The ZIP family gene proteins comprise 300–500 amino acid residues with six to nine transmembrane domains and besides, a similar membrane topology can transport various divalent cations, including Fe^{2+}, Zn^{2+}. AtIRT1 was the first individual from the ZIP protein family to be recognized in a yeast mutant defective in iron uptake through functional complementation, and it encodes a major Fe transporter at the root surface in Arabidopsis (Eide et al., 1996; Varotto et al., 2002; Vert et al., 2002).

Minor millets, being nutritiously rich, serve as vital focuses for discovering potential qualities. Foxtail millet is a food security crop in low rain-fed regions. The distinguishing proof of ZIP gene orthologs from micronutrient-rich foxtail millet will unravel their gene reservoir and in the meanwhile will furnish valuable and effective genes for the enhancement of micronutrients in other crops.

A better understanding of the roles and functions of each of the members of the *Setaria italica* ZIP family should lead to new insights into micronutrient homeostasis. Identifying and testing its potentiality in metal transport had been a primary goal of such an effort. Other important features of metal transporters that were focused on in this study are the gene structures of ZIP transporters, whether they have introns or intronless, and the regulation of tissue specific ZIP gene expression. Gaining a better understanding of the *S. italica* ZIP family should also help us better understand micronutrient nutrition in other cereal crop, as the ZIP family of transport proteins is found in all branches of life, including animals, plants, fungi, and protists (Guerinot, 2000). Palmer et al. (2014) reported a genome wide characterization of various ZIP transporters, including spatio-temporal gene expression analysis in one of the closely related C_4 plant species. To date, no or only a few members of the ZIP family have been characterized in *S. italica* regarding their transport capabilities. We try to put this work into context by stating that such findings will help in reducing malnutrition. Our study will serve as preliminary findings to characterize and functionally validate the single copy orthologs and the functions of signal peptide in plant system.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

The experiment was conducted under protected polyhouse conditions (16 h of photoperiod per day at 30°C) at a geographical location of N 21° 14′ 6.298″E 81°42′ 50.424″. Since the impact of geographical location of plants remain as potential aspect to consider in nutrient accumulation and biological activities, we mentioned the precise location of crop grown area. From the panel of millet genotypes, foxtail millet *Co (Te)7* variety which has greenish purple foliage and yellow grains and little millet cultivar (BL-4, RLM-37 and OLM-203) which has greenish foliage and dark gray grains having high Fe and Zn content was selected. Seeds were treated with 0.1% Bavistin to reduce fungal contamination before sowing. Watering was done once in a week and no nutrient supplementation was given for 3 months of the entire growth period. Completely developed grains were collected from the plants and subjected to micronutrient investigation.

Elemental Analysis-Atomic Absorption Spectrophotometry

Entire grains of foxtail millet variety and little millet cultivar seeds were physically dehusked using sand paper, followed by the estimation of micronutrients (Stangoulis and Sison, 2008). Fe and Zn concentrations were assessed according to HarvestPlus guidelines[1] using an atomic absorption spectrophotometer (AAS200) considering tomato leaf powder as standard with minor modifications.

Database Searches for ZIP Family Genes

All members of the ZIP gene family were exhaustively retrieved from the Gramene database[2] (Tello-Ruiz et al., 2016) for the two reference plant species *Arabidopsis thaliana*[3] and *Oryza sativa*.[4] The retrieved sequences were cross checked with RGAP (Kawahara et al., 2013) and TAIR (Berardini et al., 2015) database for data reliability. The result was confirmed by doing a BLAST analysis against Arabidopsis and Rice genome databases. The accession numbers of published ZIP genes from Arabidopsis and rice along with chromosome coordinates and other information are listed in Supplementary Table S2. ZIP genetic information, including the number of amino acids, cds length and chromosome locations were obtained from the Gramene database. Physical parameters of the ZIP proteins,

[1]http://www.harvestplus.org/content/crop-sampling-protocols-micronutrient-analysis

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

[3]https://www.arabidopsis.org/Blast/index.jsp

[4]http://rice.plantbiology.msu.edu/analyses_search_blast.shtml

including isoelectric point (pI), and molecular mass (kDa) were calculated using the compute pI/Mw tool in the ExPASy[5], with parameters set to 'average' (Gasteiger et al., 2005). The gene sequences *viz* CDS, intron, exon and UTR regions were used to mine SSRs in the SSR identification tool[6].

Genome Wide Investigation of ZIP Orthologs and Membrane Topology

Here we performed a genome wide survey using OrthoVenn, aimed at identifying orthologs of ZIP genes across three plant species; *O. sativa*, *A. thaliana* and *S. italica*[7] (Wang et al., 2015). Thirteen ZIP protein sequences from Rice and 16 from Arabidopsis were used to identify orthologs within a whole genome sequence of foxtail millet. The analysis parameters of OrthoVenn were as follows: cutoff for all-too-all protein similarity comparisons (*E*-value 1*e*−5); and Inflation value (1.5) to generate ortholog clusters using the Markov Cluster Algorithm (Enright et al., 2002). The putative transmembrane topology for each of the ZIP proteins was predicted using PROTTER (version 1.0)[8].

Mapping of ZIP Genes on Chromosomes and Gene Structure Prediction

The chromosome positioning of the Arabidopsis, rice and foxtail millet ZIP genes were generated using TAIR[9], Oryzabase[10] and Mapchart 2.3 (Voorrips, 2002) respectively. GSDS[11] was used to predict the exon and intron structures of the individual ZIP genes through alignment of the CDS with their corresponding genomic DNA sequences.

Molecular Modeling and Phylogenetic Analysis of ZIPs

Multiple sequence alignment of the full length amino-acid sequences of the ZIP proteins were performed by Clustal X2.0.10 (Thompson et al., 1997). An effective phylogenetic tree was developed using the W-IQ-TREE online server (Trifinopoulos et al., 2016) with default options. The SWISSMODEL workspace was used to build homology models of the ZIPs by automated protein structure modeling and the ExPASy web server.

Motif Analysis of ZIP Protein Sequences and Signal Peptide Prediction

The MEME program software, version 4.9.0 (Bailey and Elkan, 1994) was used to analyze the full length protein sequences of the ZIP genes for motif variation. The motif selection was set to 10 as the maximum number, with a minimum and maximum width of 6 and 50 amino acids, in order to locate the conserved motif. The Distribution of any number of repetitions was considered, while the other factors were of default settings. An upstream sequence of 1KB was subjected to promoter analysis through PlantPAN http://PlantPAN2.itps.ncku.edu.tw (Chow et al., 2015). Protein localization was predicted by TragetP http://www.cbs.dtu.dk/services/TargetP/ (sub-cellular localization) and SignalP http://www.cbs.dtu.dk/services/SignalP/ web servers.

Tissue Specific *In Silico* Expression Profiling of ZIP Genes in Foxtail Millet

The European Nucleotide Archive[12] was used to retrieve Illumina RNA-HiSeq reads from four tissues of foxtail millet- namely Root (SRX128223), Stem (SRX128225), Leaf (SRX128224) and Spica (SRX128226), a drought stress library (SRR629694) and its control (SRR629695) (Zhang et al., 2012; Qi et al., 2013). The NGS Toolkit[13] was employed to filter the reads, and the CLC Genomics Workbench 8[14] was used to map the reads onto the gene sequences of -*S. italica*. The normalization of the mapped reads was done using the RPKM (reads per kilobase per million) method. Based on the RPKM values, the heat map for tissue-specific expression profile was generated for each gene in all tissue samples using the TIGR MultiExperiment Viewer (MeV v 4.9) software package (Saeed et al., 2003).

Validation of Functional Orthologs

To validate our *in silico* findings, we have measured the abundance of transcript present in SiZIP orthologous genes. For validation of functional ortholog, foxtail millet seeds were surface sterilized and sown in a pot containing soil and allowed to grow for 15 days at above mentioned growth conditions. Collected tissues were frozen in liquid nitrogen and quickly stored at −80. Total RNA was isolated from the shoots of by using TRIzol reagent, according to the manufacturer's protocol (Invitrogen, USA). A one step Reverse-transcription reactions involved 1 μl of total RNA by use of the SuperScript III platinum RT-PCR system. The gene-specific primers were designed from the foxtail millet ZIP1, ZIP3, ZIP3, ZIP4, ZIP5, ZIP6, and ZIP7 genes. An RT-PCR program initially started with 55°C for 30 min; 94°C denaturation for 2 min, followed by 40 cycles of 94°C for 15 s, 60–62°C for 30 s and 68°C for 30 s, 68°C annealing for 5 min. Actin gene was used for internal control gene amplification.

Comparative Expression Analysis of SiZIP Gene Homolog in Other Millet Crop

Two foxtail millet genes were selected based on their expression level. Comparative expression analysis of two foxtail millet ZIP gene homologs (ortholog/paralog) was carried out in other millet crop, i.e., little millet (*Panicum sumatrense*) to find out the existence of SiZIP homologs and its expression level at different tissues. Experimental condition (plant growth condition and expression analysis) in little millet is same as mentioned above for foxtail millet. RNA was isolated from stem, leaf and spica at the panicle emergence stage. All tests were repeated two times,

[5]http://www.expasy.org/tools/

[6]http://www.gramene.org/db/markers/ssrtool

[7]http://probes.pw.usda.gov/OrthoVenn

[8]http://wlab.ethz.ch/protter/start/

[9]https://www.arabidopsis.org

[10]http://viewer.shigen.info/oryzavw/maptool/MapTool.do

[11]http://gsds.cbi.pku.edu.cn/

[12]http://www.ebi.ac.uk/ena

[13]http://www.nipgr.res.in/ngsqctoolkit.html

[14]https://www.qiagenbioinformatics.com/

and one of the repeats is shown in the figures. PCR products were resolved by 2.5% agarose gel electrophoresis and stained with EtBr. The gel images were captured using Bio-Rad gel documentation system.

RESULTS

Grain Nutrients and ZIP Ortholog Analysis in Foxtail Millet

Fe and Zn estimation revealed that the distribution of zinc and iron contents in foxtail millet varies with the rice. Estimated amounts of 27.19 ± 1.05 μg/g of iron and 40.40 ± 0.23 μg/g of zinc (mean and SE value of the replicated data) were present in *S. italica*.

Genome-wide analysis of orthologous clusters is an important part of comparative genomics study. Identification of overlap among orthologous clusters can enable us to elucidate the role and evolution of proteins across Arabidopsis, rice and foxtail millet species. Orthologs or orthologous genes are clusters of genes in distinct species that originated by vertical descent from a single gene in the last common ancestor. Based on the results of syntenic analysis, precise findings concerning ZIP gene family orthologs were obtained. Well-annotated and well-characterized ZIP family genes from Arabidopsis and rice were used to find orthologs from a whole genome sequence of foxtail millet. Out of 35,471 proteins in foxtail millet, 7 were found to be ZIP ortholog for rice and Arabidopsis (**Figure 1**). Among these three genomes, seven orthologous clusters were obtained. Cluster 1 had a maximum of six proteins, in which Arabidopsis shared four genes (AtIRT1, AtZIP8, AtIRT2 and AtZIP10). Three overlapping orthologous gene clusters were found in the Arabidopsis genome, whereas one was found in the rice genome and none in foxtail millet. Overlapping orthologous genes were distributed on different chromosomes from a single genome in rice and Arabidopsis. Further, there was no multi copy of orthologs found in the foxtail millet genome. The gene IDs for identified orthologs in foxtail millet are given in Supplementary Table S2. Single copy gene clusters are represented in **Figure 1**, and the predicted gene structure of these genes are shown in **Supplementary Figure S1**.

Chromosomal Distribution of the ZIP Family in Three Species Genomes

Thirty-six genes were identified as members of the ZIP gene family, including 16 genes in Arabidopsis, 13 in rice and 7 from foxtail millet. Multiple sequence alignment of predicted proteins was shown in **Supplementary Datasheet S1**. Based on these findings, the chromosomal location of ZIP genes was determined for the three species. The results showed an uneven distribution of the 36 ZIP genes on all chromosomes of the three species as shown in **Figure 2**. The genome maps of the ZIP genes showed that AtZIPs were found across all chromosomes of Arabidopsis (Chr. 1,2,3,4, and 5), while OsZIPs were distributed on 7 out of 12 chromosomes (Chr. 1, 3, 4, 5, 6, 7, and 8). In rice, chromosome 5 had the most ZIP genes (4), followed by OsChr3 (3), OsChr8 (2), and OsChr1, 6, and 7 (1-each). AtChr1 (5); AtChr2, 4, and5 (3); and AtChr4 (1) had the ZIP gene distributed discretely in each chromosome. Among the 29 genes, OsIAR1 encoded the longest protein (498 amino acids [aa]), while the shortest (326 aa) was encoded by AtZIP11. The average length of the proteins encoded by the ZIP proteins was 374 aa. The theoretical pi values of the seven proteins (AtZIP3, AtZIP10, OsZIP1, OsZIP3, OZIP4, OsIRT1, OsIRT2) were above 7, showing that they were alkaline, whereas the proteins encoded by the other genes were acidic (<7). The molecular weights of these proteins ranged from 36,021.5 to 53,578.9 Da, with an average of 39,488.42 Da. In the case of foxtail millet seven predicted ZIP genes were located in the chromosome (3, 6, 7, and 9). The detailed parameters are shown in Supplementary Table S2. Although the distributions of these ZIP genes were diverse, their genetic features and biochemical properties tended to be similar.

Phylogenetic Classification of ZIP Proteins

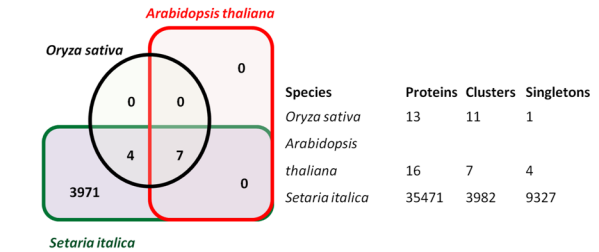
The functional similarity among 36 annotated ZIP genes was explored via phylogenetic analyses of the ZIP protein sequences using the W-IQ-TREE- a maximum likelihood based algorithm. A high bootstrap value suggested a common origin for the ZIP genes of each subgroup. Inspection of the phylogenetic tree topology showed several pairs of ZIP proteins with high homology in the terminal nodes of each subgroup, suggesting that they are putative paralogous pairs (**Figure 3**). The homology modeling structure of SiZIPs is shown in **Supplementary Figure S2**. This outcome upheld the hypothesis that these many sets of paralogous genes may have evolved from a genome duplication event.

Motif Analysis for Single Copy Gene Orthologs

Ten conserved motifs were analyzed using MEME software, and the schematic distribution of these 10 motifs among the ZIP proteins is shown in **Supplementary Figure S3**. A common motif distribution was exhibited by the closely related ZIP members as represented by different clusters in the phylogenetic tree; this suggested functional similarities among the ZIP proteins within the same sub-clusters. The distribution of motifs also highlights that the ZIP genes are supposed to be conserved during evolution. The same motif pattern was seen in (OsZIP3, OsZIP9, Si024505g, AtZIP6, OsZIP6 and Si010244). Si013901g had an identical motif compared to OsIAR1 and AtZIP1 but showed positional difference. The AtZIP1 and AtZIP5 motifs distributions were the same, with minor differences, while AtZIP1 had one extra motif. Promoter analyses revealed that the SiZIP family gene has a maximum of b ZIP, b HLH and NAC Fe and Zn responsive transcription factors (**Figure 4**).

Validation of Functional Orthologs

To further confirming our hypothesis and investigate the role of the ZIP family genes in influencing grain Fe and Zn contents, the expression levels of genes were quantified by RNA-seq analysis. Expression of these genes was analyzed among



Species	Proteins	Clusters	Singletons
Oryza sativa	13	11	1
Arabidopsis thaliana	16	7	4
Setaria italica	35471	3982	9327

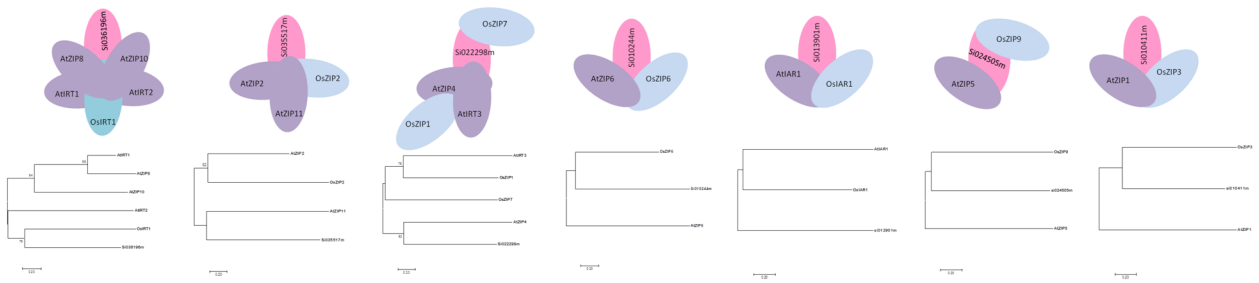


FIGURE 1 | Ortholgous genes for the ZIP family have been identified among three species viz. Arabidopsis, rice and foxtail millet. Seven orthologous clusters along with its Phylogenetic tree have been presented.

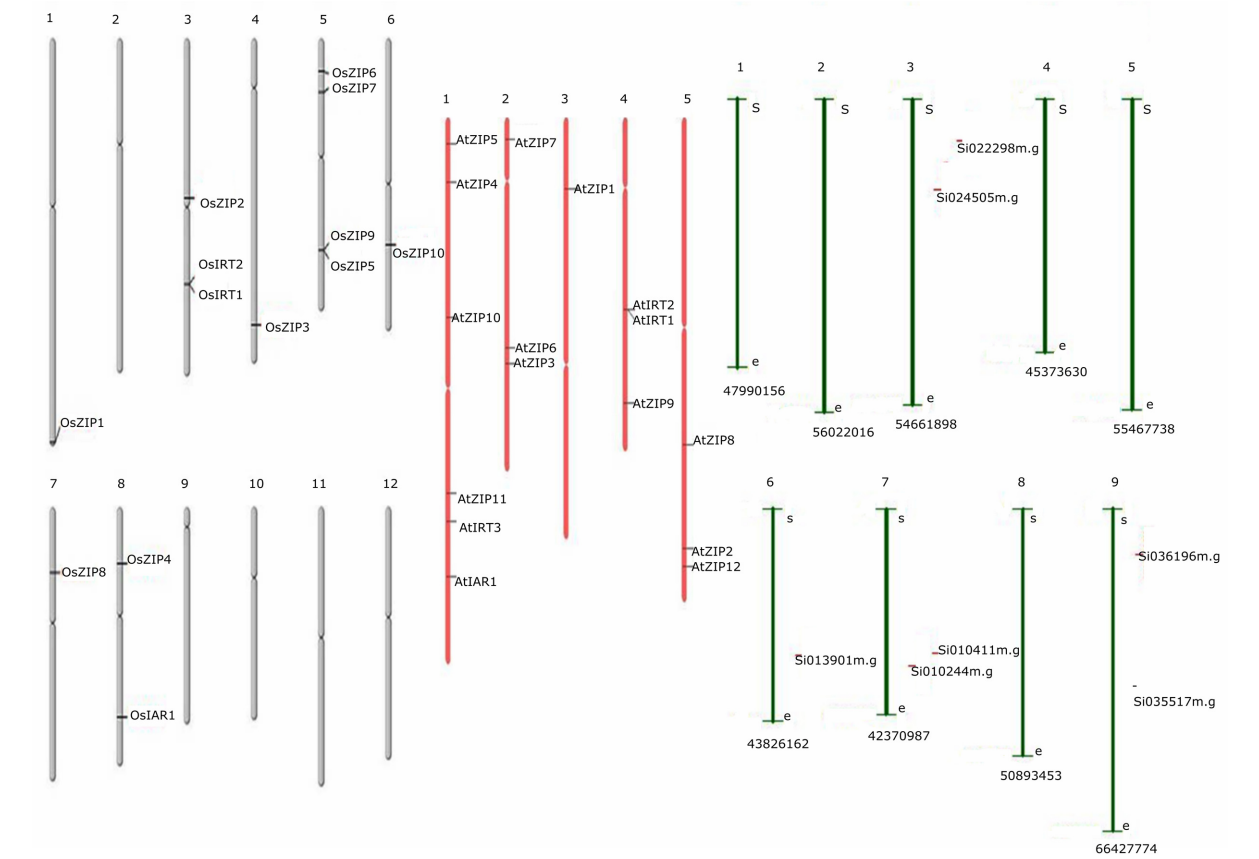


FIGURE 2 | The distribution of ZIPs across the length of the chromosome, along with its name is mentioned on the right side of every chromosomal segment. There were a total of 36 genes, 16 from Arabidopsis, 13 from rice and 7 from foxtail millet. Gray color chromosomes indicates rice, whereas pink and green for Arabidopsis and foxtail millet, respectively.

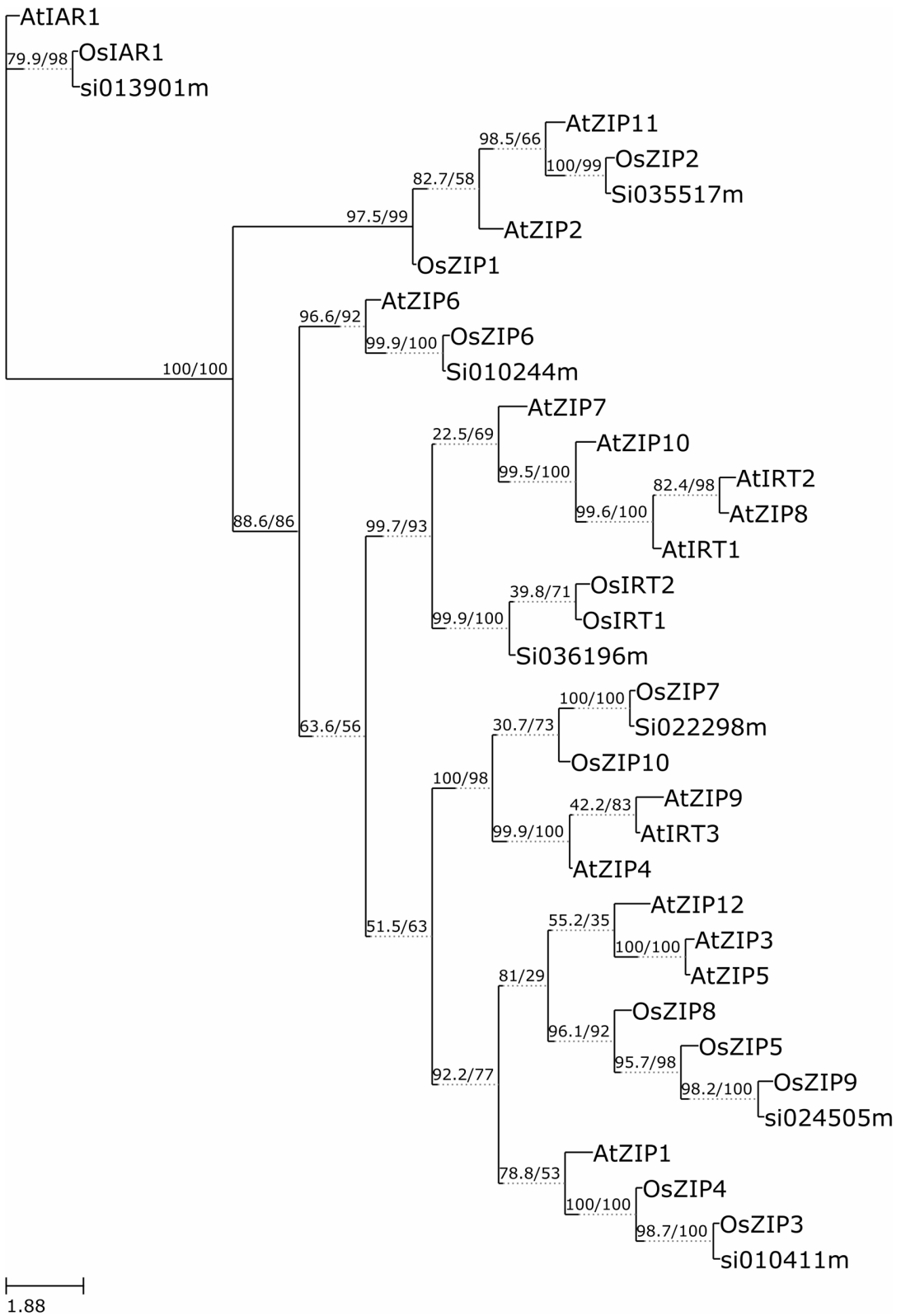


FIGURE 3 | Molecular Phylogenetic analysis by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on W-IQ-TREE. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985). The analysis involved 36 amino acid sequences.

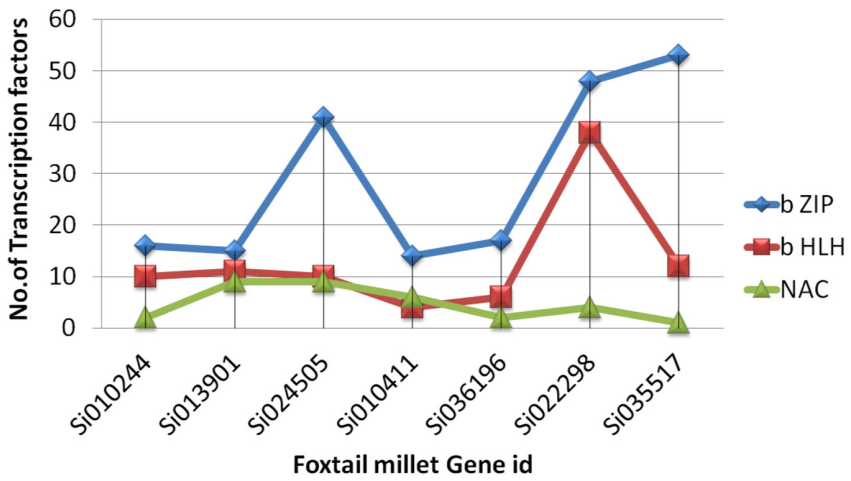


FIGURE 4 | The distribution of zinc and iron responsive (TF) transcription factors across ZIP in *S. italica*.

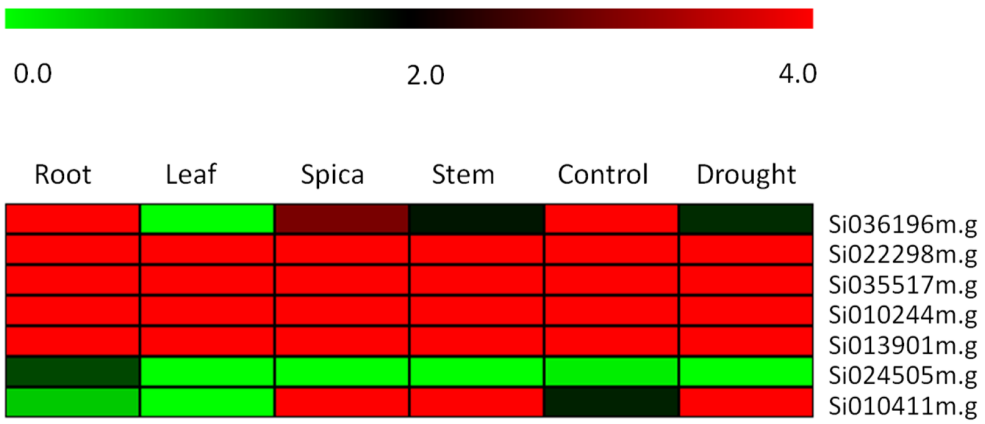


FIGURE 5 | Heatmap showing the expression pattern of *SiZIP* genes in four tissues, namely root, leaf, spica and stem along with control and drought stress library of *Setaria italica*. The colored bar at top left represents relative expression value, where 0.0, 2.0, and 4.0 denotes low, medium, and high expression, respectively.

different tissue types and developmental stages to reveal their precise involvement in the transport, remobilization and grain loading of micronutrients. The expression levels of zinc and iron transporter genes were detected by comparing the RPKM value of different genes expressed in the transcriptome of different tissues. In **Figure 5**, RPKM values are represented as a heat map for each identified ZIP transporter gene. The expression of four genes- Si022298m.g, Si035517m.g, Si010244m.g, and Si013901m.g- in four different tissues remains unique and this was found to be maximum with minor variance. In contrast, in Si010411m.g the expression was upregulated in root, leaf, stem and spica tissues. A negligible level of expression was seen in Si024505m.g in all tissues, and exceptionally moderate expression was observed in the root. Meanwhile, the expression of the Si036196m.g gene in leaf tissue was negligible, while a gradual increase was maintained in stem, spica and root tissues' expression. Interestingly, under stress conditions, the Si036196m.g gene was moderately expressed with maximum expression in control; the opposite was true in Si010411m.g.

For wet lab validation, as well as to confirm the SiZIPs' protein synthesis in SiZIP genes shoot tissues were examined for their transcript abundance. Of seven ZIP genes, four (Si022298m.g, Si035517m.g, Si010244m.g, and Si013901m.g) had relatively good expression levels (**Figure 6**), and they could be used for the bio-fortification process. The list of primers used in this study is given in Supplementary Table S1.

Comparative Expression Analysis of SiZIP Family Gene Homolog in Other Millet Crop (*Panicum sumatrense*)

Of seven SiZIP genes, we observed four different levels of expression pattern in foxtail millet. Hence, to compare the SiZIP homologs expression data with those of another millet crop, we selected little millet (*P. sumatrense*). Three genotypes BL-4,

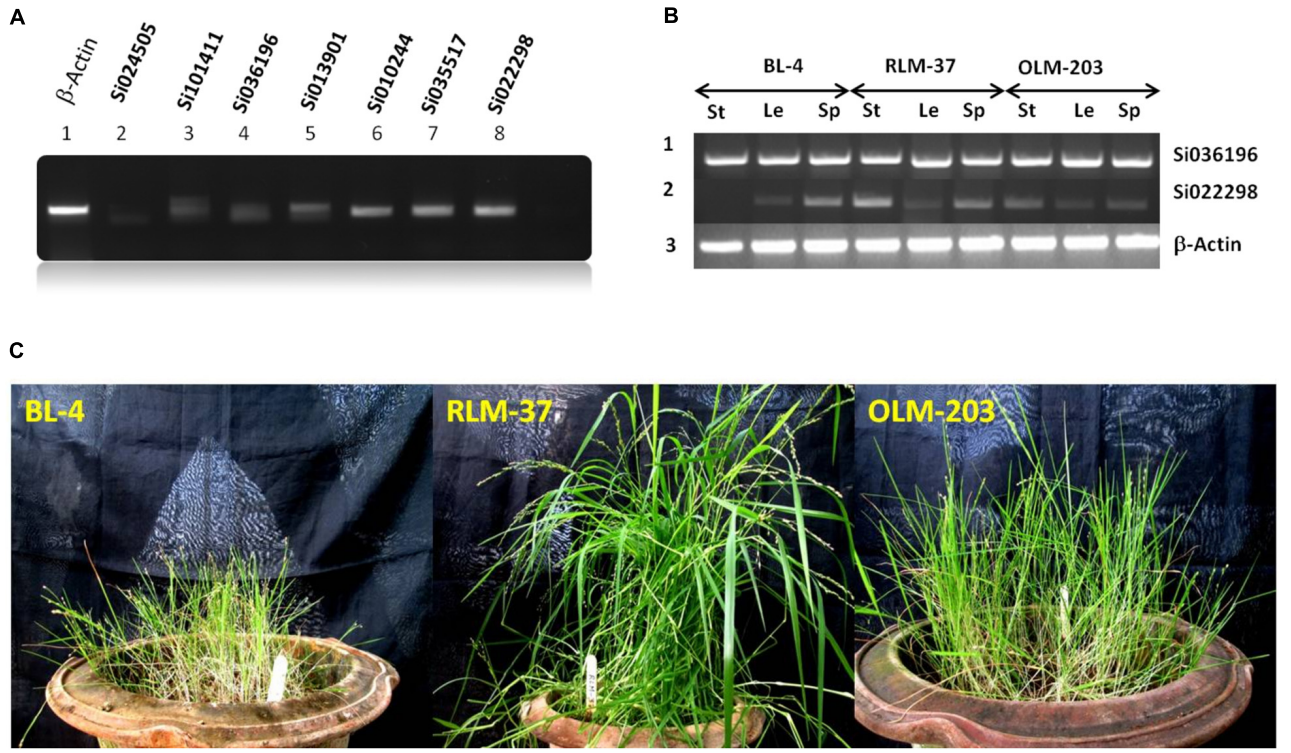


FIGURE 6 | (A) RT-PCR expression analysis of seven ZIP genes from *S. italica* in shoot tissues. **(B)** Differential expression analysis of ZIP family gene in little millet. Le, leaf; St, stem and Sp, spica. **(C)** Standing crop view of little millet genotypes.

RLM-37 and OLM-203 (**Figure 6**) were chosen for differential expression analysis using the two SiZIP family gene primers (Si036196 and Si022298). Although Si036196 expression was up regulated in foxtail millet, its expression level was same and maximal in stem, leaf and spica tissues of all three genotypes in little millet. Gene Si02298 exhibited maximum expression in foxtail millet of all tissue types, but in the case of little millet, medium to low levels of expression were observed (**Figure 6**). Homologs of Si036196 and Si022298 had contrasting types of expression level in little millet. Our comparative analysis and results provides a preliminary data for the cloning of potential ZIP genes from non-sequenced crop like little millet.

Analysis of a Signal Peptide on the Multi-Species (ZIP) Proteins' N-Terminus

Signal peptides are typically cleaved from the mature proteins during transport and/or processing through the endoplasmic reticulum. When comparatively analyzed with cleavable signal peptides in other organisms, the N-termini of ZIP proteins from rice, Arabidopsis, and foxtail millet showed the presence of signal peptide regions in plant iron/zinc transport proteins. The predicted sub- cellular localization and signal peptide cleavage site of ZIP proteins are presented in Supplementary Table S2 and the signal peptide is shown in **Supplementary Figure S4**. From the plant species studied here, it was posited that most transport proteins have signal peptides, and this suggests a role in the active transport of ion molecules, as supported by the creation of mutant lines of signal peptide in *Malus xiaojinensis* (Zhang et al., 2014).

DISCUSSION

The ZIP family genes have been identified in many plants like Arabidopsis and rice. These genes are responsible for the transport of Zn and Fe and are known to play a role in Mn transport. The complete genome sequence of foxtail millet has allowed researchers to identify and characterize various gene families in foxtail millet (Lata et al., 2014; Mishra et al., 2014; Yadav et al., 2014; Muthamilarasan et al., 2015a,b). Although ZIPs have been characterized in many plants, to the best of our insight, there were few or no reports on the functional characterization of the ZIPs in foxtail millet, though it has high micronutrients. The nutritional profile (Muthamilarasan et al., 2016) and genetic improvement of cereal crops using foxtail millet genome has been extensively reviewed (Muthamilarasan and Prasad, 2015).

In our study, we identified and characterized seven SiZIP from *S. italica*. Comparative analysis of SiZIPs with other species showed that foxtail millet has fewer orthologous genes (7) then rice (8) or Arabidopsis (12). The protein properties of SiZIP, AtZIP and OsZIP revealed many differences in amino acid

length, isoelectric point, molecular weight and trans-membrane domains. In addition, the proteins had different signaling peptides, which could result from the presence of novel splice varients.

To analyze the evolutionary relationship of SiZIPs, a fair phylogenetic tree comprising ZIPs from Arabidopsis and rice was constructed. It was found that predicted amino acids were closely related to other plant species and existed as an ortholog in Arabidopsis and rice.

Meanwhile, we investigated the expression of genes in *S. italica* in response to stress and normal growth conditions. We hypothesized that genes that show differential expression under various stress exposures are more likely to be involved in metal homeostasis. Most of these genes are differentially expressed due to downstream changes in the physiological status of plants as a result of changes in metal homeostasis, although a couple of genes are directly involved in regulating metal homeostasis. The expression patterns of SiZIP genes reflect their diverse functions during Zn and or Fe translocation. It has been reported that the ZIP genes display various expression profiles due to tissue specificity and in response to fluctuating environmental Zn and Fe conditions. For instance, OsZIP7a was induced in Fe-deficient root, while OsZIP8 was stimulated in Zn-deficient shoots and roots (Yang et al., 2007). Histochemical confinement analysis showed that the mRNA of OsZIP4 was more in the vascular bundles of leaves and roots and phloem cells of the stem and the meristems (Ishimaru et al., 2005). Hence, these results demonstrated that SiZIP genes encode Zn or Fe transporters and have various functions associated with uptake and translocation, detoxification and storage of Zn and or Fe in plant cells.

Homology modeling of SiZIP revealed that proteins with similar patterns are not equally expressed (Si036196m.g, Si024505m.g and Si010411m.g), whereas proteins with different patterns were found to be equally expressed, irrespective of tissue type (Si022298m.g, Si035517m.g, Si010244m.g and Si013901m.g), which could be due to effects of gene homologs present in the same species. Comparative expression analysis of SiZIP with little millet showed that Si036196 gene homolog can be further cloned for functional characterization. Sequencing nutritionally rich crops like little millet will provide a better platform to enhance the micronutrient content in other cereal crops.

SSR identification in SiZIPs revealed that SiZIP2-Intronic $(TC)_6$, SiZIP3-CDS $(CCA)_5$, SiZIP5-Intronic $(TC)_5$ and SiZIP7-Cds $(GC)_5$ have repeat motifs in their CDS and intronic regions Supplementary Table S3. These SSR repeats could be used for allele mining of the respective genes in Marker Assisted Selection (MAS) crop breeding programs.

Trafficking of the multi-pass PM proteins such as ZIP typically requires an N-terminal signal peptide (Blobel and Dobberstein, 1975; Martoglio and Dobberstein, 1998). A multi pass integral PM protein, *S. italica* ZIP family is predicted to comprise 5 to 9 TMs, with a putative N-terminal signal peptide of 1–30 amino acids with cleavage site ranging from 25 to 30 amino acids. Here we hypothesize that a maximum of ZIP genes in plants might transport and maintain homeostasis with the help of signal peptide present in their N-terminal end. Altogether, our study provides a computational framework for *in silico* characterization of any particular gene family, which can be utilized in MAS breeding and genetic engineering of field crops. Care must be taken before selecting a gene of interest in downstream analysis. Further the effect of homologs in the transgenic plants can be minimized by transfer of single copy orthologs.

CONCLUSION

Our results propose that SiZIP genes encode functional Zn and or Fe transporters that may regulate the uptake, translocation and storage of divalent metal ion in plant cells and mainly endosperm. Transcriptome analysis hints the sustained expression level of identified gene orthologs among spica, leaf, stem and root tissues. The present study provides new insights into the evolutionary relationship and putative functional divergence of the ZIP gene family during the growth and development of rice, Arabidopsis and foxtail millet. *In silico* functional characterization viz., expression analysis, transcription factor mining, homology modeling, phylogenetic analysis and wet lab validation of SiZIP family showed that Si02298m.g, Si03557m.g, Si010244m.g and Si013901m.g genes could serve as potential genes in Fe and Zn biofortification. Comparative expression analysis of SiZIP homolog in little millet showed the existence of potential ZIP genes for biofortification. Further characterization of identified orthologs in *S. italica* and their functional validation in a panel of genotypes under varying nutrient supplement will help in divulging new sources of nutritionally important genes for improvement of staple food crops.

AUTHOR CONTRIBUTIONS

GC and GA conceived and designed the experiment. GA wrote the manuscript, performed the *in silico* experiments and MD maintained plant materials and has done the nutritional analysis and wet lab validation. All the authors have revised the final draft of the manuscript.

ACKNOWLEDGMENTS

The budgetary backing for the author GA accorded by Department of Biotechnology, Government of India, New Delhi is thankfully acknowledged. The authors acknowledge the assistance from Mr. Muthamilarasan Mehanathan of Dr. Manoj Prasad's laboratory, National Institute of Plant Genome Research, New Delhi, India in analyzing RNA-seq derived expression patterns.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00775/full#supplementary-material

FIGURE S1 | Intron-exon structure of SiZIPs as predicted by GSDS.

FIGURE S2 | Homology modeling of 7 SiZIPs as predicted by SWISS-MODEL given in 3-D structure.

FIGURE S3 | Motif analysis has been done among the 12 single copy ortholog ZIP gene family members of Arabidopsis, rice and foxtail millet. The different color bars in one gene indicate different motifs.

FIGURE S4 | Membrane topology of ZIP proteins across rice, Arabidopsis and foxtail millet. Red color amino acids represent signal peptides of respective protein.

DATASHEET S1 | Multiple sequence alignment of predicted amino acid sequences of SiZIP, AtZIP and OsZIP proteins.

REFERENCES

Bailey, T. L., and Elkan, C. (1994). *Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers*. La Jolla, CA: Dept. of Computer Science and Engineering.

Barberon, M., Dubeaux, G., Kolb, C., Isono, E., Zelazny, E., and Vert, G. (2014). Polarization of IRON-REGULATED TRANSPORTER 1 (IRT1) to the plant-soil interface plays crucial role in metal homeostasis. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8293–8298. doi: 10.1073/pnas.1402262111

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* 53, 474–485. doi: 10.1002/dvg.22877

Blobel, G., and Dobberstein, B. (1975). Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* 67, 835–851. doi: 10.1083/jcb.67.3.835

Chandel, G., Banerjee, S., Vasconcelos, M., and Grusak, M. A. (2010). Characterization of the root transcriptome for iron and zinc homeostasis-related genes in indica rice (*Oryza sativa* L). *J. Plant Biochem. Biotechnol.* 19, 145–152. doi: 10.1007/bf03263334

Chow, C.-N., Zheng, H.-Q., Wu, N.-Y., Chien, C.-H., Huang, H.-D., Lee, T.-Y., et al. (2015). PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.* 44, D1154–D1160. doi: 10.1093/nar/gkv1035

Colangelo, E. P., and Guerinot, M. L. (2006). Put the metal to the petal: metal uptake and transport throughout plants. *Curr. Opin. Plant Biol.* 9, 322–330. doi: 10.1016/j.pbi.2006.03.015

Eide, D., Broderius, M., Fett, J., and Guerinot, M. L. (1996). A novel iron-regulated metal transporter from plants identified by functional expression in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 93, 5624–5628. doi: 10.1073/pnas.93.11.5624

Enright, A. J., Van Dongen, S., and Quzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.2307/2408678

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., et al. (2005). "Protein identification and analysis tools on the ExPASy server," in *Proteomics Protocols Handbook*, ed. J. M. Walker (New York, NY: Humana Press), 571–607. doi: 10.1385/1-59259-890-0:571

Grotz, N., and Guerinot, M. L. (2006). Molecular aspects of Cu, Fe and Zn homeostasis in plants. *Biochim. Biophys. Acta* 1763, 595–608. doi: 10.1016/j.bbamcr.2006.05.014

Guerinot, M. L. (2000). The ZIP family of metal transporters. *Biochim. Biophys. Acta* 1465, 190–198. doi: 10.1016/S0005-2736(00)00138-3

Haydon, M. J., and Cobbett, C. S. (2007). Transporters of ligands for essential metal ions in plants: research review. *New Phytol.* 174, 499–506. doi: 10.1111/j.1469-8137.2007.02051.x

Ishimaru, Y., Suzuki, M., Kobayashi, T., Takahashi, M., Nakanishi, H., Mori, S., et al. (2005). OsZIP4, a novel zinc-regulated zinc transporter in rice. *J. Exp. Bot.* 56, 3207–3214. doi: 10.1093/jxb/eri317

Kabir, A. H., Swaraz, A. M., and Stangoulis, J. (2014). Zinc-deficiency resistance and biofortification in plants. *J. Plant Nutr. Soil Sci.* 177, 311–319. doi: 10.1002/jpln.201300326

Kambe, T., Yamaguchi-Iwai, Y., Sasaki, R., and Nagao, M. (2004). Overview of mammalian zinc transporters. *Cell. Mol. Life Sci.* 61, 49–68. doi: 10.1007/s00018-003-3148-y

Kawahara, Y., Bastide, M. D. L., Hamilton, J. P., Kanamori, H., Mccombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-4

Lata, C., Mishra, A. K., Muthamilarasan, M., Bonthala, V. S., Khan, Y., and Prasad, M. (2014). Genome-wide investigation and expression profiling of AP2/ERF transcription factor superfamily in foxtail millet (*Setaria italica* L.). *PLoS ONE* 9:e113092. doi: 10.1371/journal.pone.0113092

Lee, S., Jeong, H. J., Kim, S. A., Lee, J., Guerinot, M. L., and An, G. (2010a). OsZIP5 is a plasma membrane zinc transporter in rice. *Plant Mol. Biol.* 73, 507–517. doi: 10.1007/s11103-010-9637-0

Lee, S., Kim, S. A., Lee, J., Guerinot, M. L., and An, G. (2010b). Zinc deficiency-inducible OsZIP8 encodes a plasma membrane-localized zinc transporter in rice. *Mol. Cells* 29, 551–558. doi: 10.1007/s10059-010-0069-0

Martoglio, B., and Dobberstein, B. (1998). Signal sequences: more than just greasy peptides. *Trends Cell Biol.* 8, 410–415. doi: 10.1016/s0962-8924(98)01360-9

Milner, M. J., Seamon, J., Craft, E., and Kochian, L. V. (2012). Transport properties of members of the ZIP family in plants and their role in Zn and Mn homeostasis. *J. Exp. Bot.* 64, 369–381. doi: 10.1093/jxb/ers315

Mishra, A. K., Muthamilarasan, M., Khan, Y., Parida, S. K., and Prasad, M. (2014). Genome-wide investigation and expression analyses of WD40 protein family in the model plant foxtail millet (*Setaria italica* L.). *PLoS ONE* 9:e86852. doi: 10.1371/journal.pone.0086852

Muthamilarasan, M., Bonthala, V. S., Khandelwal, R., Jaishankar, J., Shweta, S., Nawaz, K., et al. (2015a). Global analysis of WRKY transcription factor superfamily in Setaria identifies potential candidates involved in abiotic stress signaling. *Front. Plant Sci.* 6:910. doi: 10.3389/fpls.2015.00910

Muthamilarasan, M., Khan, Y., Jaishankar, J., Shweta, S., Lata, C., and Prasad, M. (2015b). Integrative analysis and expression profiling of secondary cell wall genes in C4 biofuel model Setaria italica reveals targets for lignocellulose bioengineering. *Front. Plant Sci.* 6:965. doi: 10.3389/fpls.2015.00965

Muthamilarasan, M., Dhaka, A., Yadav, R., and Prasad, M. (2016). Plant Science Exploration of millet models for developing nutrient rich graminaceous crops. *Plant Sci.* 242, 89–97. doi: 10.1016/j.plantsci.2015.08.023

Muthamilarasan, M., and Prasad, M. (2015). Advances in *Setaria* genomics for genetic improvement of cereals and bioenergy grasses. *Theor. Appl. Genet.* 128, 1–14. doi: 10.1007/s00122-014-2399-3

Pahlsson, A.-M. B. (1989). Toxicity of heavy metals (Zn, Cu, Cd, Pb) to vascular plants. *Water Air Soil Pollut.* 47, 287–319. doi: 10.1007/bf00279329

Palmer, N. A., Saathoff, A. J., Waters, B. M., Donze, T., Heng-Moss, T. M., Twigg, P., et al. (2014). Global changes in mineral transporters in tetraploid switchgrasses (*Panicum virgatum* L.). *Front. Plant Sci.* 4:549. doi: 10.3389/fpls.2013.00549

Palmgren, M. G., Clemens, S., Williams, L. E., Krämer, U., Borg, S., Schjørring, J. K., et al. (2008). Zinc biofortification of cereals: problems and solutions. *Trends Plant Sci.* 13, 464–473. doi: 10.1016/j.tplants.2008.06.005

Price, A. H., and Hendry, G. A. F. (1991). Iron-catalysed oxygen radical formation and its possible contribution to drought damage in nine native grasses and three cereals. *Plant Cell Environ.* 14, 477–484. doi: 10.1111/j.1365-3040.1991.tb01517.x

Qi, X., Xie, S., Liu, Y., Yi, F., and Yu, J. (2013). Genome-wide annotation of genes and noncoding RNAs of foxtail millet in response to simulated drought stress by deep sequencing. *Plant Mol. Biol.* 83, 459–473. doi: 10.1007/s11103-013-0104-6

Rhodes, D., and Klug, A. (1993). Zinc fingers. *Sci. Am.* 268, 56–65. doi: 10.1038/scientificamerican0293-56

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374–378.

Samira, R., Stallmann, A., Massenburg, L. N., and Long, T. A. (2013). Ironing out the issues: integrated approaches to understanding iron homeostasis in plants. *Plant Sci.* 210, 250–259. doi: 10.1016/j.plantsci.2013.06.004

Stangoulis, J., and Sison, C. (2008). *Crop Sampling Protocols for Micronutrient Analysis HarvestPlus Technical Monographs 7.* Washington, DC: International Food Policy Research Institute.

Taylor, K. M., Morgan, H. E., Johnson, A., and Nicholson, R. I. (2004). Structure-function analysis of HKE4, a member of the new LIV-1 subfamily of zinc transporters. *Biochem. J.* 377, 131–139. doi: 10.1042/bj20031183

Tello-Ruiz, M. K., Stein, J., Wei, S., Preece, J., Olson, A., Naithani, S., et al. (2016). Gramene 2016: comparative plant genomics and pathway resources. *Nucleic Acids Res.* 44, 1133–1140. doi: 10.1093/nar/gkv1179

Thakur, S., Singh, L., Wahid, Z. A., Siddiqui, M. F., Atnaw, S. M., and Din, M. F. (2016). Plant-driven removal of heavy metals from soil: uptake, translocation, tolerance mechanism, challenges, and future perspectives. *Environ. Monit. Assess.* 188, 206. doi: 10.1007/s10661-016-5211-9

Thompson, J., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Trifinopoulos, J., Nguyen, L.-T., Von Haeseler, A., and Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44, W232–W235. doi: 10.1093/nar/gkw256

Vallee, B. L., and Auld, D. S. (1990). Zinc coordination, function, and structure of zinc enzymes and other proteins. *Biochemistry* 29, 5647–5659. doi: 10.1021/bi00476a001

Vallee, B. L., and Falchuk, K. H. (1993). The biochemical basis of zinc physiology. *Physiol. Rev.* 73, 79–118.

Varotto, C., Maiwald, D., Pesaresi, P., Jahns, P., Salamini, F., and Leister, D. (2002). The metal ion transporter IRT1 is necessary for iron homeostasis and efficient photosynthesis in *Arabidopsis thaliana*. *Plant J.* 31, 589–599. doi: 10.1046/j.1365-313x.2002.01381.x

Vert, G., Barberon, M., Zelazny, E., Séguéla, M., Briat, J.-F., and Curie, C. (2009). *Arabidopsis* IRT2 cooperates with the high-affinity iron uptake system to maintain iron homeostasis in root epidermal cells. *Planta* 229, 1171–1179. doi: 10.1007/s00425-009-0904-8

Vert, G., Grotz, N., Dédaldéchamp, F., Gaymard, F., Guerinot, M. L., Briat, J. F., et al. (2002). IRT1, an Arabidopsis transporter essential for iron uptake from the soil and for plant growth. *Plant Cell* 14, 1223–1233. doi: 10.1105/tpc.001388

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, Y., Coleman-Derr, D., Chen, G., and Gu, Y. Q. (2015). OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 43, W78–W84. doi: 10.1093/nar/gkv487

Yadav, C. B., Bonthala, V. S., Muthamilarasan, M., Pandey, G., Khan, Y., and Prasad, M. (2014). Genome-wide development of transposable elements-based markers in foxtail millet and construction of an integrated database. *DNA Res.* 22, 79–90. doi: 10.1093/dnares/dsu039

Yang, X., Huang, J., Jiang, Y., and Zhang, H.-S. (2007). Cloning and functional identification of two members of the ZIP (Zrt, Irt-like protein) gene family in rice (*Oryza sativa* L.). *Mol. Biol. Rep.* 36, 281–287. doi: 10.1007/s11033-007-9177-0

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., et al. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* 30, 549–554. doi: 10.1038/nbt.2195

Zhang, P., Tan, S., Berry, J., Li, P., Ren, N., Li, S., et al. (2014). An uncleaved signal peptide directs the *Malus xiaojinensis* iron transporter protein Mx IRT1 into the ER for the PM secretory pathway. *Int. J. Mol. Sci.* 15, 20413–20433. doi: 10.3390/ijms151120413

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.

The reviewer VB and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Genetics

REVIEW
published: 07 July 2017
doi: 10.3389/fgene.2017.00094

Evaluation of Quality Assessment Protocols for High Throughput Genome Resequencing Data

*Matteo Chiara and Giulio Pavesi**

Dipartimento di Bioscienze, Università di Milano, Milan, Italy

Large-scale initiatives aiming to recover the complete sequence of thousands of human genomes are currently being undertaken worldwide, concurring to the generation of a comprehensive catalog of human genetic variation. The ultimate and most ambitious goal of human population scale genomics is the characterization of the so-called human "variome," through the identification of causal mutations or haplotypes. Several research institutions worldwide currently use genotyping assays based on Next-Generation Sequencing (NGS) for diagnostics and clinical screenings, and the widespread application of such technologies promises major revolutions in medical science. Bioinformatic analysis of human resequencing data is one of the main factors limiting the effectiveness and general applicability of NGS for clinical studies. The requirement for multiple tools, to be combined in dedicated protocols in order to accommodate different types of data (gene panels, exomes, or whole genomes) and the high variability of the data makes difficult the establishment of a ultimate strategy of general use. While there already exist several studies comparing sensitivity and accuracy of bioinformatic pipelines for the identification of single nucleotide variants from resequencing data, little is known about the impact of quality assessment and reads pre-processing strategies. In this work we discuss major strengths and limitations of the various genome resequencing protocols are currently used in molecular diagnostics and for the discovery of novel disease-causing mutations. By taking advantage of publicly available data we devise and suggest a series of best practices for the pre-processing of the data that consistently improve the outcome of genotyping with minimal impacts on computational costs.

Keywords: precision medicine, next-generation sequencing read quality, genome resequencing, whole exome sequencing, molecular diagnostics

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Zeeshan Ahmed,
University of Connecticut Health Center, United States
Zhaohui Steve Qin,
Emory University, United States

***Correspondence:**
Giulio Pavesi
giulio.pavesi@unimi.it

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

Received: *28 March 2017*
Accepted: *21 June 2017*
Published: *07 July 2017*

Citation:
Chiara M and Pavesi G (2017) Evaluation of Quality Assessment Protocols for High Throughput Genome Resequencing Data. Front. Genet. 8:94. doi: 10.3389/fgene.2017.00094

INTRODUCTION

The steady reduction in sequencing costs associated with the advent of the new generation of ultra-high throughput sequencing platforms, collectively known as Next-Generation Sequencing (NGS) technologies, is one of the major drivers of the so called "genomic revolution." Consequent to the development of these novel ultra efficient sequencing technologies [see (Goodwin et al., 2016) for a comprehensive review] the number of publicly available human genome and exome sequences is now in the hundreds of thousands, steadily increasing on a daily basis (Stephens et al., 2015). The characterization and fine scale annotation of the human *variome*, that is, the ensemble of genetic

variants in the human population, is one of the most ambitious goals of massive human genome sequencing projects. The possibility to link genetic variants and haplotypes with the corresponding phenotypes and discover causal relationships or calculate risk factors is instrumental for the development of more informed approaches to medical science, such as precision medicine (Lu et al., 2014) where patients can be treated based on their genetic background, or predictive medicine (Kotze et al., 2015) where risks factors for various diseases can be calculated beforehand and suitable measures can be instituted in order to prevent the disease or decrease its severity. Numerous countries worldwide are currently undertaking or are planning to launch large-scale projects aiming to sequence an increasing proportion of their population: by example England (UK10K Consortium et al., 2015) and Saudi Arabia (Alkuraya, 2014) have both announced 100.000 individuals sequencing projects and researchers from the United States[1] and China (Cyranoski, 2016) aim to sequence 1 million genomes in the next few years. In the meanwhile, various "pilot" (Gurdasani et al., 2015; Nagasaki et al., 2015; Sidore et al., 2015; The 1000 Genomes Project Consortium et al., 2015; Lek et al., 2016) projects, sequencing thousands of human genomes and exomes have been successfully undertaken which demonstrate the power of big data genomics for the identification of deleterious mutations and providing a substantial contribution to the understanding of the evolutionary processes that shape the genomes of modern human populations.

While the possibility to sequence an unprecedented number of individual genomes could serve as the basis for a new revolution in medical science and genetics, the need to handle, analyze and store huge amounts of data is posing major challenges to genomics and bioinformatics which at present remain largely unresolved. A possibly incomplete catalog of all known NGS sequencing platforms[2] provide evidence for the presence of almost 2200 instruments worldwide, distributed in 1027 sequencing facilities across 62 countries. A conservative estimate of sequencing capacity based on the manufacturer specifications of the various instruments suggests that, if used at full scale, NGS platforms could generate in the excess of 35 petabases of sequencing data per year. In a recent paper Stephens et al. (2015) suggest that, at the current rate, worldwide sequencing capacity could possibly reach zettabases of sequencing in the next 10 years, corresponding a number of complete human genomic sequences ranging from 100 million to 2 billion.

Bioinformatic analysis is currently one of the major bottlenecks in the processing of human resequencing data (Alyass et al., 2015). The need to integrate multiple tools into dedicated and sometimes complex analysis procedures requires a substantial amount of manual work and represents a major hindrance that limits the speed and general applicability of genotyping strategies. While best practices, procedures, and guidelines, defining the basic principles for the analysis and the annotation of the data have been introduced (Richards et al., 2015) and a large collection of bioinformatic tools for the identification of simple nucleotide variants (SNV) and/or small indels from NGS resequencing data is currently available (Pabinger et al., 2014), there exists no golden standard approach, and comparative studies evaluating different genotyping pipelines reached contrasting conclusions (Cornish and Guda, 2015; Hwang et al., 2015; Zhang et al., 2015). Consensus call-set based approaches, integrating predictions from multiple tools, can improve the accuracy and sensitivity of genotyping strategies (Bao et al., 2014). They, however, require additional computational costs, which might not always be justified by improvements in accuracy. Also, notwithstanding the high standardization of laboratory protocols and kits used in their production, the variability of NGS sequencing data remains high, and systematic biases resulting in the so called "batch effects" (Taub et al., 2010) can limit the extent of any bioinformatic approach, preventing the development of a conclusive strategy.

While several studies that compare the performances of various tools and pipelines are currently available [see (Pabinger et al., 2014), for a comprehensive review], at the time being we are not aware of dedicated studies evaluating the effects of quality assessment and reads pre-processing strategies in genotyping studies. Fine scale optimization of such procedures can on one hand improve the accuracy of the results, and on the other reduce significantly the computational requirements. This step is therefore essential as the starting point in the development of highly scalable workflows for the analysis of human resequencing data. In this article we discuss major strength and limitations of the different types of resequencing protocols that are currently used in molecular diagnostics and population scale genomics, and, by taking advantage of publicly available data, we devise guidelines and best practices for the pre-processing of the sequences.

RESEQUENCING STRATEGIES, APPLICATIONS AND LIMITATIONS

The ability to perform population scale studies of large genomes is highly interlinked with the advent of modern ultra-high throughput DNA sequencing technologies and the substantial reduction in sequencing costs. At present, Illumina accounts for the largest share of the sequencing market. The recent release of the ultra-high throughput Illumina X Ten sequencing system[3], which permits the sequencing of 1000s of human genomes per year for less than 1000 USD per genome, represents a major breakthrough in this field, and at the time being all the most ambitious population scale human genome resequencing projects are based on this technology. One of the major constraints of the second generation of ultra-high throughput sequencing technologies is the reduced size of the reads (a few hundreds bps), that poses limits to the possibility of reconstructing accurately long haplotypes and resolve repetitive and complex regions, which represent approximately 60% of the human genome (De Koning et al., 2011). Such limitations are being superseded by the development of a third generation of NGS sequencing platforms such as the PacBio (Eid et al., 2009) and Oxford Nanopore

[1] http://www.whitehouse.gov/precision-medicine

[2] http://omicsmaps.com/

[3] https://www.illumina.com/systems/hiseq-x-sequencing-system.html

(Clarke et al., 2009) sequencing systems, that can produce sequences of stretches of DNA ranging from a few to hundreds of kilobases in size. Such long reads can span complex or repetitive regions with a single continuous read, thus eliminating ambiguity in the positions or size of genomic elements. A recent resequencing of the human GRCh37 reference genome based on long-read sequencing technologies (Chaisson et al., 2015) recovered more than 1 Mb of novel sequence and identified more than 26,000 relatively long (≥50 bp) indels, providing one of the most comprehensive genome reference sequences available. Apart from simply improving reference genomes, long reads are more effective than short reads in the identification of clinically relevant structural variations and in the reconstruction of long haplotypes (Ammar et al., 2015).

The comparatively higher error rate (15% on the average) and the increased cost with respect to short read technologies are considerable disadvantages that limit the application of long-read sequencing in current large-scale genome resequencing studies. However, the continuous improvements in sequencing chemistries and base calling algorithms, and the development of novel sequencing platforms with reduced operating costs such as the PacBio Sequel or the Oxford Nanopore Promethion might open the possibility of applying long-read sequencing technologies to population scale resequencing projects in the near future. In this respect, the recent development of sequencing strategies for the generation of synthetic long reads relying on existing Illumina platforms, with error profiles and throughput similar to those of current Illumina devices, might represent a valid alternative to long-read sequencing technologies. Two different systems for generating synthetic long-reads are currently available: the Illumina synthetic long-read sequencing platform (McCoy et al., 2014) (formerly known as Moleculo) and the 10X Genomics emulsion-based system[4]. Both platforms rely on a similar strategy, where a specialized library preparation method based on extreme dilution and DNA barcoding is applied to a size selected DNA library in order to mimic single molecule sequencing. Input DNA is first sheared into kilobase-long fragments, which are then randomly distributed across a small number of containers. The contents of each container are then sheared further into shorter fragments and are assigned a unique barcode before being pooled together for sequencing. After sequencing, reads are demultiplexed using the barcodes. Each container may be assembled separately with a short-read assembler, which produces multiple kilobase-long sequences in each well; this approach is referred to as subassembly. Alternatively, the contents of each container may be sequenced at a relatively low coverage and resulting reads might be used to assist in tasks such as genome phasing and scaffolding. The main difference between the two approaches consists in that while the Illumina system is aimed at the precise reconstruction of each long DNA fragment and is designed for genome assembly, the 10X strategy does not attempt gapless, end-to-end coverage of single DNA fragments, and is generally used for haplotyping and scaffolding. Considerations on sequencing costs, however, suggest that, for the time being, complete genome assembly of complex genomes based on synthetic long-reads technologies remains unfeasible, even if using the most advanced sequencing machines.

Hybrid approaches based on the combination of long or synthetic long and short reads have proven themselves to be highly effective in the generation of high-quality assemblies of large and complex genomes at a relatively low cost, resulting in the detection of complex structural rearrangements and in the accurate reconstruction of haplotypes (Mostovoy et al., 2016; Collins et al., 2017; Weisenfeld et al., 2017). With the ongoing steady reduction in sequencing costs it is not unfeasible to imagine that strategies of this kind will end up to be applied also to large-scale sequencing studies, resulting in a more accurate reconstruction of individual genomes and extending our understanding of the human variome.

Large-scale genome resequencing studies are nowadays usually performed by two alternative approaches: Whole Genome Shotgun sequencing (WGS), that is, the sequencing of complete genomes, or targeted resequencing, where high-throughput sequencing is applied to a predefined subset of genomic loci, usually selected on the base of their annotation (i.e., exons) or their association with pathological conditions. WGS clearly offers a more comprehensive, virtually complete, catalog of the genetic variation of an individual and is not limited by prior knowledge of the sequence, permitting the reconstruction of complex genomic rearrangements and large insertions. On the other hand, targeted resequencing, by limiting the size of the genomic material used, makes possible the sequencing of several samples within a sequencing run, increasing both the breadth and the depth of a genomic study on the selected loci. Another considerable advantage of targeted resequencing is that newly identified variants are more easy to interpret and characterize, since target regions usually correspond to functionally annotated genomic loci. Considering that the majority of known disease causing mutations is found in protein coding genes, the wealth of data produced by WGS approaches might result excessive and sometimes even misleading for clinical and diagnostic applications. Recent studies (Belkadi et al., 2015), however, suggest that WGS resequencing data are in general of better quality than the targeted resequencing counterpart, resulting in a slightly improved power in the detection of novel mutations even within targeted regions. Moreover, targeted resequencing can interrogate only predefined regions of the genome, and is therefore clearly ineffectual in the detection of large chromosomal rearrangements and large structural variants.

Whole Exome Sequencing

The deep sequencing of all the exons of a genome, known as Whole Exome Sequencing (WES) (Ng et al., 2009), is probably the most popular and widely used targeted resequencing approach. As the name suggests, it is based on exome capture, that is, the construction of DNA libraries enriched for the exonic fraction of the genome. DNA samples are randomly fragmented and oligonucleotide probes (baits) are used to capture the target regions by DNA hybridization. The resulting DNA sample is

[4]https://community.10xgenomics.com/t5/10x-Blog/A-basic-introduction-to-linked-reads/ba-p/95

then subjected to library construction and sequencing. Whole-exome methods generally capture from 35 to 100 megabases of DNA target regions, depending on the reference annotation system used in the design of the probes and on the inclusion of 3′ or 5′ untranslated regions (UTRs) in the experimental design. Agilent, Nimblegen, Illumina are the main suppliers of exome-enrichment kits for exome capture. The most relevant differences between these technologies are in the choice of target regions, in bait lengths and density, in the molecules used for capture, and in the genome fragmentation method. At present NimbleGen offers the largest target region set, covering 96 Mb (64 Mb coding + 32 Mb UTR), compared to 75 Mb (50 Mb coding + 25 Mb UTR) of Agilent and 62 Mb (42 Mb coding + 20 Mb UTR) of Illumina. However, given the continuous improvements in sequencing throughput of NGS sequencing technologies, the range of genomic regions targeted by exome capture kits is constantly expanding, resulting in the inclusion in target regions of promoter regions and intron-exon junctions (Samuels et al., 2013).

Whole Exome Sequencing in Clinical Studies

Since the majority of known disease-causing mutations are found in protein coding genes, WES is becoming an increasingly attractive alternative to WGS for clinical applications. As an additional advantage, genotyping assays performed by exome sequencing have a narrow breadth if compared with WGS approaches, and thus require less computational resources for the analysis and the storage of the data. Moreover, novel genomic variants discovered by exome sequencing are restricted to functionally annotated genomic regions, thus enabling a rapid inference of potential functional effects. Finally, notwithstanding the additional costs required for the capture kits, exome sequencing remains more economic than WGS, making possible the sequencing of a higher number of samples with an increased depth of coverage. For all these reasons, despite the increasing number of completely sequenced human genomes, WES sequencing remains today the preferential strategy for large-scale sequencing studies and for clinical applications of genome sequencing, and indeed the majority of available human resequencing data is in the form of exomes (Lek et al., 2016). This is also reflected in primary repositories of human genetic variation data, as in the latest release of the dbSNP database (Sherry et al., 2001), where the number of SNPs falling into protein coding genes surpasses by far the number of those found in intergenic regions (dbSNP build 141).

The relatively heterogeneous profile of read coverage over target regions is one of the major bottlenecks that reduce the sensitivity and applicability of exome capture assays. Experimental biases resulting in the so called "batch effects" are generally introduced both during exome capture and in the library preparation steps (Chilamakuri et al., 2014; Shigemizu et al., 2015). Such biases are specific and intrinsic to the different capture kits and library preparation protocols, and therefore limit the possibility of comparing WES experiments performed by means of different capture kits or by different sequencing providers. Also, in a typical exome sequencing study, approximately 40–60% of the reads derive from genomic regions outside of the designed targets, resulting in a substantial reduction of the theoretical coverage. Exome capture efficiency is highly variable and influenced by multiple factors related both to the design of the capture kit (length of the probes, probes density, probes design) and to experimental conditions affecting the efficiency of DNA fragmentation and PCR amplification of the DNA library (García-García et al., 2016).

Another relevant bias introduced by the capture hybridization step in WES sequencing consists in the preferential capture of reference sequence alleles, which hinders the detection of alternate alleles at heterozygous polymorphic sites by shifting the allele distribution (Guo et al., 2013). Highly polymorphic and heterozygous genomic regions are thus captured at lower efficiency than highly conserved genomic intervals, resulting again in a systematic bias in the coverage profile. Moreover, all library preparation protocols for exome sequencing require PCR amplification, which tends to lower coverage in GC rich regions due to annealing during amplification (Aird et al., 2011). Fluctuations in the coverage profile have a deep impact on the sensitivity of WES, and in particular in the detection of heterozygous variants (Belkadi et al., 2015). It has been estimated that 15X mapped read depth of WGS samples would be sufficient to detect almost all homozygous SNPs and 33X for almost all heterozygous SNPs (Bentley et al., 2008). Depending on the capture kit, it has been shown that WES required 80X mean on-target depth to reach the common threshold of 10X per-site depth in 90% or more of all targeted regions (Clark et al., 2011), which represents the minimal requirement for clinical applications of the WES technology.

Gene Panels

Gene panels are another popular form of targeted resequencing which is often used in large diagnostic screenings. This approach leverages on prior knowledge about the association of a set of genomic loci with phenotypic traits of interest (typically a disease) in order to perform highly focused sequencing of a very specific portion of the genome (Katsanis and Katsanis, 2013). Loci of interest, ranging to a few kilobases to several megabases in size, are usually enriched either by DNA hybridization capture or by targeted amplification (amplicon sequencing), and then sequenced with high-throughput sequencing platforms. Enrichment systems based on PCR amplification require a very limited quantity of DNA for the construction of the sequencing library, thus enabling the analysis of relatively tiny tissue samples which are common in medical applications. The capture of target regions is highly specific and does not suffer from off-target DNA contamination, offering a substantially higher coverage. Differential PCR amplification of the target regions, however, can introduce relevant biases, resulting in a highly heterogeneous coverage profile (Samorodnitsky et al., 2015). Depending on PCR primers design and of DNA fragmentation accuracy, target regions might end up to be covered only by reads obtained from a single DNA strand, resulting in a considerably higher error rate due to the fact that second generation NGS technologies are affected by systematic context-specific sequencing errors (Schirmer et al., 2015).

Capture systems based on DNA hybridization show better coverage uniformity, higher sensitivity and better accuracy than amplicon-based methods. Moreover, since the size of target regions is not strictly limited by PCR primers, capture by hybridization can recover also relatively small (depending on the size of the baits) structural variants, which are systematically missed by amplicon sequencing techniques. The amount of off-target capture is comparable to WES, accounting for about 40–50% of the reads: this proportion can, however, vary greatly according to the design of the array, since low complexity and micro-satellite regions, which are often found in intronic sequences, can sensibly reduce the specificity of the capture. Capture hybridization systems are more expensive and require a considerably larger amount of DNA for library construction if compared with equivalent amplicon-based strategies, making them a less attractive option for high-throughput genotyping of large cohorts of samples.

Gene Panels in Clinical Studies

Gene panels are particularly suited for diagnostic screenings, since they provide a consistent reduction in costs and turnaround times and offer the possibility to customize the design of the panel in order to include complete genes or specific intronic sequences.

High-throughput sequencing of a limited number of carefully selected loci enables the characterization of wide cohorts of patients, virtually querying the presence of all known causal mutations and therefore providing an invaluable tool for diagnostics. Large-scale screenings based on carefully designed gene panels show a diagnostic power comparable, or even superior to that of WES, as the reduction in sequencing costs permits the sequencing of larger cohorts of patients (Saudi Mendeliome Group, 2015). This strategy, however, requires a substantial knowledge of the molecular basis of the condition/disease under study, and is not clearly applicable to the discovery of novel disease-causing mutations affecting genes not previously associated with the condition of interest.

Gene panels sequencing typically result in a very large of coverage of the target regions, exceeding 1000X in most cases. This coverage depth surpasses by far the minimum requirements for genotyping applications, and enables the reliable detection of somatic variants that might be present in a minority of the cellular population. The possibility to detect somatic variants in heterogeneous cellular populations is a very powerful tool for cancer genomics. Since carcinogenesis is an evolutionary process driven by natural selection, tumors of all types consist of cellular populations that are highly diverse at the genetic, epigenetic, and phenotypic levels. Tumor heterogeneity is a major cause of therapy failure and disease resistance, and is a subject of the utmost biological and clinical relevance. Ultra-high coverage targeted resequencing of panels of known tumor related genes thus enables the characterization of cancer cell populations and the detection of somatic cancer mutations, including those possibly linked with drug resistance (Gerlinger et al., 2012; Kim et al., 2015; Au et al., 2016; De Leng et al., 2016), a process that can be instrumental for the correct formulation of personalized anti-tumoral therapies. Repeated sequencing over time permits to monitor the evolution of the tumoral population in response to therapies, both for the evaluation of the efficacy of the therapy, by studying the prevalence of "founder mutations," and for the identification of possible new resistance inducing variants. Therapies can be thus adapted accordingly, maximizing their efficacy.

Whole Genome Sequencing

Whole genome shotgun sequencing (WGS) is rapidly becoming the method of choice for the study of human genetic variation at population scale level. Indeed, recent studies (Belkadi et al., 2015; Meienberg et al., 2016) suggest that, beside the capacity to interrogate a substantially larger fraction of the genome, WGS can offer major advantages and data of superior quality with respect to targeted resequencing approaches.

Whole genome shotgun-based strategies are not based on prior knowledge of the reference genome and can (in principle) address any type of complex genomic structural variant, including inversions, large insertions and deletions. The relevance of structural events of this type has been largely underestimated, since it is now clear that they contribute more than SNVs to the variability of individual genomes (Huddleston et al., 2016), where they can constitute up to 75% of the individual specific genomic material.

Whole genome shotgun sequencing libraries require a simpler and more streamlined preparation, where the most recent protocols do not require PCR amplification resulting in a substantially more homogeneous coverage profile (Meienberg et al., 2016). Target regions capture, that as previously discussed can introduce significant amounts of technical variability, is in turn not required by WGS. As a consequence, WGS data show consistently superior coverage uniformity with respect to WES, and a substantially lower average read depth is required to achieve the same breadth of coverage (Belkadi et al., 2015). These facts permit a more consistent identification of heterozygous mutations, and a more reliable discovery of copy number variants (Belkadi et al., 2015; Meienberg et al., 2016). More importantly, WGS does not suffer from reference bias capture, resulting in a more accurate calling of heterozygous variants. Finally, while all the commercially available exome capture kits are prone to systematic biases that are in large part platform specific (Chilamakuri et al., 2014; Shigemizu et al., 2015; García-García et al., 2016) and limit the possibility to compare data across different systems and kits, WGS data are to some extent more reproducible and comparable, facilitating the comparison of data produced by different sequencing facilities at different times.

Whole Genome Sequencing in Clinical Studies

The major factors limiting the adoption of WGS technologies in clinical practice are not only related to the increase in costs with respect to targeted resequencing, but also to the computational resources required for the bioinformatic analysis and interpretation of the data. A typical human genome contains approximately 4 million SNV or small indels (Eberle et al., 2016), the vast majority of which is confined within intergenic or un-annotated genomic regions. This poses a limit to the systematic functional classification of variants and a prompt identification of putative disease-causing mutations. The equivalent figure for

an exome is in the order of about 80,000 variants per individual. Importantly, all these variants fall by definition within or close to functional genomic regions, and their effects can be predicted on the base of existing genomic annotations. The significance of a large number of intergenic variants detected by WGS remains unclear in large-scale clinical set-ups, as more than 80% of currently known disease-causing mutations are found within protein coding genes. This figure could be, however, an over-estimate due to ascertainment bias, since the majority of large-scale human genome resequencing projects aimed at the detection of disease-causing mutations have been carried out by means of WES sequencing, and a significant proportion of the studies was focused on rare monogenic Mendelian diseases. In this respect data, produced by WGS offer a more granular representation of the genomic variability, facilitating a more accurate reconstruction of the haplotypes which can be instrumental for the detection of genomic loci associated with complex phenotypic traits, including diseases like atherosclerosis, diabetes, and hypertension. Population genetics studies can benefit greatly from the wealth of genetic markers recovered from WGS sequencing, resulting in a more precise reconstruction of the evolutionary history of closely related populations.

Finally, another important advantage of WGS strategies is that they are not limited by any particular genomic annotation, and as such will probably form a better legacy for future investigations including newly discovered functional genomic elements. Indeed, although the current annotation of the human genome can be considered to be of high-quality, such a possibility can not be excluded as demonstrated by the recent explosion of the number of long non-coding RNA genes (Chen et al., 2016).

COMPARATIVE EVALUATION OF READ PRE-PROCESSING STRATEGIES

Since the application of high-throughput sequencing technologies for the study of human genome variability at population scale level is becoming more and more commonplace, the development of standardized bioinformatics pipelines for an effective analysis the data is becoming crucial. Ideally, these pipelines should be fast, in order to cope with the increasing volumes of data, yet at the same time highly accurate as required by clinical applications. While implications of the usage of different combinations of tools for the alignment of short reads to the genome and for variant calling have been debated in depth (Pabinger et al., 2014), the evaluation of how quality assessment procedures can concur to the improvement of bioinformatic strategies for genotyping has been so far a little bit neglected. Indeed, good practices for quality assessment and pre-processing of the reads can contribute significantly to the optimization of downstream genotyping strategies, both by reducing computational requirements and by possibly lowering false positive rates. Three major approaches are commonly used for the pre-processing of reads obtained from large-scale resequencing studies: *quality trimming*, that is the polishing of the reads based on descriptive statistics calculated on their quality scores; *PCR de-duplication*, consisting in the elimination of identical reads or read pairs that might derive from PCR amplification of the same DNA fragment; *merging of overlapping pairs*, that consolidates pairs of reads originating from DNA fragments shorter than the combined length of the mates, into a longer, non-redundant sequence.

Materials and Methods

In order to explore the impact of reads pre-processing strategies on genotyping workflows and devise guidelines and suggestions for its optimization, we took advantage of a collection of publicly available genome and exome (Nextera kit) sequencing data derived form the platinum genome NA12878 (Eberle et al., 2016). Reference call-sets along with genome and exome sequencing data were retrieved from the Illumina BaseSpace Sequence Hub[5]. Reads were preprocessed by using nine different pipelines (summarized in **Table 1**), based on the combination of three progressive quality trimming stringency levels, and by adopting or discarding PCR de-duplication and read merging steps. Computations were performed on a Centos linux server with

[5]https://blog.basespace.illumina.com/category/datasets/

TABLE 1 | Read pre-processing strategies used in this study.

Pre-processing strategy*	Quality trimming**	Merging of overlapped pairs***	PCR de-duplication****
Lax	Lead:Q20, Trail:Q15, Wlen:10,Q15	No	No
Medium	Lead:Q25, Trail:Q20, Wlen:15,Q20	No	No
Hard	Lead:Q25, Trail:Q25, Wlen:20,Q25	No	No
Lax + Ovl	Lead:Q20, Trail:Q15, Wlen:10,Q15	Min Ovl 15 bp	No
Medium + Ovl	Lead:Q25, Trail:Q20, Wlen:15,Q20	Min Ovl 15 bp	No
Hard + Ovl	Lead:Q25, Trail:Q25, Wlen:20,Q25	Min Ovl 15 bp	No
Lax + PCR	Lead:Q20, Trail:Q15, Wlen:10,Q15	No	MDR = 0.03
Medium + PCR	Lead:Q25, Trail:Q20, Wlen:15,Q20	No	MDR = 0.03
Hard + PCR	Lead:Q25, Trail:Q25, Wlen:20,Q25	No	MDR = 0.03

From left to right, columns contain:
**Description of the strategy, as described in main text and figures.*
***Trimmomatic parameters for quality trimming. Q, quality score cut-off; Wlen, length of the window for sliding window operations.*
****Pear parameters for read merging. Min Ovl, minimum overlap required for merging of read pairs.*
*****MarkDuplicates parameters for PCR de-duplication. MDR (MAX DIFF RATE), overall mismatch rate for non-duplicated reads.*

64 Gb of RAM and 24 CPU cores, using a limit of 12 CPU cores and 32 Gb of RAM for each step of the pipelines.

Quality trimming was carried out using the Trimmomatic software (Bolger et al., 2014). Three different quality trimming procedures, with increasing levels of stringency, were used for the quality trimming of the raw reads. All the procedures were based on the same combination of Trimmomatic operations: "Leading" which removes nucleotides from the 5′ end of the reads if their quality score falls below a predefined cutoff, "Trailing" which performs the equivalent operations on the 3′ end of the reads, and "Slidingwindows," which evaluates the average quality score of the reads along sliding windows of fixed length, cutting the read if the average quality score within a window falls below a given threshold. Reads resulting in less than 50 bps after quality trimming were not incorporated in the subsequent stages of the analyses. Different levels of stringency were implemented with the following parameters:

- Lax: Leading Qs >= 20; Trailing Qs > 15; Slidingwindows, windows length 10, Qs > 15.
- Medium: Leading Qs >= 25; Trailing Qs > 20; Slidingwindows, windows length 15, Qs > 20.
- Hard: Leading Qs >= 25; Trailing Qs > 25; Slidingwindows, windows length 20, Qs > 25.

Merging of overlapping paired end reads was performed with the PEAR (Zhang et al., 2014) program, using default parameters. Removal of potential PCR duplicates from exome sequencing data was performed with the MarkDuplicate module of the Picard software (Wysoker et al., 2013) with default parameters. Genotyping was performed by using the GATK workflow (DePristo et al., 2011), Varscan2 (Koboldt et al., 2012), and Freebayes (Garrison and Marth, 2012). Only variants supported by at least two methods were included in the final call-sets. Intersections and comparisons of call-sets were performed by means of the vcf-tools merge utility (Danecek et al., 2011) and bedtools intersect program (Quinlan and Hall, 2010). Reads were mapped to the reference hg38 human assembly using Bowtie2 (Langmead et al., 2009), and resulting bam files were preprocessed following the GATK best practices recommendations. Different levels of coverage (20–90x) were simulated by sub-sampling the reads. Pipelines were evaluated both in terms of computational requirements, accuracy and specificity of the results, by comparing the respective call-sets with the golden standard sets of variants provided by Illumina. For exome data, only variants falling within the target regions were considered.

Impact of Reads Pre-processing Strategies on Variant Calling

The results are summarized in **Figure 1** (whole genome sequencing), **Figure 2** (WES) and detailed in Supplementary Tables S1–S5. Sensitivity and specificity of the call-sets obtained starting from the quality trimming strategies used in this study and described in the previous section are represented in **Figures 1A,B** for WGS and **Figures 2A,B** for WES datasets. Consistently with previous observations (Belkadi et al., 2015), WGS call-sets show a substantially higher sensitivity than WES, regardless of the (simulated) coverage level. This fact is probably due to a more uniform coverage profile (**Figures 1D**, **2D**), where we can observe a considerable reduction in coverage of GC rich regions in the WES data. Interestingly, regions with a high GC content are affected by a systematic reduction in coverage across all the simulated depths of sequencing, and even at 90x we observe that only about half (45%) of the regions with a GC composition greater than 60% reach the minimal coverage of 20x required for the confident identification of heterozygous variants. Notably, quality trimming seems to cause a substantial reduction of GC rich regions, and in particular hard filters cause a pronounced decrease in coverage in the most GC rich genomic regions, on both WES and WGS data sets.

Per base quality score distributions did not indicate any significant differences between the overall quality profile of the calls corresponding to each nucleotide (not shown). However, we can notice a considerable reduction in the average quality scores of GC rich reads (content in GC >= 60%), suggesting that the drop in sequencing quality is restricted to specific sequence contexts. This is again consistent with previous observations, as the accuracy of Illumina sequencing technology is known to deteriorate slightly in the presence of GC rich sequences (Ross et al., 2013).

The specificity of the call-sets is between 99 and 99.5% and nearly identical across all coverage levels (**Figures 2A,B**), suggesting that the genotyping strategy used in this study is robust and can produce consistent results.

On the other hand, higher coverage levels are associated with a steady increase in sensitivity both for the WES and WGS data, suggesting that an adequate coverage is a key factor for the correct identification of genetic variants. Consistently with this observation, call-sets based on lax quality trimming, which resulted in a moderated reduction of the nominal coverage (**Figures 1A**, **2A** and Supplementary Tables S4, S5), show a better sensitivity than equivalent sets where a more stringent quality trimming procedure was applied, and recover a higher proportion of "true" small indels and SNVs. More aggressive quality trimming can result in drastic reduction in coverage, with a systematic loss of accuracy in the most GC rich portions of the genome. This is particularly evident on the exome dataset, where the coverage is more skewed and influenced by GC composition. Interestingly, quality trimming resulted in a higher proportion of uniquely mapped reads, compared with the untrimmed data (Supplementary Tables S4, S5), suggesting that the removal of sequencing errors can improve the mappability of short reads.

Merging of overlapping reads pairs resulted in a small but general improvement of the sensitivity, leading to the identification of thousands of additional true variants (average 0.34% corresponding to 14,086 variants). In our experimental setup the application of this procedure yielded also a small (from 0.3 to 0.6%) but consistent increase in the number of reads mapping to the reference genome. Importantly, the majority (72.81%) of such reads were mapped to scarcely covered regions (average coverage 7.8x), which explains the increased sensitivity observed. Also, a significant proportion of the additional variants (52%) recovered by pipelines with overlapping read merging

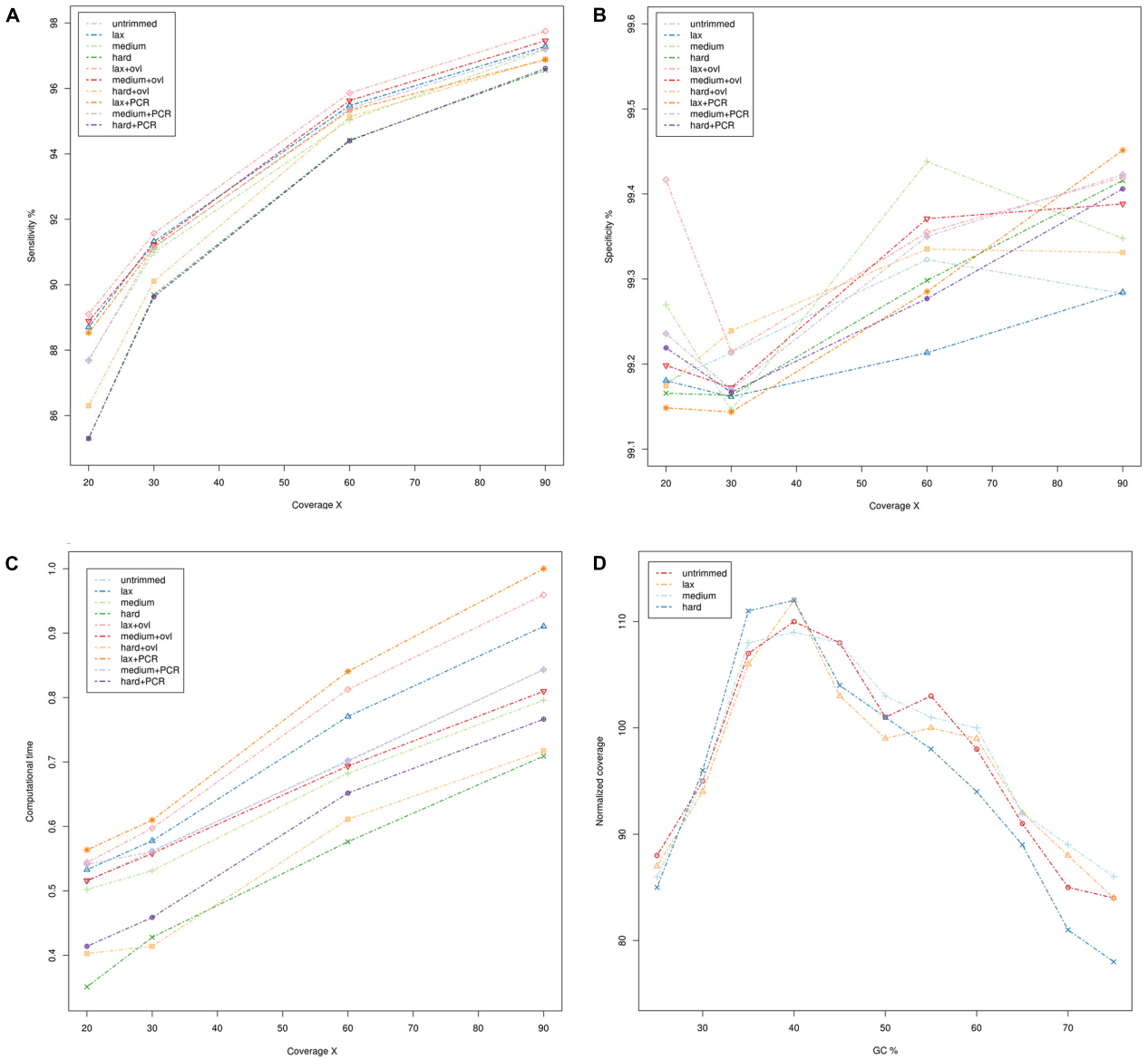


FIGURE 1 | Comparison of quality assessment pipelines on Whole Genome Sequencing data. **(A)** Sensitivity of quality assessment pipelines at different levels of coverage. **(B)** Specificity of quality assessment pipelines at different levels of coverage. **(C)** Computation time required by the pipelines. Times reported have been scaled with unity-based normalization. **(D)** Normalized coverage levels with respect to GC composition. Coverage levels and GC composition were calculated on genomic windows of 200 bp, overlapping by 100 bp. Coverage levels were normalized using the upper-quartile normalization and scaled to 100 (where 100 represents the expected coverage).

is represented by small indels, and occurs in highly variable genomic regions containing a complex combination of relatively short variants (28%). This suggests that the increased length of the merged reads can improve the alignment of such reads over highly polymorphic regions and facilitate the detection of complex events.

The usage of PCR de-duplication procedures seems to have, if any, only negative effects in all datasets analyzed in the present study: call-sets derived from pipelines where this step was applied show a marginal reduction in sensitivity at the cost of a general increase in computational times. Notably, we observe that even if all WGS resequencing data used in this study were produced by the means of a PCR free protocol, potential PCR duplicate reads are still identified even on these datasets, suggesting that the PCR de-duplication algorithm used in this study might be too stringent in the detection of potentially duplicated reads.

All the pipelines required comparable computational times (**Figures 1C**, **2C** and Supplementary Tables S1, S2), and the additional computational overheads needed to perform the various pre-processing steps did not result in any relevant increase in computational resources. PCR de-duplication and merging of overlapping reads were the most demanding steps in

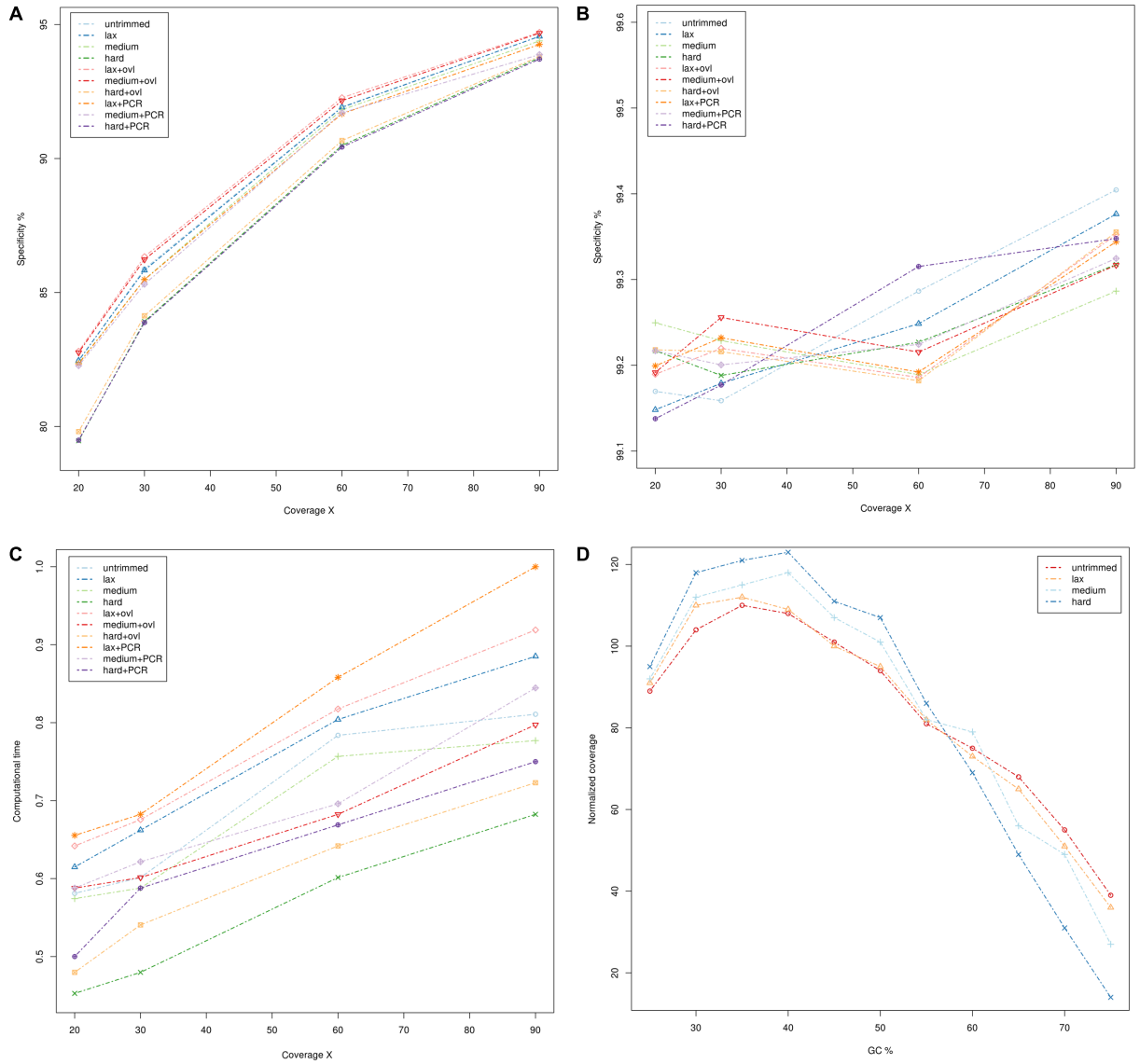


FIGURE 2 | Comparison of quality assessment pipelines on Whole Exome Sequencing data. **(A)** Sensitivity of quality assessment pipelines at different levels of coverage. **(B)** Specificity of quality assessment pipelines at different levels of coverage. **(C)** Computation time required by the pipelines. Times reported have been scaled with unity-based normalization. **(D)** Normalized coverage levels with respect to GC composition. Coverage levels and GC composition were calculated on genomic windows of 200 bp, overlapping by 100 bp. Coverage levels were normalized using the upper-quartile normalization and scaled to 100 (where 100 represents the expected coverage).

this respect, yielding an average increase of the computational time by 28 and 31%, respectively. On the other hand, very stringent quality trimming resulted in a consistent reduction of computational requirements, however, at the price of a considerable drop in sensitivity.

Although we do not observe marked differences between the nine pipelines tested in this study, the pipeline based on permissive quality trimming of the reads coupled with merging of overlapping read pairs achieved a slightly improved sensitivity over the others, resulting in the identification of about 8,000 unique true variants that were otherwise missed. The pipeline based on medium stringency quality trimming and on merging of overlapping reads came as a close second, consistent with the idea that merging of overlapping reads can contribute to increase systematically the sensitivity of genotyping procedures, yet by a small margin.

CONCLUSION

While population scale genome sequencing projects promise major revolutions in medical science, the need to efficiently analyze, handle and store such an unprecedented stream of data poses some major challenges which are still mostly unresolved.

Indeed, bioinformatic analysis of NGS resequencing data is currently one of the major bottlenecks, limiting the development of swift and effective diagnostic tools based on of large-scale sequencing.

The analysis usually requires elaborate pipelines, where multiple tools need to be properly combined with the right parameters in order to obtain reliable results. Human genome resequencing data can be highly heterogeneous due to inherent biases introduced by different library preparation protocols, sequencing platforms and experimental strategies. As a consequence, the optimization of bioinformatic pipelines for the analysis of NGS resequencing data is highly desirable, as it can contribute to the reduction of the impact of systematic biases, and simultaneously maximize the outcome of the experiments. While several studies evaluated the performances of different variant calling pipelines, as of today we are not aware of dedicated studies performing a systematic evaluation of read quality pre-processing procedures. Good practices for quality assessment and pre-processing of the reads can contribute significantly to the optimization of genotyping strategies, both by reducing computational requirements and by lowering false positive rates. In this article we discussed advantages and limitations of state of the art genome resequencing techniques, and by taking advantage of publicly available data we devised a series of suggestions and good practices for the pre-processing of sequencing reads that can improve systematically the efficacy of bioinformatic genotyping strategies. These suggestions are of general applicability, have a minimal impact on computational resources, and therefore they can be instrumental for the future design of highly scalable and efficient genotyping systems.

By comparing the results achieved by nine different pre-processing pipelines on a golden standard reference genome for which more than 4 millions highly accurate SNVs are available (Eberle et al., 2016), we evaluated the effects of common quality assessment procedures on genotyping, both in terms of accuracy of the resulting variant call-sets and in terms of the required computational resources. Pipelines were evaluated by simulating various levels of coverage depth, from shallow to deep. Unsurprisingly, we observed that the sensitivity of the genotyping assays increased with the depth of sequencing levels, suggesting that an adequate coverage is the key for the identification of genetic variants.

The genotyping workflows tested in this study, which are based on a combination of three popular variant calling algorithms, achieved a steady level of accuracy under all the scenarios herein tested, allowing an unbiased comparison of their sensitivity. Notably, quality trimming procedures that were applied using different level of stringency did not have any major impact on the accuracy of the call-sets, suggesting that variant calling algorithms are generally robust to sequencing errors. However, we also observed that after quality trimming a higher proportion of reads could be mapped unambiguously on the reference genome, supporting the idea that the removal of sequencing errors can facilitate reads mapping. Highly stringent quality trimming filters, discarding a significant proportion of the reads (average 36% of the reads, 41% of the total amount of sequence), resulted in a substantial reduction in coverage and as a consequence in a permanent deterioration in sensitivity. GC rich regions, where the composition in GC exceeded 55–60%, were more affected by stringent quality trimming, resulting in a systematic loss of coverage. This trend was particularly evident in WES data, where compositional biases in the coverage profile with a reduced coverage of GC rich regions are commonly introduced by PCR amplification (Ross et al., 2013).

Removal of reads potentially deriving from PCR duplication artifacts did not have any significant impact on the results for the datasets analyzed in this study, yielding only a modest reduction in sensitivity (due to loss of coverage) at the cost of a consistent increase in computational resources. This is in accordance to previous reports showing that PCR-deduplication has little impact on the overall accuracy of genotyping assays (Ebbert et al., 2016). This is, however, also probably due to the high-quality of the sequencing libraries used in the course of studies of this kind (including the current), and to the presence of a limited number of duplicated reads. PCR de-duplication is an important quality assessment step, which can be used to assess systematically the overall quality of a sequencing experiment. For this reason we do not advise to remove PCR de-duplication from bioinformatic workflows for quality assessment of NGS reads. On the other hand, we noticed that for DNA libraries with low PCR duplication levels such process can be detrimental to variant identification. In such cases is probably better to avoid PCR de-duplication at all, and perform variant calling directly from non-de-duplicated bam files.

Datasets where merging of overlapping reads pairs was performed resulted in small but steady increased sensitivity level, facilitating the identification of genetic variants falling in highly polymorphic genomic regions. Longer reads produced by this process were preferentially mapped to genomic regions that were scarcely covered by shorter un-merged reads, resulting in an increase of coverage in highly heterogeneous regions of the genome. The increase in computational resources required by this procedure is on the other hand moderate, and is fully justified in the light of the improvements in sensitivity, yielding the discovery of thousands of otherwise missed "true" genetic variants. Interestingly, we noticed that stringent quality trimming filters, by shortening the reads, can lead to a considerable reduction in the number of pairs of overlapping reads that can be merged with confidence (lax trimming 10.2%; hard trimming 5.8%). This indicates that merging of overlapping reads should be preferentially performed before the application of quality filters. In such a scenario, a Smith-Waterman alignment between the 3′ ends (where sequencing errors are more frequent) and the 5′ ends of the R2 reads (which are generally of higher quality) can be used as an effective error correction procedure.

In conclusion, our experiments suggest that quality assessment procedures can have a considerable impact on the accuracy and sensitivity of human genome genotyping based on NGS sequencing. Variant calling algorithms are generally robust to sequencing error and a high level of accuracy can be achieved when the prediction of multiple tools are combined. Coverage levels seem to be the most important factor affecting the sensitivity of this type of genotyping assays. In the light of

these considerations, quality assessment procedures based on relaxed quality trimming of the reads combined with merging of overlapping reads pairs seems ideal, as it can contribute a systematic improvement of the coverage of specific genomic regions, resulting in the identification of an increased number of true variants in highly polymorphic genomic contexts.

AUTHOR CONTRIBUTIONS

MC devised the study, performed bioinformatic analyses and wrote the manuscript. GP devised the study and wrote the manuscript.

FUNDING

This work was funded by the Italian Ministry of health, Bando di ricerca finalizzata e giovani ricercatori GR-2011-02347129, "Next-generation sequencing to study the penetrance of dominantly inherited porphyrias."

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00094/full#supplementary-material

REFERENCES

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18.

Alkuraya, F. S. (2014). Genetics and genomic medicine in Saudi Arabia. *Mol. Genet. Genomic Med.* 2, 369–378. doi: 10.1002/mgg3.97

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med. Genomics* 8, 33. doi: 10.1186/s12920-015-0108-y

Ammar, R., Paton, T. A., Torti, D., Shlien, A., and Bader, G. D. (2015). Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research.* 4, 17. doi: 10.12688/f1000research.6037.1

Au, C. H., Wa, A., Ho, D. N., Chan, T. L., and Ma, E. S. (2016). Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. *Diagn. Pathol.* 11, 11. doi: 10.1186/s13000-016-0456-8

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. A., Jiang, H., et al. (2014). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.* 13(Suppl. 2), 67–82. doi: 10.4137/CIN.S13779

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., et al. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U.S.A.* 28, 5473–5478. doi: 10.1073/pnas.1418631112

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611. doi: 10.1038/nature13907

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2016). Long non-coding RNAs and complex diseases:from experimental results to computational models. *Brief Bioinform.* doi: 10.1093/bib/bbw060 [Epub ahead of print].

Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., et al. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 15:449. doi: 10.1186/1471-2164-15-449

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., et al. (2011). Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* 29, 908–914. doi: 10.1038/nbt.1975

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4, 265–270. doi: 10.1038/nnano.2009.12

Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., et al. (2017). Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* 18, 36. doi: 10.1186/s13059-017-1158-6

Cornish, A., and Guda, C. (2015). A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Res. Int.* 2015, 11. doi: 10.1155/2015/456479

Cyranoski, D. (2016). China embraces precision medicine on a massive scale. *Nature* 7, 9–10. doi: 10.1038/529009a

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

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

De Leng, W. W. J., Gadellaa-van Hooijdonk, C. G., Barendregt-Smouter, F. A. S., Koudijs, M. J., Nijman, I., Hinrichs, J. W., et al. (2016). Targeted next generation sequencing as a reliable diagnostic assay for the detection of somatic mutations in tumours using minimal DNA amounts from formalin fixed paraffin embedded material. *PLoS ONE* 11:e0149405. doi: 10.1371/journal.pone.0149405

DePristo, M., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., et al. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* 17(Suppl. 7), 239. doi: 10.1186/s12859-016-1097-3

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., et al. (2016). A reference data set of 5.4 million human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* 27, 157–164. doi: 10.1101/gr.210500.116

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 2, 133–138. doi: 10.1126/science.1162986

García-García, G., Baux, D., Faugère, V., Moclyn, M., Koenig, M., Claustres, M., et al. (2016). Assessment of the latest NGS enrichment capture methods in clinical context. *Sci. Rep.* 6:20948. doi: 10.1038/srep20948

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi: 10.1056/NEJMoa1113205

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Guo, Y., Samuels, D. C., Li, J., Clark, T., Li, C. I., and Shyr, Y. (2013). Evaluation of allele frequency estimation using pooled sequencing data simulation. *Sci. World J.* 2013, 895496. doi: 10.1155/2013/895496

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The african genome variation project shapes medical genetics in Africa. *Nature* 517, 327–332. doi: 10.1038/nature13997

Huddleston, J., Chaisson, M. J., Meltz Steinberg, K., Warren, W., Hoekzema, K., Gordon, D., et al. (2016). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685. doi: 10.1101/gr.214007.116

Hwang, S., Kim, E., Lee, I., and Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* 5:17875. doi: 10.1038/srep17875

Katsanis, S. H., and Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* 14, 415–426. doi: 10.1038/nrg3493

Kim, S. T., Lee, W. S., Lanman, R. B., Mortimer, S., Zill, O. A., Kim, K. M., et al. (2015). Prospective blinded study of somatic mutation detection in cell-free DNA utilizing a targeted 54-gene next generation sequencing panel in metastatic solid tumor patients. *Oncotarget* 6, 40360–40369. doi: 10.18632/oncotarget.5465

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Kotze, M. J., Lückhoff, H. K., Peeters, A. V., Baatjes, K., Schoeman, M., van der Merwe, L., et al. (2015). Genomic medicine and risk prediction across the disease spectrum. *Crit. Rev. Clin. Lab. Sci.* 52, 120–137. doi: 10.3109/10408363.2014.997930

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

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 18, 285–291. doi: 10.1038/nature19057

Lu, Y. F., Goldstein, D. B., Angrist, M., and Cavalleri, G. (2014). Personalized medicine and human genetic diversity. *Cold Spring Harb. Perspect. Med.* 4, a008581. doi: 10.1101/cshperspect.a008581

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., et al. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS ONE* 9:e106689. doi: 10.1371/journal.pone.0106689

Meienberg, J., Bruggmann, R., Oexle, K., and Matyas, G. (2016). Clinical sequencing: is WGS the better WES? *Hum. Genet.* 135, 359–362. doi: 10.1007/s00439-015-1631-9

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., et al. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* 13, 587–590. doi: 10.1038/nmeth.3865

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., et al. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 japanese individuals. *Nat. Commun.* 6, 8018. doi: 10.1038/ncomms9018

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of twelve human exomes. *Nature* 461, 272–276. doi: 10.1038/nature08250

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., et al. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* 15, 256–278. doi: 10.1093/bib/bbs086

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., et al. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.* 14:R51. doi: 10.1186/gb-2013-14-5-r51

Samorodnitsky, E., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., Lyon, E., et al. (2015). Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum. Mutat.* 36, 903–914. doi: 10.1002/humu.22825

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., et al. (2013). Finding the lost treasures in exome sequencing data. *Trends Genet.* 29, 593–599. doi: 10.1016/j.tig.2013.07.006

Saudi Mendeliome Group (2015). Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biol.* 16, 134. doi: 10.1186/s13059-015-0693-2

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43, e37. doi: 10.1093/nar/gku1341

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311.

Shigemizu, D., Momozawa, Y., Abe, T., Morizono, T., Boroevich, K. A., Takata, S., et al. (2015). Performance comparison of four commercial human whole-exome capture platforms. *Sci. Rep.* 5:12742. doi: 10.1038/srep12742

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., et al. (2015). Genome sequencing elucidates sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* 47, 1272–1281.111. doi: 10.1038/ng.3368

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pbio.1002195

Taub, M. A., Corrada Bravo, H., and Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.* 2, 87. doi: 10.1186/gm208

The 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., and Memari, Y. (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526, 82–90. doi: 10.1038/nature14962

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Res.* 27, 757–767. doi: 10.1101/gr.214874.116

Wysoker, A., Tibbetts, K., and Fennell, T. (2013). *Picard Tools Version 1.90*. Available at: http://picard.sourceforge.net

Zhang, G., Wang, J., Yang, J., Li, W., Deng, Y., Li, J., et al. (2015). Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. *BMC Genomics* 16, 581. doi: 10.1186/s12864-015-1796-6

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

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 Genetics

REVIEW
published: 14 June 2017
doi: 10.3389/fgene.2017.00083

Computational Methods for Characterizing Cancer Mutational Heterogeneity

Fabio Vandin *

Department of Information Engineering, University of Padova, Padova, Italy

Advances in DNA sequencing technologies have allowed the characterization of somatic mutations in a large number of cancer genomes at an unprecedented level of detail, revealing the extreme genetic heterogeneity of cancer at two different levels: inter-tumor, with different patients of the same cancer type presenting different collections of somatic mutations, and intra-tumor, with different clones coexisting within the same tumor. Both inter-tumor and intra-tumor heterogeneity have crucial implications for clinical practices. Here, we review computational methods that use somatic alterations measured through next-generation DNA sequencing technologies for characterizing tumor heterogeneity and its association with clinical variables. We first review computational methods for studying inter-tumor heterogeneity, focusing on methods that attempt to summarize cancer heterogeneity by discovering pathways that are commonly mutated across different patients of the same cancer type. We then review computational methods for characterizing intra-tumor heterogeneity using information from bulk sequencing data or from single cell sequencing data. Finally, we present some of the recent computational methodologies that have been proposed to identify and assess the association between inter- or intra-tumor heterogeneity with clinical variables.

Keywords: cancer heterogeneity, mutations, cancer pathways, mutual exclusivity, clinical association

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Luciano Cascione,
Institute of Oncology Research, Switzerland
Matteo D'Antonio,
University of California, San Diego, United States
Faraz Hach,
University of British Columbia, Canada

***Correspondence:**
Fabio Vandin
fabio.vandin@unipd.it

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

Received: *07 February 2017*
Accepted: *30 May 2017*
Published: *14 June 2017*

Citation:
Vandin F (2017) Computational Methods for Characterizing Cancer Mutational Heterogeneity. Front. Genet. 8:83. doi: 10.3389/fgene.2017.00083

1. INTRODUCTION

Somatic mutations, alterations of the DNA which accumulate during the lifetime of an individual, are the most common cause of cancer. High-throughput sequencing technologies now allow to identify and catalog the entire complement of somatic mutations in a tumor (Mardis, 2008; Meyerson et al., 2010) and many studies, including the ones from TCGA[1] and ICGC [2], have used these technologies to measure mutations in the whole exome or whole genome of hundreds or thousands of tumors (e.g., see The Cancer Genome Atlas Research Network, 2017a,b for recent studies). These studies provide a detailed characterization of the landscape of somatic mutations in cancer, describing the hundreds-thousands of somatic mutations appearing in each tumor. Such somatic mutations include *single nucleotide variants* (SNVs) as well as *copy number aberrations* (CNAs), larger scale events which modify (by amplifications or deletions) the number of copies of a DNA region. Only a handful of all somatic mutations, called *driver* mutations, confer selecting advantage to cancer cells, while most somatic mutations are *passenger* mutations not contributing to the disease (Garraway and Lander, 2013; Vogelstein et al., 2013).

[1]https://cancergenome.nih.gov
[2]http://icgc.org

One of the most striking features of cancer mutational landscape is its *inter-tumor heterogeneity* (**Figure 1**): no two cancer genomes bear the same collection of somatic mutations, with many pairs of tumors having no mutation in common (Stratton et al., 2009), and a limited number of mutations appear in a large fraction of tumors, with most genes being mutated (by SNVs or CNAs) in $< 5\%$ of all patients with a given cancer type (Ciriello et al., 2013; Kandoth et al., 2013; Tamborero et al., 2013). Inter-tumor heterogeneity hinders efforts to identify *driver genes*, bearing driver mutations, by detecting frequently mutated genes, i.e., genes mutated in a significantly high fraction of patients (Dees et al., 2012; Lawrence et al., 2013). In addition, frequency-based methods may result in several false positives (D'Antonio and Ciccarelli, 2013) since genomic features not related to the disease, including (normal) gene expression levels and replication time (Lawrence et al., 2013), can nonetheless lead to a high mutation frequency for a gene and must therefore be taken into account to identify significantly mutated genes (Lawrence et al., 2014).

One of the causes of inter-tumor heterogeneity is the fact that driver mutations target signaling and molecular *pathways* (Vogelstein and Kinzler, 2004; Vogelstein et al., 2013), groups of interacting proteins and genes performing specific functions in a cell. Mutations in genes belonging to cancer pathways lead to the acquisition of the biological capabilities (e.g., resisting cell death and inducing angiogenesis) or *hallmarks* (Hanahan and Weinberg, 2000, 2011) featured by cancer cells. A cancer pathway may be altered by mutations in any of its genes, leading to a wide spectrum of mutation frequencies for genes in the same cancer pathway, with one or few genes mutated with relatively high frequency and many genes mutated at much smaller frequency, which may not be sufficient for detection by frequency-based methods. In addition, each cancer genome is exposed to different mutational processes characterized by different combinations of mutations or *signatures* (Alexandrov et al., 2013b; Petljak and Alexandrov, 2016), with different cancer types presenting different mixtures of such signatures (Nik-Zainal et al., 2012a, 2016; Alexandrov et al., 2013a, 2015, 2016). Studying and characterizing mutations at the level of pathways is therefore crucial to deal with heterogeneity for the identification of driver mutations and to identify common themes extending the "rulebook" of cancer (McGranahan and Swanton, 2015), with important implications in prognosis and therapy (Swanton, 2016).

In addition to uncover such *inter-tumor heterogeneity*, cancer genome sequencing has also uncovered *intra-tumor heterogeneity* (**Figure 2**): a tumor is often composed by different populations of cancer cells (Anderson et al., 2011; Gerlinger et al., 2012, 2014; Schuh et al., 2012; Newburger et al., 2013; Bolli et al., 2014; Brastianos et al., 2015; Gundem et al., 2015; Ling et al., 2015; Sottoriva et al., 2015), called *clones*, arising from the evolutionary process (Nowell, 1976) which starting from a normal cell leads, through somatic mutations, to a collection of related but different cancer cells (Greaves and Maley, 2012; Swanton, 2012). While only providing measurements at the level of the entire cell population, deep (e.g., >100-fold) bulk sequencing offers the opportunity to study intra-tumor heterogeneity: the variant allele frequency (VAF), or fraction of reads supporting a variant among all the reads mapped to the same genomic location, of a heterozygous variant in a diploid region is proportional to the fraction of cells with the variant among all cells in the sample. VAFs from a tumor can then be used to identify the various clones present in a tumor. In addition, since the VAFs in a cell are constrained by evolutionary relationships among the clones in a tumor, they can be used to infer the evolutionary

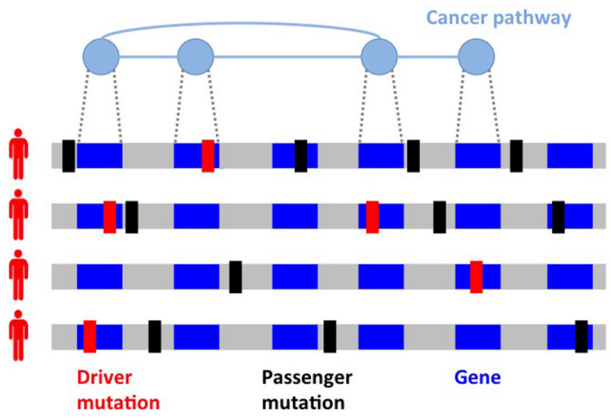


FIGURE 1 | Inter-tumor heterogeneity and its causes. Driver mutations (in red) target genes which are members of different cancer *pathways*, sets of interacting genes which perform specific functions and are altered in cancer. Passenger mutations (in black) not related to the disease comprise the majority of mutations in a tumor. Different mutated genes in cancer pathways and different passenger mutations are observed in tumors of the same type, with two cancer genomes often having no mutation in common.

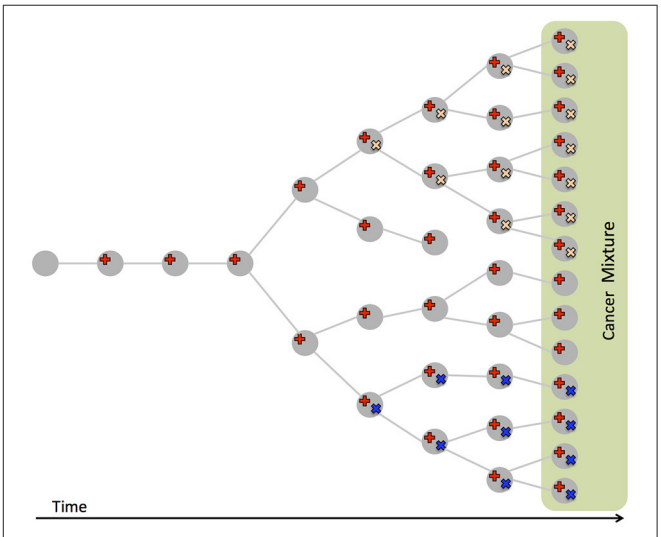


FIGURE 2 | Intra-tumor heterogeneity and its causes. Cancer evolves from a normal cell that accumulates mutations (in red, yellow, and blue), leading to different *clones*, populations of cells of different genotypes, coexisting in the same tumor. Bulk sequencing measures mutations from a sample of the resulting cell mixture, that also comprises normal cells. The fraction of reads supporting a mutation (VAF) is proportional to the number of cells with the mutation in the sample.

trajectory followed by the observed tumor (Ding et al., 2012; Nik-Zainal et al., 2012b; Yates and Campbell, 2012; Burrell et al., 2013). Understanding the clonal composition of a tumor is crucial for prognosis and therapy (Greaves and Maley, 2012; Swanton, 2012), since different clones may present drug resistant mutations (Greaves and Maley, 2012; Swanton, 2012) and the reliable characterization of the evolutionary history of a tumor is needed to predict the development of the disease (Yachida et al., 2010; Lipinski et al., 2016).

A more direct approach to study intra-tumor heterogeneity is single-cell sequencing (Hou et al., 2012; Xu et al., 2012; Wang et al., 2014; Navin, 2015a,b). Single-cell sequencing allows the direct observation of the cooccurrence of mutations within cells from different clones. However, single-cell data is currently noisy, with high false-positive and false-negative rates for mutation calls, and the number of cells that can be assessed is still limited compared to the billions of cells in a tumor.

In this review we describe bioinformatic and computational approaches to characterize cancer heterogeneity from next-generation sequencing data. We consider methods to deal with three aspects of cancer heterogeneity, after alterations such as SVNs and/or CNAs have been identified in a tumor or in multiple tumors (Raphael et al., 2014). First, we consider methods that tackle inter-tumor heterogeneity by characterizing tumor mutations at the pathway level. Second, we describe methods to characterize intra-tumor heterogeneity by using mutations from bulk sequencing or single-cell sequencing. Third, we describe some methods to relate cancer heterogeneity with clinical variables. The computational characterization of cancer heterogeneity is a topic which has spurred a lot of work in recent years and we only cover some of the tools that have been recently proposed. In particular, we only focus on methods assuming that somatic variants have already been called using one of the many methods currently available (e.g., Lawrence et al., 2013), and we refer the reader to other reviews discussing methods for variant calling in cancer (e.g., Raphael et al., 2014). The methods discussed in this review are mostly complementary, describing different characteristics of inter- or intra-tumor heterogeneity, which we believe constitute useful, multi-faceted information for cancer researchers and practitioners.

2. METHODS FOR INTER-TUMOR HETEROGENEITY

Several methods have been designed to characterize inter-tumor heterogeneity by identifying pathways and processes altered in a significant number of patients. These approaches can be categorized into 3 classes (**Figure 3**): methods based on predefined pathways; methods that extract pathways from a large interaction network of genes or proteins; *de novo* methods that do not use prior information of interactions among genes. Below we review some of the representative methods in each class. In general, the input to each method can be a list of genes mutated in the patients cohort or a score (e.g., frequency of mutation, a score reflecting the significance of the fraction of mutated genes in the cohort; Lawrence et al., 2013, etc.) for each gene in the cohort. As described below, while some of the methods require in input a list of putative driver mutations, identified for example by frequency-based approaches (e.g., Dees et al., 2012; Lawrence et al., 2013), other methods try to leverage the information regarding interactions among genes/proteins to identify novel driver genes which cannot be identified by frequency-based approaches. We highlight here the main methods that produce, in output, pathways, or sets genes summarizing inter-patient heterogeneity, while we do not consider methods which provide instead a ranking of genes (e.g., Vanunu et al., 2010; Shrestha et al., 2014), or which focus on patients stratification (e.g., Hofree et al., 2013), or which combine mutations with other data types (e.g., Vaske et al., 2010; McPherson et al., 2012; Paull et al., 2013). See Creixell et al. (2015) for a more comprehensive review of network approaches to analyze cancer genomes.

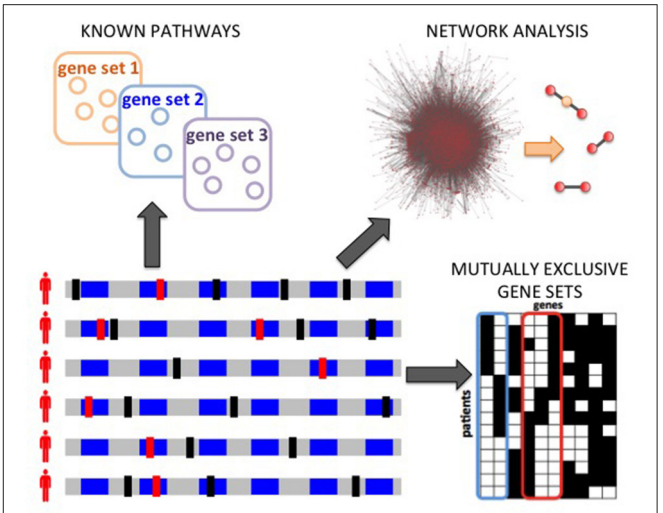


FIGURE 3 | Computational analyses to characterize inter-tumor heterogeneity. Starting from somatic mutations measured in many patients, different types of analyses are possible: annotation and enrichment analysis for known pathways; network analyses to discover significantly mutated subnetworks of a large interaction networks; the *de novo* identification of pathways, based for example on the identification of mutual exclusivity patterns.

2.1. Pathway-Based Approaches

A common way to identify significantly mutated pathways is to use *predefined pathways*, obtained from databases such as KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2015, 2017) and MSigDB (Subramanian et al., 2005; Liberzon et al., 2015), an then asses whether the set of genes in a predefined pathway is significantly enriched for mutated genes or scores compared to the entire set of genes. The simplest approach is to assess whether a list of mutated genes is enriched for genes in predefined set of genes, for example by using an hypergeometric test on the overlap of the intersection among the list of genes and the gene set. There are several tools [e.g., DAVID (Huang et al., 2009), g:Profiler (Reimand et al., 2016)], some of which originally designed for gene expression data, that can be used for gene lists obtained from mutation data. A common feature of these approaches is that they require the definition of the

list of mutated genes, commonly based on thresholds based on frequency or statistical significance of single genes. An alternative is to use Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), a general methodology to assess the association of a ranking of genes with a given gene set. The rank of the genes can be obtained from the various tools mentioned above; for example, Lin et al. (2007) used the Cancer Mutation Prevalence (CaMP) scores (Sjöblom et al., 2006), but other scores can be used. A different approach is taken by PathScan (Wendl et al., 2011), that computes a *p*-value for the enrichment of mutations in a given set separately for each patient, and then combines the *p*-values across all patients. Similarly, the method from Boca et al. (2010) defines, given a gene set, a score for each patient and then combines such scores across all patients.

The methods above are useful to characterize inter-tumor heterogeneity using known sets of genes and pathways, but have some major limitations. First, they require the *a priori* definition of the list of mutated genes, and therefore, while they are useful in organizing a list of mutated genes into pathways, they cannot be used to reliably identify novel driver genes. Second, some of the genes sets from datasets are extremely large (>300 genes). With such large gene sets it may not be possible to identify a small subset associated with the disease. Third, these methods ignore the interactions among genes in a network, considering all genes in a pathway equally, without including the topology of network in their analysis. Fourth, they consider each set of genes as a separate entity, while it is well-known that there is cross-talk among pathways, which interact into larger networks (McCormick, 1999).

2.2. Network-Based Approaches

A different approach to characterize cancer inter-tumor heterogeneity at the pathway level while not restricting to known sets of genes is to use a genome-scale protein-protein interaction network. Several computational methods that combine mutation data and networks to infer gene sets have been designed. A first class of methods (e.g., NetBox; Cerami et al., 2010) looks for significant network modules among a list of genes which is provided as input. Such approaches require to define a score threshold to include genes in the analysis, limiting the possibility of the method to identify novel driver genes. A different approach is to identify significant subnetworks (comprising connected genes) that are significantly mutated in the patients cohort. While allowing to expand from predefined sets of interacting genes to general interacting subnetworks, the identification of significantly mutated subnetwork presents computational and statistical challenges. There is a huge number of subnetworks which need to be screened and which need to be considered into a multiple hypothesis testing framework to identify the significantly mutated ones, therefore naïve approaches (e.g., the enumeration and testing of all subnetworks) cannot be employed and more sophisticated techniques are required.

HotNet (Vandin et al., 2011) and HotNet2 (Leiserson et al., 2015a) address the challenges above by using a diffusion process on a graph to combine gene scores with the network topology while capturing the local structure of the network. A novel statistical test is used by HotNet and HotNet2, allowing the identification of a set of subnetworks while bounding the false discovery rate (FDR). The combination of gene scores and network topology solves the issue of choosing a threshold for the inclusion of genes in the analysis and allows the identification of subnetworks whose significance is due to the mutation scores of the genes *and* the local topology of a subnetwork. In the analysis of >3,000 samples from 12 cancer types from TCGA (Leiserson et al., 2015a), HotNet2 identified 16 significantly mutated subnetworks that comprise well-known cancer pathways as well as subnetworks with less established contributions to cancer, including the cohesin complex.

MEMo (Ciriello et al., 2012) is an algorithm that uses a different approach to identify subnetworks: provided in input with a relative short of list of (frequently mutated) genes from which subnetworks (called modules) are to be found, it identifies groups of genes sharing several neighbors in the interaction network and showing significant mutual exclusivity of mutations in the patients cohort. MEMo therefore identifies modules summarize inter-patient heterogeneity through mutual exclusivity, but it is unlikely to include in its modules genes that are not significantly mutated on their own. MEMCover (Kim et al., 2015) is a different algorithm that combines network information and mutual exclusivity of mutations to identify modules of mutated genes. MEMCover employs a greedy strategy to identify high scoring subnetworks, where a subnetwork score is a combination of the number of patients with at a least a mutated subnetwork member and of the mutual exclusivity of mutations in the subnetwork genes. Babur et al. (2015) present a greedy approach to find gene sets sharing a common down-stream target in the network and showing high mutual exclusivity. They assess mutual exclusivity by comparing each gene in the set with the union of the other genes.

Network-based approaches are useful to characterize inter-tumor heterogeneity without restricting to know sets of genes and pathways, but they suffer from the limitations of currently available network. Such networks have only partial coverage of genes and interactions: some genes have no interactions in current networks, and interactions of different genes may have been assayed to different extents, with genes known to be associated to diseases that are likely to have been more thoroughly assayed for interactions (ascertainment bias). In addition, current networks include interactions that occur among proteins in different tissues or at different phases of the cell cycle. Improved methods are needed to integrate additional information (e.g., co-location of proteins in cells) with the interaction information provided by currently available networks.

2.3. *De novo* Approaches

Previous approaches are based on knowledge of the interactions among genes/proteins. A different class of methods characterize inter-tumor heterogeneity by finding groups of genes or pathways without restricting to predefined sets or to groups of interacting genes in a large network. The *de novo* extraction of pathways poses enormous computational and statistical challenges, since every subset of genes is a candidate which may need to be considered. However, some methods use

combinatorial properties (Yeang et al., 2008) of important mutations in cancer to restrict the set of potential candidates. One such property is *mutual exclusivity*, with sets of genes displaying at most 1 mutation in many patients. Mutual exclusivity of mutations has been observed in various cancer types (Kandoth et al., 2013) and may be due to the relatively low number of driver mutations in each tumor and to the fact that driver mutations target different pathways (Hanahan and Weinberg, 2011; Garraway and Lander, 2013; Vogelstein et al., 2013).

Several methods have been recently designed to identify gene sets with high mutual exclusivity. Since most genes are mutated with low frequency in a cohort of patients, it is easy to find a set of unrelated genes with high mutual exclusivity. For this reason, one needs to assess the statistical significance of the gene set, assessing whether the observed mutual exclusivity is likely to be due to chance alone. RME (Miller et al., 2011) identifies mutually exclusive sets using a score derived from information theory, and starts from pairs of genes to build larger sets. It includes only frequently mutated genes (>10%), limiting its applicability to characterize inter-tumor heterogeneity. Dendrix (Vandin et al., 2012b) defines a gene set score that combines the number of patients with at least a mutation in the set and the mutual exclusivity of mutations in the set, and uses a Markov Chain Monte Carlo (MCMC) approach for identifying mutually exclusive gene sets altered in a large fraction of the patients. Multi-Dendrix (Leiserson et al., 2013) employs the same score as Dendrix and extends it to multiple sets, and uses an integer linear program (ILP) based algorithm to simultaneously find multiple sets of mutually exclusive genes. CoMET (Leiserson et al., 2015b) uses a generalization of Fisher exact test to higher dimensional contingency tables to define a score that better characterizes mutually exclusive gene sets altered in relatively low fraction of the samples, and uses an efficient MCMC approach to identify such sets. WExT (Leiserson et al., 2015b) generalizes the test from CoMET to incorporate individual gene weights (probabilities) for each mutation in each sample, and provides an efficient way to assess the statistical significance of the sets using a saddle-point approximation. Similarly, WeSME (Kim Y.A. et al., 2016) introduces a test which incorporates the mutation rates of patients and genes and uses a fast permutation approach to assess the statistical significance of the sets. TiMEx (Constantinescu et al., 2015) assumes a generative model for mutations and defines a test to assess the null hypothesis that mutual exclusivity of a gene set is due to the interplay between waiting times to alterations and the time at which the tumor is sequenced. The test is used to assess pairs of genes, and larger sets are built from significant pairs and then assessed using the same test. As mentioned above, MEMo and the method from Babur et al. (2015) employ mutual exclusivity to find gene sets, but use an interaction network to limit the candidate gene sets. The method by Raphael and Vandin (2015) and PathTIMEx (Cristea et al., 2016) introduce an additional dimension to the characterization of inter-tumor heterogeneity, by reconstructing the order in which mutually exclusive gene sets are mutated. Kim J.W. et al. (2016) recently developed REVEALER, a method to identify mutually exclusive genes sets associated with functional phenotypes (see Section 4).

While the approaches above allow the *de novo* discovery of cancer gene sets, there are challenges that remain to be solved. For example, larger sample sizes than currently available may be needed to discover low frequency cancer pathways by using mutual exclusivity (Vandin et al., 2012c, 2016). The methods above are in general computationally intensive, mainly due to the large search space that must be explored, and more effective exploration strategies may be needed for larger datasets.

3. METHODS FOR INTRA-TUMOR HETEROGENEITY

In recent years, several methods have been proposed to characterize intra-tumor heterogeneity. Such methods can be classified into three classes (**Figure 4**). First, methods that use mutation data from bulk sequencing to reconstruct the *clonal* composition of a tumor, thus identifying the different *clones*, populations of cells, present in a tumor sample and quantifying the fraction that each clone contributes to the tumor. Second, methods that use mutation data from bulk sequencing to reconstruct the evolutionary relationships among different clones and mutations in the tumor. Third, more recent methods that use mutation data from single cell sequencing to infer the evolution of a tumor at the single cell level. Due to space constraints, below we describe some of the methods in the three classes; we point the reader to the recent reviews by Schwartz and Schäffer (2017) and by Kuipers et al. (2017) for more details on approaches to infer tumor evolution.

3.1. Inference of Clonal Composition from Bulk Sequencing

Bulk sequencing data provides information regarding the fraction of cells containing a mutation, and, therefore, regarding

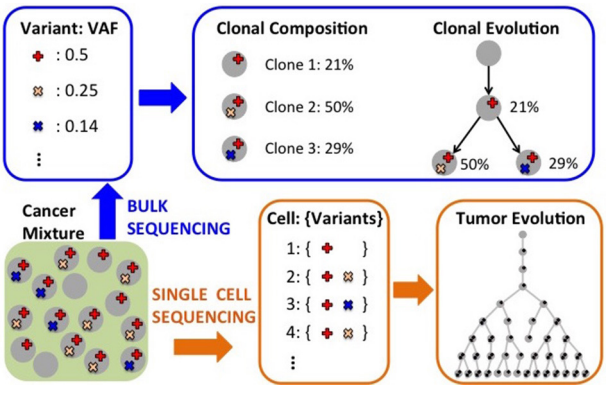


FIGURE 4 | Computational analyses to characterize intra-tumor heterogeneity. Starting from measurements (e.g.,VAFs) obtained from bulk sequencing of one or more tumor samples, one can infer the clonal composition of the sample and also the evolutionary relationships among clones. Single-cell sequencing can be use to infer the evolutionary relationships among the individual cells for which mutations have been assayed.

the fraction of cells defining the clone with the given mutation. In fact, for a heterozygous mutation in a copy neutral region the expected number of reads supporting the mutation (VAF) equals half of the clone frequency in the sample, since the mutation appears in only one of the two copies of the DNA. However, there are many confounders that make the identification of the clones not straightforward. First, the relation above holds only in expectation or for infinite coverage, while with finite coverage the actual VAF can deviate substantially from the corresponding clone frequency. Second, there are experimental biases in sequencing technologies that can change the relation between VAF and clonal frequency. Third, CNAs are quite common in cancer and nullify the relation above, making the inference of clones much more complex. Andor et al. (2016) have recently shown that the number of clones in a tumor is associated with mortality risk, which increases when between 2 and 4 clones are present in a tumor, while it decreases when >4 clones are present. The accurate characterization of the clonal composition of a tumor is therefore extremely important for diagnosis and therapy.

Several methods have been developed to identify the different clones, or cell populations, in a tumor starting from mutation data obtained from bulk sequencing. PyClone (Roth et al., 2014) identifies clones and their abundances by considering VAFs and allele-specific copy number data. It uses a beta-binomial model for VAFs and identifies clusters of mutations and their frequencies in a tumor sample with Bayesian nonparametric clustering which simultaneously infers clusters and the number of clusters. SciClone (Miller et al., 2014) considers VAFs in copy number neutral, loss of heterozygosity (LOH) free regions of the genome, and uses a variational Bayesian mixture model to infer clones and their frequency in the sample. Zare et al. (2014) present an algorithm to infer groups of mutations and their frequency in a tumor using mutation data from multiple sections of a tumor at a given time point. Their method is based a generative binomial model to incorporate information from the multiple sections and employs an expectation-maximization (EM) algorithm to estimate clones and their relative frequencies. BayClone (Sengupta et al., 2015) defines a class of nonparametric models, the categorical Indian buffet process, and uses bayesian inference to obtain posterior probabilities for the number clones, their genotypes, and their proportions, in a tumor sample.

With the coverage (30x–40x) used in many large scale cancer studies, there is a high variance in the number of reads covering a given position in the genome, weakening the relation between VAF and clonal frequency. In contrast, each copy number aberration perturbs many reads, and can provide a more reliable signal for clonal inference for tumors in which clones present different copy number profiles. THetA (Oesper et al., 2013) uses CNAs profiles from whole genome sequencing to characterize clones and their frequencies in a tumor mixture. It defines and optimizes an explicit probabilistic model for the generation of the observed sequencing data from a mixture of normal cells and different clones, and uses a BIC criteria to choose among the many models that may explain the data while balancing the likelihood of the data and the model complexity. THetA2 (Oesper et al., 2014) extends THetA in various directions, including the possibility to consider whole exome sequencing data and the use of B-allele frequencies (which indicates the relative quantity of the one allele compared to the other) to distinguish among several clonal population models consistent with the data. A different approach is taken by TITAN (Ha et al., 2014), which employs a generative factorial hidden Markov model framework to simultaneously infer CNA and LOH segments from read depths and digital allele ratios at heterozygous variant loci in the genome from whole genome sequencing data. CloneHD (Fischer et al., 2014) provides a statistical framework using read depth, B-allele frequencies, and VAFs to infer the clonal population structure of a tumor, allowing the simultaneous analysis of multiple samples from different regions of the same tumor or from longitudinal sequencing of the same tumor.

3.2. Inference of Clonal Evolution from Bulk Sequencing

While methods to infer clones, their mutations, and their abundance, provide important and clinically relevant insights into intra-tumor heterogeneity, they do not explicitly provide information about the evolutionary relations among mutations and clones in a tumor. In addition to expanding our understanding of how a tumor arises, such information can provide extremely important information for clinical intervention. For example, the order in which mutations arise can influence the prognosis of a patient (Ortmann et al., 2015). Moreover, the characterization of the evolutionary paths followed by tumors is crucial to be able to predict the development of the disease for future patients (Yachida et al., 2010; Lipinski et al., 2016).

The computational reconstruction of the evolutionary relations among clones in a tumor from bulk sequencing data is a challenging task, due to several reasons. First, we do not directly observe clones in a tumor, but bulk sequencing provide aggregate information, in the form of VAFs, from a mixture of clones. Second, a natural model to describe tumor evolution is provided by phylogenetic or evolutionary trees, but there are in general several evolutionary trees consistent with the data from a single tumor sample. In most cases this may be mitigated by sequencing several sections of the same tumor, but reconciling the information from the different sections is a complex problem. Third, VAFs in regions affected by CNAs and LOH can be significantly different from VAFs of other mutations in the same clone, complicating the reliable identification of clones and their relations.

Many methods have been designed to reconstruct the evolutionary history of a tumor from bulk sequencing of one or more sections of the tumor and address the challenges above. TrAp (Strino et al., 2013) is a method designed to infer clones, their abundance, and clones' evolutionary paths using VAFs for SNVs from a single tumor sample. It first groups together mutations with similar frequencies, and then uses an iterative procedure to build evolutionary paths for such groups, starting from simple (height 1) trees. PhyloSub (Jiao

et al., 2014) considers VAFs from deep sequencing experiments to infer the evolutionary relationship of clones, and uses a Dirichlet process prior over phylogenetic trees to group SNVs into clones. It employs Bayesian inference, based on MCMC sampling, to infer a distribution over possible evolutionary trees. PhyloWGS (Deshwar et al., 2015) builds on PhyloSub and allows the reconstruction of tumor evolution from SNVs and CNAs obtained from whole genome sequencing data. CITUP (Malikic et al., 2015) proposes a combinatorial model for the problem of inferring clonal evolution from SNVs obtained from multiple tumor samples, and designs an exact algorithm based on a quadratic integer programming to solve the problem, which may require high computational resources when the tumor contains a large number of clones. LICHeE (Popic et al., 2015) is another method to reconstruct clones, abundances, and their evolutionary relationships starting from SNVs measured in multiple samples of a tumor. LICHeE first groups SNVs and identifies clusters of SNVs based on VAFs, and then uses a network to represent VAFs constraints imposed by the evolutionary process. It then identifies an evolutionary model by looking for the spanning tree that best supports the cluster VAF data. BitPhylogeny (Yuan et al., 2015) provides a probabilistic framework that allows the joint inference of the number and composition of clones in a tumor, as well as the most probable tree representing their evolutionary relationship. SPRUCE (El-Kebir et al., 2016) infers evolutionary trees jointly from SNVs and CNAs from multiple tumor samples, with CNAs that are modeled as multi-state alterations, in which alterations can only mutate to a given state at most once in the tree. SPRUCE starts from clusters of SNVs and copy number mixing proportions, and derives a compatibility graph describing the compatibility of state trees for pairs of clusters. The evolutionary trees compatible with the input data are derived by enumerating all spanning trees with appropriate constraints in a labeled multi-graph constructed starting from the compatibility graph. The application of SPRUCE on real data show that many evolutionary trees are compatible with data from multiple samples, cautioning on drawing strong conclusions on any single such tree (Hu and Curtis, 2016). Canopy (Jiang et al., 2016) is a related method to infer evolutionary trees using both CNAs and SNVs from one or more samples, but it starts from raw copy number ratios estimated from CNA segmentation programs. It uses a statistical model and a MCMC algorithm to sample from the space of evolutionary trees, providing a confidence assessment from the posterior distribution. Additional methods to infer clonal evolution are presented in Hajirasouliha et al. (2014), Donmez et al. (2016), Qiao et al. (2014), and El-Kebir et al. (2015).

While each method displays specific features addressing one or more of the challenges above, they are all based, in one form or the other, on the infinite-site assumption: the same site is not mutated twice during the evolutionary history of a tumor. Such assumption may be violated in tumors with high genomic instability, undermining the accuracy of the inferred evolutionary trees. However, without such assumption the inference problem becomes computationally intractable even assuming perfect knowledge of mutations in each clone.

3.3. Inference from Single Cell Sequencing

While bulk sequencing provides some information to infer the evolutionary tree describing a tumor history, the best way to elucidate such history is from single-cell data, which provides direct measurements for some of the leaves of a tumor evolutionary tree. Single-cell sequencing technology has been improving in recent years and datasets with SNVs from >40 single-cells are now available (Hou et al., 2012; Xu et al., 2012; Wang et al., 2014). However, mutation calls from single-cell sequencing still suffer from high false positive and false negative rate and missing values, due to various technical reasons (e.g., allele dropout; Kuipers et al., 2017). In addition, while obtaining measurements from hundreds of single-cells is an incredible advance, such cells still represent an extremely small fraction of all cells in a tumor ($> 10^9$ in advanced tumors). For these reasons, standard phylogenetic approaches cannot be used to infer evolutionary trees from single cell data.

Few methods have been designed to infer the evolutionary relationships among single cells. Youn and Simon (2011) develop a method to infer a *mutation tree*, in which each node corresponds to a mutation and the tree relations describe the relative order among the appearance of mutations in a sample. The mutation tree is reconstructed by using a pairwise test to define the order for pairs of mutations. While the restriction to pairs of genes makes the method efficient, it discards the information among high order relations among mutations. SCITE (Jahn et al., 2016) identifies evolutionary trees from noisy and incomplete mutation data from single-cell sequencing. SCITE uses a statistical model and an MCMC approach to sample trees, error rates, and placement of single cells in the tree. While providing interesting insights, the method is fairly expensive computationally, allowing proper inference only for the limited number of cells available in current datasets. OncoNEM (Ross and Markowetz, 2016) is a related method that uses a nested effects model for the data and employs a heuristic local search algorithm to explore possible tree topologies. While appropriate for current dataset sizes, for much larger dataset such a search algorithm may be too expensive.

4. ASSESSING THE ASSOCIATION OF CANCER HETEROGENEITY WITH CLINICAL VARIABLES

A major goal in characterizing inter- and intra-tumor heterogeneity is to understand its impact on prognosis and therapy. In most case, clinical data has been used after the computational characterization of tumor heterogeneity, as a post-processing step testing whether heterogeneity-related features are associated to or predictive for some clinical variable, mostly survival time. For example: survival data or other clinical information are used to evaluate the results of patients stratification methods (Hofree et al., 2013); Andor et al. (2016) computationally assessed the clonal composition of $>1{,}000$ samples of various cancer types and then assessed the association between the number of clones in a sample with overall and progression-free survival; Chowdhury et al. (2014,

2015) designed and used a novel algorithm to reconstruct trees describing cancer evolution from single cell copy number data obtained by fluorescence *in situ* hybridization (FISH), and showed that improved prediction accuracy is obtained for classification tasks (e.g., distinguishing primary vs. metastases in the same patient) when features from the cancer evolutionary tree are considered.

The discovery of mutations or mutated groups associated with clinical data starting from genome-wide measurements poses several challenges, due to the peculiar characteristic of genomic data, including the relatively low frequency of individual mutations (Vandin et al., 2015). A standard analysis (Gross et al., 2014) is to first identify driver mutations or pathways and then assess the association of the mutated genes or group of genes with a clinical variable (e.g., survival time). While providing useful information on the clinical relevance of the driver genes and pathway identified by approaches above, such methods may not identify groups of genes with low mutation frequency whose mutations are collectively associated with survival. Few methods have been developed to directly leverage clinical information to identify gene sets associated with clinical data. Vandin et al. (2012a) use gene scores derived from the *p*-values for the association of individual gene mutations with survival as input to HotNet to identify subnetworks associated with survival, but do not provide a method to directly identify gene sets associated with survival. HyperModules (Reimand and Bader, 2013) looks for subnetworks of a large interaction network that are associated with survival using a local search algorithm that builds a subnetwork by starting from one seed vertex and then greedily adds neighbors (at distance at most 2) from the seed. Leung et al. (2014) used it to find subnetworks of a kinase-substrate interaction network with phosphorylation-associated mutations associated with survival. NoMAS (Hansen and Vandin, 2016) is an efficient method based on graph color-coding which identifies subnetworks with mutations associated with survival by looking for subnetworks maximizing the log-rank statistic of subnetworks (**Figure 5**). NoMAS identifies subnetworks with stronger association with survival compared to greedy procedures, and also reports valid permutational *p*-values. REVEALER (Kim J.W. et al., 2016) is a computational method to identify groups of mutually exclusive genes correlated with a functional phenotype, for example sensitivity to a drug treatment. REVEALER uses a gene set score derived from mutual information and employs a greedy strategy to find genes sets associated with the target functional phenotype.

The methods above provide initial steps to discover gene sets driven by inter-tumor heterogeneity and associated with clinical features, but much more work is required to identify clinically relevant features from tumor heterogeneity.

5. CONCLUSIONS AND FUTURE PERSPECTIVE

This review described some of the challenges that arise in studying and characterizing cancer inter- and intra-tumor heterogeneity. We focused on some computational methods which characterize inter-tumor heterogeneity at the level of

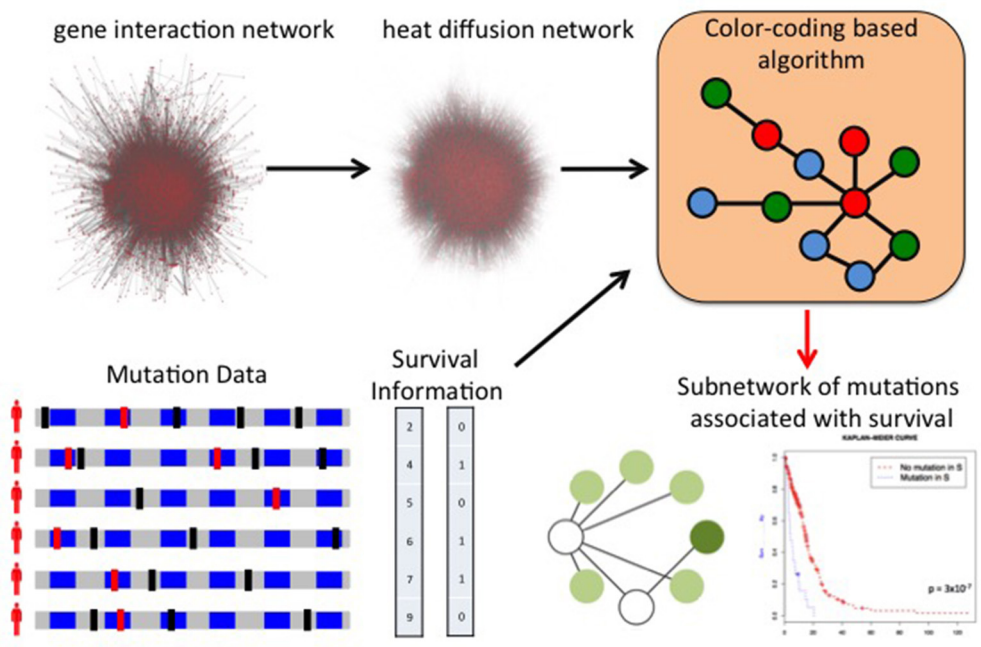


FIGURE 5 | Network of Mutations Associated with Survival (NoMAS). NoMAS combines mutations measured in many patients and the corresponding survival time with a large interaction network to identify subnetworks of genes with significant association to survival. NoMAS is based on an efficient graph color-coding algorithm, and uses a permutation test to correct for multiple hypothesis testing.

pathways, infer intra-tumor heterogeneity from bulk or single-cell sequencing, and identify pathways associated with clinical variables. These and other methods are increasingly used to characterize heterogeneity in large sequencing studies and for individual patients. Given its importance for therapeutic decisions, the fast and precise characterization of cancer heterogeneity is likely to remain a key step in precision medicine.

The methods we described have significantly advanced our understanding of cancer heterogeneity and its importance in patient prognosis and treatment, but there still challenges to be addressed. First, while recent studies have shown that intra-tumor heterogeneity has clinical implications (McGranahan and Swanton, 2015, 2017; Andor et al., 2016), it is still unclear which ones among its features are key determinants for therapeutic decisions. The development of more precise computational methods to infer intra-tumor clonal composition and evolution is a necessary step to properly assess the relevance of each aspect for therapy and inform effort for noninvasive monitoring of tumors (e.g., liquid biopsies; Diaz and Bardelli, 2014). Second, the extensive intra-tumor heterogeneity and the stochasticity of some of the processes shaping the evolution of a tumor may limit the ability to accurately predict the future behavior of an individual cancer. Studies (e.g., Jamal-Hanjani et al., 2014) that are collecting molecular and clinical measurements at different time points during treatment for a large number of patients will provide the data necessary to understand the extent of the diversity in the evolutionary paths explored by different tumors, but substantially different computational methods are needed to rigorously and effectively analyze such datasets. Third, current methods for inferring a tumor evolution from single-cell data are computationally intensive, and will not be able to analyze much larger datasets which may soon be available. Fourth, current methods for analyzing bulk sequencing and single-cell sequencing data are orthogonal, but the two technologies provide complementary information about the same tumor. ddClone (Salehi et al., 2017) is a recent method which combines data from the two technologies, but the development of additional methods may be crucial in fully exploiting the power of next-generation sequencing to characterize cancer heterogeneity. Fifth, methods for inter-patient heterogeneity focus mostly on coding variants, while noncoding variants are known to be recurrently mutated in cancer (Weinhold et al., 2014; Melton et al., 2015; Puente et al., 2015), with the mutation in the promoter region of the TERT gene in melanoma (Huang et al., 2013) and other cancer types (Fredriksson et al., 2014; Melton et al., 2015) being a prominent example. Finally, other data types, including RNA sequencing, methylation data, and chromatin modifications need to be considered to understand the genomic heterogeneity of cancer. While there are some methods that integrate some of these data types with mutation data (Vaske et al., 2010; McPherson et al., 2012; Paull et al., 2013), additional work is required to characterize cancer heterogeneity by the full integration of the various data types. All these challenges need to be addressed to reach true precision medicine, and computational methods will continue to play a key role in advancing our understanding of cancer heterogeneity.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work is supported, in part, by MIUR of Italy under project AMANDA and by NSF grant IIS-1247581.

REFERENCES

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.* 47, 1402–1407. doi: 10.1038/ng.3441

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science* 354, 618–622. doi: 10.1126/science.aag0299

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013a). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 3, 246–259. doi: 10.1016/j.celrep.2012.12.008

Anderson, K., Lutz, C., Van Delft, F. W., Bateman, C. M., Guo, Y., Colman, S. M., et al. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* 469, 356–361. doi: 10.1038/nature09650

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., et al. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* 22, 105–113. doi: 10.1038/nm.3984

Babur, Ö., Gönen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., et al. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* 16, 45. doi: 10.1186/s13059-015-0612-6

Boca, S. M., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., and Parmigiani, G. (2010). Patient-oriented gene set analysis for cancer mutation data. *Genome Biol.* 11:R112. doi: 10.1186/gb-2010-11-11-r112

Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., et al. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* 5:2997. doi: 10.1038/ncomms3997

Brastianos, P. K., Carter, S. L., Santagata, S., Cahill, D. P., Taylor-Weiner, A., Jones, R. T., et al. (2015). Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* 5, 1164–1177. doi: 10.1158/2159-8290.CD-15-0369

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE* 5:e8918. doi: 10.1371/journal.pone.0008918

Chowdhury, S. A., Gertz, E. M., Wangsa, D., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. A., et al. (2015). Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics* 31, i258–i267. doi: 10.1093/bioinformatics/btv233

Chowdhury, S. A., Shackney, S. E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. A., and Schwartz, R. (2014). Algorithms to model single

gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comput. Biol.* 10:e1003740. doi: 10.1371/journal.pcbi.1003740

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., and Beerenwinkel, N. (2015). Timex: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics* 32, 968–975. doi: 10.1093/bioinformatics/btv400

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., et al. (2015). Pathway and network analysis of cancer genomes. *Nat. Methods* 12:615. doi: 10.1038/nmeth.3440

Cristea, S., Kuipers, J., and Beerenwinkel, N. (2016). pathtimex: joint inference of mutually exclusive cancer pathways and their progression dynamics. *J. Comput. Biol.* 24, 603–615. doi: 10.1089/cmb.2016.0171

D'Antonio, M., and Ciccarelli, F. D. (2013). Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.* 14:R52. doi: 10.1186/gb-2013-14-5-r52

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). Music: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 16, 35. doi: 10.1186/s13059-015-0602-8

Diaz, L. A., and Bardelli, A. (2014). Liquid biopsies: genotyping circulating tumor dna. *J. Clin. Oncol.* 32, 579–586. doi: 10.1200/JCO.2012.45.2011

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481, 506–510. doi: 10.1038/nature 10738

Donmez, N., Malikic, S., Wyatt, A. W., Gleave, M. E., Collins, C. C., and Sahinalp, S. C. (2016). "Clonality inference from single tumor samples using low coverage sequence data," in *International Conference on Research in Computational Molecular Biology* (Santa Monica: Springer), 83–94.

El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 31, i62–i70. doi: 10.1093/bioinformatics/btv261

El-Kebir, M., Satas, G., Oesper, L., and Raphael, B. J. (2016). Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.* 3, 43–53. doi: 10.1016/j.cels.2016.07.004

Fischer, A., Vázquez-García, I., Illingworth, C. J., and Mustonen, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep.* 7, 1740–1752. doi: 10.1016/j.celrep.2014.04.055

Fredriksson, N. J., Ny, L., Nilsson, J. A., and Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* 46, 1258–1263. doi: 10.1038/ng.3141

Garraway, L. A., and Lander, E. S. (2013). Lessons from the cancer genome. *Cell* 153, 17–37. doi: 10.1016/j.cell.2013.03.002

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* 46, 225–233. doi: 10.1038/ng.2891

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 2012, 883–892. doi: 10.1056/NEJMoa1113205

Greaves, M., and Maley, C. C. (2012). Clonal evolution in cancer. *Nature* 481, 306–313. doi: 10.1038/nature10762

Gross, A. M., Orosco, R. K., Shen, J. P., Egloff, A. M., Carter, H., Hofree, M., et al. (2014). Multi-tiered genomic analysis of head and neck cancer ties tp53 mutation to 3p loss. *Nat. Genet.* 46, 939–943. doi: 10.1038/ng.3051

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M., Papaemmanuil, E., et al. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature* 520, 353–357. doi: 10.1038/nature14347

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., et al. (2014). Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* 24, 1881–1893. doi: 10.1101/gr.180281.114

Hajirasouliha, I., Mahmoody, A., and Raphael, B. J. (2014). A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics* 30, i78–i86. doi: 10.1093/bioinformatics/btu284

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hansen, T., and Vandin, F. (2016). "Finding mutated subnetworks associated with survival time in cancer," in *20th Annual Conference on Research in Computational Molecular Biology*. Santa Monica.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., et al. (2012). Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell* 148, 873–885. doi: 10.1016/j.cell.2012.02.028

Hu, Z., and Curtis, C. (2016). Inferring tumor phylogenies from multi-region sequencing. *Cell Syst.* 3, 12–14. doi: 10.1016/j.cels.2016.07.007

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., and Garraway, L. A. (2013). Highly recurrent tert promoter mutations in human melanoma. *Science* 339, 957–959. doi: 10.1126/science.1229259

Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biol.* 17:86. doi: 10.1186/s13059-016-0936-x

Jamal-Hanjani, M., Hackshaw, A., Ngai, Y., Shaw, J., Dive, C., Quezada, S., et al. (2014). Tracking genomic cancer evolution for precision medicine: the lung tracerx study. *PLoS Biol.* 12:e1001906. doi: 10.1371/journal.pbio.1001906

Jiang, Y., Qiu, Y., Minn, A. J., and Zhang, N. R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5528–E5537. doi: 10.1073/pnas.1522203113

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* 15:35. doi: 10.1186/1471-2105-15-35

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kanehisa, M., and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070

Kim, J. W., Botvinnik, O. B., Abudayyeh, O., Birger, C., Rosenbluh, J., Shrestha, Y., et al. (2016). Characterizing genomic alterations in cancer by complementary functional associations. *Nat. Biotechnol.* 34, 539-546. doi: 10.1038/nbt.3527

Kim, Y.-A., Cho, D.-Y., Dao, P., and Przytycka, T. M. (2015). Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* 31, i284–i292. doi: 10.1093/bioinformatics/btv247

Kim, Y.-A., Madan, S., and Przytycka, T. M. (2016). Wesme: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* 33, 814–821. doi: 10.1093/bioinformatics/btw242

Kuipers, J., Jahn, K., and Beerenwinkel, N. (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochim. Biophys. Acta* 1867, 127–138. doi: 10.1016/j.bbcan.2017.02.001

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Leiserson, M. D., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9:e1003054. doi: 10.1371/journal.pcbi.1003054

Leiserson, M. D., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015a). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Leiserson, M. D., Wu, H.-T., Vandin, F., and Raphael, B. J. (2015b). Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 16:160. doi: 10.1186/s13059-015-0700-7

Leung, A., Bader, G. D., and Reimand, J. (2014). Hypermodules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. *Bioinformatics* 30, 2230–2232. doi: 10.1093/bioinformatics/btu172

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjöblom, T., Wood, L. D., et al. (2007). A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.* 17, 1304–1318. doi: 10.1101/gr.6431107

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6496–E6505. doi: 10.1073/pnas.1519556112

Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., and Gerlinger, M. (2016). Cancer evolution and the limits of predictability in precision cancer medicine. *Trends Cancer* 2, 49–63. doi: 10.1016/j.trecan.2015.11.003

Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* 31, 1349–1356. doi: 10.1093/bioinformatics/btv003

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

McCormick, F. (1999). Signalling networks that cause cancer. *Trends Biochem. Sci.* 24, M53–M56. doi: 10.1016/S0968-0004(99)01480-2

McGranahan, N., and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* 27, 15–26. doi: 10.1016/j.ccell.2014.12.001

McGranahan, N., and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 168, 613–628. doi: 10.1016/j.cell.2017.01.018

McPherson, A., Wu, C., Wyatt, A. W., Shah, S., Collins, C., and Sahinalp, S. C. (2012). nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res.* 22, 2250–2261. doi: 10.1101/gr.136572.111

Melton, C., Reuter, J. A., Spacek, D. V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* 47, 710–716. doi: 10.1038/ng.3332

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11, 685–696. doi: 10.1038/nrg2841

Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D., and Milosavljevic, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics* 4:34. doi: 10.1186/1755-8794-4-34

Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., et al. (2014). Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* 10:e1003665. doi: 10.1371/journal.pcbi.1003665

Navin, N. E. (2015a). Delineating cancer evolution with single-cell sequencing. *Sci. Transl. Med.* 7, 296fs29. doi: 10.1126/scitranslmed.aac8319

Navin, N. E. (2015b). The first five years of single-cell cancer genomics and beyond. *Genome Res.* 25, 1499–1507. doi: 10.1101/gr.191098.115

Newburger, D. E., Kashef-Haghighi, D., Weng, Z., Salari, R., Sweeney, R. T., Brunner, A. L., et al. (2013). Genome evolution during progression to breast cancer. *Genome Res.* 23, 1097–1108. doi: 10.1101/gr.151670.112

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., et al. (2012a). Mutational processes molding the genomes of 21 breast cancers. *Cell* 149, 979–93. doi: 10.1016/j.cell.2012.04.024

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47–54. doi: 10.1038/nature17676

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., et al. (2012b). The life history of 21 breast cancers. *Cell* 149, 994–1007. doi: 10.1016/j.cell.2012.04.023

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science* 194, 23–28. doi: 10.1126/science.959840

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol.* 14:R80. doi: 10.1186/gb-2013-14-7-r80

Oesper, L., Satas, G., and Raphael, B. J. (2014). Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* 30, 3532–3540. doi: 10.1093/bioinformatics/btu651

Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., et al. (2015). Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* 372, 601–612. doi: 10.1056/NEJMoa1412098

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics* 29, 2757–2764. doi: 10.1093/bioinformatics/btt471

Petljak, M., and Alexandrov, L. B. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* 37, 531–540. doi: 10.1093/carcin/bgw055

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* 16:91. doi: 10.1186/s13059-015-0647-8

Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526:519–524. doi: 10.1038/nature14666

Qiao, Y., Quinlan, A. R., Jazaeri, A. A., Verhaak, R. G., Wheeler, D. A., and Marth, G. T. (2014). Subcloneseeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol.* 15:443. doi: 10.1186/s13059-014-0443-x

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* 6:5. doi: 10.1186/gm524

Raphael, B. J., and Vandin, F. (2015). Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. *J. Comput. Biol.* 22, 510–527. doi: 10.1089/cmb.2014.0161

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., et al. (2016). g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 44, W83–W89. doi: 10.1093/nar/gkw199

Reimand, J., and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* 9:637. doi: 10.1038/msb.2012.68

Ross, E. M., and Markowetz, F. (2016). Onconem: inferring tumor evolution from single-cell sequencing data. *Genome Biol.* 17:69. doi: 10.1186/s13059-016-0929-9

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., et al. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nat. Methods* 11, 396–398. doi: 10.1038/nmeth.2883

Salehi, S., Steif, A., Roth, A., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2017). ddclone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.* 18:44. doi: 10.1186/s13059-017-1169-3

Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., et al. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* 120, 4191–4196. doi: 10.1182/blood-2012-05-433540

Schwartz, R., and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* 18, 213–229. doi: 10.1038/nrg.2016.170

Sengupta, S., Wang, J., Lee, J., Müller, P., Gulukota, K., Banerjee, A., et al. (2015). "Bayclone: Bayesian nonparametric inference of tumor subclones using ngs data," in *Pacific Symposium on Biocomputing* (Big Island), Vol. 20:467.

Shrestha, R., Hodzic, E., Yeung, J., Wang, K., Sauerwald, T., Dao, P., et al. (2014). "Hit'ndrive: multi-driver gene prioritization based on hitting time," in *International Conference on Research in Computational Molecular Biology* (Pittsburgh, PA: Springer), 293–306.

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* 314, 268–274. doi: 10.1126/science.1133427

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., et al. (2015). A big bang model of human colorectal tumor growth. *Nat. Genet.* 47, 209–216. doi: 10.1038/ng.3214

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458, 719–724. doi: 10.1038/nature07943

Strino, F., Parisi, F., Micsinai, M., and Kluger, Y. (2013). Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.* 41:e165. doi: 10.1093/nar/gkt641

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Swanton, C. (2012). Intratumor heterogeneity: evolution through space and time. *Cancer Res.* 72, 4875–4882. doi: 10.1158/0008-5472.CAN-12-2217

Swanton, C. (2016). Tumor evolutionary principles: how intratumor heterogeneity influences cancer treatment and outcome. *Am. Soc. Clin. Oncol.* 35, e141–e149. doi: 10.14694/EDBK_158930

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* 3:2650. doi: 10.1038/srep02650

The Cancer Genome Atlas Research Network (2017a). Integrated genomic and molecular characterization of cervical cancer. *Nature* 543, 378–384. doi: 10.1038/nature21386

The Cancer Genome Atlas Research Network, (2017b). Integrated genomic characterization of oesophageal carcinoma. *Nature* 541, 169–175. doi: 10.1038/nature20805

Vandin, F., Clay, P., Upfal, E., and Raphael, B. J. (2012a). Discovery of mutated subnetworks associated with clinical data in cancer. *Pac. Symp. Biocomput.* 2012, 55–66. doi: 10.1142/9789814366496_0006

Vandin, F., Papoutsaki, A., Raphael, B. J., and Upfal, E. (2015). Accurate computation of survival statistics in genome-wide studies. *PLoS Comput. Biol.* 11:e1004071. doi: 10.1371/journal.pcbi.1004071

Vandin, F., Raphael, B. J., and Upfal, E. (2016). On the sample complexity of cancer pathways identification. *J. Comput. Biol.* 23, 30–41. doi: 10.1089/cmb.2015.0100

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522.

Vandin, F., Upfal, E., and Raphael, B. J. (2012b). *De novo* discovery of mutated driver pathways in cancer. *Genome Res.* 22, 375–385. doi: 10.1101/gr.120477.111

Vandin, F., Upfal, E., and Raphael, B. J. (2012c). Finding driver pathways in cancer: models and algorithms. *Algorithms Mol. Biol.* 7:23. doi: 10.1186/1748-7188-7-23

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* 6:e1000641. doi: 10.1371/journal.pcbi.1000641

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* 26, i237–i245. doi: 10.1093/bioinformatics/btq182

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* 10, 789–799. doi: 10.1038/nm1087

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512, 155–160. doi: 10.1038/nature13600

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* 46, 1160–1165. doi: 10.1038/ng.3101

Wendl, M. C., Wallis, J. W., Lin, L., Kandoth, C., Mardis, E. R., Wilson, R. K., et al. (2011). Pathscan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* 27, 1595–1602. doi: 10.1093/bioinformatics/btr193

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148, 886–895. doi: 10.1016/j.cell.2012.02.025

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* 467, 1114–1117. doi: 10.1038/nature09515

Yates, L. R., and Campbell, P. J. (2012). Evolution of the cancer genome. *Nat. Rev. Genet.* 13, 795–806. doi: 10.1038/nrg3317

Yeang, C.-H., McCormick, F., and Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.* 22, 2605–2622. doi: 10.1096/fj.08-108985

Youn, A., and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 27, 175–181. doi: 10.1093/bioinformatics/btq630

Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* 16, 36. doi: 10.1186/s13059-015-0592-6

Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., et al. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* 10:e1003703. doi: 10.1371/journal.pcbi.1003703

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

frontiers in Plant Science

MINI REVIEW
published: 09 January 2018
doi: 10.3389/fpls.2017.02241

Current Knowledge and Computational Techniques for Grapevine Meta-Omics Analysis

Salvatore Alaimo[1], *Gioacchino P. Marceca*[1], *Rosalba Giugno*[2], *Alfredo Ferro*[1] and *Alfredo Pulvirenti*[1]*

[1] Bioinformatics Unit, Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy, [2] Department of Computer Science, University of Verona, Verona, Italy

Growing grapevine (*Vitis vinifera*) is a key contribution to the economy of many countries. Tools provided by genomics and bioinformatics did help researchers in obtaining biological knowledge about the different cultivars. Several genetic markers for common diseases were identified. Recently, the impact of microbiome has been proved to be of fundamental importance both in humans and in plants for its ability to confer protection or induce diseases. In this review we report current knowledge about grapevine microbiome, together with a description of the available computational methodologies for meta-omics analysis.

Keywords: ***Vitis vinifera*, microbiome, metagenomics, metatranscriptomics, bioinformatic tools and databases**

OPEN ACCESS

Edited by:
Shrikant S. Mantri, National Agri-Food Biotechnology Institute, India

Reviewed by:
Kashmir Singh, Panjab University, Chandigarh, India
Lei Song, National Cancer Institute (NIH), United States

***Correspondence:**
Alfredo Pulvirenti
apulvirenti@dmi.unict.it

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

Received: *16 July 2017*
Accepted: *20 December 2017*
Published: *09 January 2018*

Citation:
Alaimo S, Marceca GP, Giugno R, Ferro A and Pulvirenti A (2018) Current Knowledge and Computational Techniques for Grapevine Meta-Omics Analysis. Front. Plant Sci. 8:2241. doi: 10.3389/fpls.2017.02241

1. INTRODUCTION

Vitis vinifera is one of the most important plant in modern agriculture. Its economic and cultural impact is undeniable (Mullins et al., 1992; Pulvirenti et al., 2015). Almost 8 million hectares of vineyards (Vivier and Pretorius, 2000) together with 5,000 estimated cultivars (Jackson, 1994) make grapevine economic contribution to wine-producing countries very significant. Improving wine quality and increasing grapevine pathogens and environmental stress resistance is crucial for wine industry (Vivier and Pretorius, 2002).

In 2007, the French-Italian Public Consortium for Grapevine Genome Characterization sequenced the first genome of *V. vinifera* Jaillon et al. (2007). Grapes genome has driven a multitude of studies on the genetics of grapes (Tomkins et al., 2001; Adam-Blondon et al., 2005, 2011; Lamoureux et al., 2006). Research attention has been also focused on grapevine transcriptome analysis. Gene expressions and trascriptional profiling data analysis have shed light on several *V. vinifera* biological processes. In particular the following biological functions have been considered: (i) ripening and maturation (Fortes et al., 2011; Guillaumie et al., 2011; Fasoli et al., 2012; Lijavetzky et al., 2012); (ii) dormancy transitioning (Sreekantan et al., 2010); (iii) and resistance to pathogens and environmental conditions (Grimplet et al., 2009; Polesani et al., 2010; Tillett et al., 2011; Pulvirenti et al., 2015). Nowadays, an increasing amount of knowledge proves the importance of beneficial plant-associated microbes in plant health, growth, nutrition and stress resistance, as well as in increasing crop quality thanks to their biostimulant properties (Rouphael et al., 2015; Pieterse et al., 2016). From a functional point of view, plant microbiome is comparable to gut microbiome in mammals, and it has been defined as the plant's second genome. In this perspective, plant fitness (a quantitative description of survival and reproductive success of a plant in a given environment) is the point of convergence of two components: the plant itself and its associated microbiota, which collectively form a holobiont (Vandenkoornhuyse et al., 2015). Indeed, it has been shown that

grapevine growth and survival are significantly impacted by its microflora. This latter has been identified as a key factor influencing not only plant fitness, but also many vine traits, positively contributing viticulture economy. Many viticulturists are now persuaded that wine organoleptic properties are partly due to microbes influence.

One of the most accepted principle in this field concerns the beneficial effects of *mycorrhizal symbiosis*. Indeed, grapevine growth is remarkably dependent on mycorrhizae, since this plant has low-density roots and few root hairs. Mycorrhizal fungi improve water use efficiency, soil nutrient uptake, and biomass production in grapevine. Furthermore, arbuscular mycorrhizal communities composition in vineyards is largely influenced by surrounding vegetation (Holland et al., 2014). According to the latest studies, based principally on culture-independent methods, the eukaryotic microbiome of *V. vinifera* is characterized by fungi belonging to early diverging fungal lineages, *Ascomycota* and *Basidiomycota*. On the other hand, prokaryotic microbiome is prevalently composed by *Proteobacteria*, followed by *Firmicutes*, *Actinobacteria* and *Bacteroidetes*. At species level, these microorganisms, either *epiphytes* or *endophytes*, are not uniformly distributed along the vine. Belowground microbial communities significantly differ from aboveground ones. More precisely, the level of biodiversity decreases from belowground to aboveground. Furthermore, aboveground microbiota is subject to temporal variation during grapevine vegetative cycle and agricultural practices like integrated pest management (Pancher et al., 2012; Martins et al., 2013; Campisano et al., 2014; Pinto et al., 2014; Zarraonaindia et al., 2015). In healthy conditions, within the aboveground fungal microorganisms, the *Aureobasidium* genus results to be predominant and ubiquitous all over vine's organs, including leaves and barks. In this group, *A. pullulans* is by far the most abundant fungal species. *Cryptococcus* spp. and *Rhodotorula* spp. are frequently found on leaves and grapes with a high relative abundance, while *Candida* spp. and *Pichia* spp. are more present on grapes than other parts of the plant. All together, these microorganisms are capable of exerting antibacterial activity, inhibiting spore germination and mold growth, resulting in a protector effect on vine and grapes (Raspor et al., 2010; Mousa and Raizada, 2013). Similarly, bacterial species of the genera *Pseudomonas* and *Bacillus* are widely spread on flowers, leaves and grapes. Genera *Burkholderia*, *Sphingomonas*, *Serratia* and *Streptococcus* are more present on leaves and grapes. Finally, *Erwinia* spp. are dominant in flowers. These bacterial genera are among the most abundant in grapevine and some of them are well established bacterial and yeast antagonists (De Vleesschauwer and Höfte, 2003; Trotel-Aziz et al., 2008; Elshafie et al., 2012). Besides beneficial microbes, metagenomic analysis also revealed the presence of phytopathogen microorganism living on grapevines, albeit normally with low abundance, like fungal species of the genera *Phomopsis*, *Cryptovalsa*, and *Botryotinia* (Pinto et al., 2014). Keeping a balance between beneficial microorganisms, in terms of abundance and richness, is of crucial importance for grapevine health and biocontrol of pathogens.

Grapevine microbiota is influenced by several factors, including pedoclimatic and biogeographic conditions. The concept of a region-specific microbiota gives strength to the concept of "terroir," on which viticulturists often heavily rely. Not all regions and vineyards are microbiologically unique, but evidences prove the existence of patterns at least on a large scale for microbial communities in grapes and musts. These correlate with climate, soil type and crop management (Bokulich et al., 2016). On the other hand, it is important to stress that studies on the biogeographic structural characterization of the grapevine microbiota have been mainly carried on grape samples. Furthermore, there is a lack of information regarding the role of soil microbiome in defining the terroir at a local scale, as well as the influence exerted by grapevine roots on soil microorganisms and vice versa. A recent study suggests that soil microbiome could represent a potential reservoir, since about 40% of bacterial OTUs obtained from grape, leaf and flower samples are also present in root samples. Moreover, about 48% of prokaryotic OTUs from three different types of belowground samples overlap (Zarraonaindia et al., 2015). Rolli et al. reported that plant growth promoting bacteria (PGPB) of different geographical origins and different crop plants rapidly colonizes the root system of grapevine. This allows the growth of grapevine under different field conditions. However, authors found contrasting results in the literature concerning the effect of PGPB on different plants. This clearly suggests that a more in dept analysis is needed (Rolli et al., 2017).

2. COMPUTATIONAL METHODS FOR MICROBIOME ANALYSIS

Given the recent interest in meta-omics sciences, many computational methodologies are rising to allow the interpretation of the large amount of data produced by high-throughput techniques. The primary purpose of these studies has been the identification of microorganisms present in a environmental sample, determining their activity and interaction with host plant.

Metagenomic methodologies, developed to establish microbiome composition, are divided in two classes: DNA Metabarcoding techniques and Genome Relative Abundance (GRA) estimation techniques.

DNA Metabarcoding aims at identifying a set of *operational taxonomic units* (OTUs) present in a single environmental sample. However, this method requires specific algorithmic techniques capable of handling large amounts of data. QIIME (Quantitative Insights into Microbial Ecology: Caporaso et al., 2010) is a suite of tools combined to define standard pipelines for metabarcoding analysis. It provides a set of analysis and prediction algorithms, along with graphical reports, allowing a simplified analysis of the results. OBITools (Boyer et al., 2016) is another tool enabling analysis from raw sequencing data up to taxon assignment. PRINSEQ (Schmieder and Edwards, 2011) offers functionality similar to QIIME and OBITools through a web interface. Several other tools are available: UPARSE

(Edgar, 2013) aims to detect *de novo* OTUs from NGS reads achieving high accuracy in biological sequence recovery, and improving richness estimate; MOTHUR (Schloss et al., 2009) is a comprehensive software package, which analyzes community sequencing data; DADA2 (Callahan et al., 2016) is a model-based approach to correct amplicon errors without constructing OTUs. See **Table 1** for an overview of described methods.

The main shortcoming of metabarcoding is the classification of OTUs using an existing reference, and the preparation of specific sequencing libraries (Somervuo et al., 2016). Recently, techniques have been developed to detect microbial composition directly from shotgun sequencing. These approaches can be divided into two categories: compositional-based and alignment-based (Xia et al., 2011). In compositional-based approaches, *k*-mer frequency measurements are used to classify metagenomic reads. Methods such as TETRA (Teeling et al., 2004), CompostBin (Chatterji et al., 2008) and TACOA (Diaz et al., 2009) organize sequences in clusters (*k*-mer frequency is used to build a feature vector to compute distances between reads). Next, they assign an unique taxon to each cluster through a set of references computed on known genomes. However, none of these approaches is able to estimate Genomes Relative Abundance (GRAs) for microbial communities. AbundanceBin (Wu and Ye, 2011) uses the content of *k*-mers in the reads to estimate abundance of the genomes. The main assumption in this process is that reads are sampled from genomes following a Poisson distribution. However, detection efficiency decreases when a uniform distribution of species is present in a sample.

Unlike compositional-based algorithms, alignment-based methods use tools such as BLAST to find similarity to a reference species database, while estimating the relative abundance of each genome. MEGAN (Huson et al., 2007) uses BLAST to assign a species to each read, tracing the lowest common ancestor for those with multiple assignments. Then, it estimates the relative abundance using reads distribution normalized by taking into account ambiguous ones. GRAMMy (Genome Relative Abundance using Mixture Models: Xia et al., 2011) uses BLAST to perform an initial assignment of the species to each read. Next, it uses Expectation Maximization (EM) technique to establish probability assignment of the reads, modeling ambiguities, and accurately identifying the relative quantities of each species. See **Table 1** for an overview of described methods.

Although the metagenomic approach can estimate a profile of a sample microbial community, it only allows comparative studies in different conditions, without providing insight on their actual activity (Simon and Daniel, 2011). A complementary view is given by metatranscriptomics. It gives details on the expression profiles and regulation mechanisms in the identified microorganisms (See **Table 3** for an overview of all methods).

This produces details concerning progress and intensity of biological and metabolic processes, elucidating the means by which a microbial community interacts with its host. Several tools can analyze RNA-seq data to extract information about microbial transcriptome. In Leimena et al. (2013), authors propose a comprehensive platform for metatranscriptomics analysis. It allows removal of rRNA sequences, prediction of taxonomic origin and assignment of a function to mRNAs in a sample. HUMAnN2 (Abubucker et al., 2012) can detect the presence, absence, and abundance of microbial pathways through sequencing data. The purpose of the suite is describing the metabolic potential of a microbial community and its members, establishing a functional profile. MetaTrans (Martinez et al., 2016) is an open-source pipeline, which analyzes the structure and function of an active microbial community in a sample. It was developed to analyze large amounts of data produced by sequencing leveraging parallel computing techniques. COMAN (Ni et al., 2016) is a tool for determining the metatranscriptome through an easy-to-use web interface. The primary purpose of the platform is providing tools for quality control, metatranscripts counting, and several statistical analyzes. It can be used in the absence of sufficient computational resources and without any programming expertise, since it is based on a web interface. See **Table 2** for an overview of described methods.

An important limitation of the methods analyzed so far is the use of functional enrichment to establish microbial pathway activity. Recently, a new paradigm has emerged. Indeed, considering both gene expression and their interaction network can lead to more accurate results (Tarca et al., 2008; Alaimo et al., 2016, 2017). These tools have great potential for microbial activity analysis and quantifying microbial interactions with the host, through its pathways.

3. MANIPULATING GRAPEVINE MICROBIOME: FROM *IN SILICO* TO THE FIELD

Traditionally, to overcome the low economic returns caused by Grapevine Trunk Diseases (GTDs), viticulturists treat plants with pesticides. Commercially available microbial inoculants are also available. Most of these inoculants includes individual bacterial or fungal strains, aiming to contrast grapevine pathogens without causing environmental pollution (McSpadden Gardener and Fravel, 2002). However, despite the availability of registered biocontrol products, recently published data report that their adoption in viticulture is still limited mainly due to the belief that they are less effective than traditional pesticides (Gramaje and Di Marco, 2015).

The growing understanding of microbial influence on plants is driving us toward an innovative sustainable viticulture where microbes will replace pesticides, improving grapevine traits. In this new scenario, culture-independent molecular techniques and *in silico* analysis are the new protagonists. Metagenomics, metabarcoding, metatranscriptomics and other molecular approaches applied to healthy, pesticides-treated, and disease-affected grapevines are revealing specific patterns of colonizing microorganisms. This will enable the development of prediction models based on structure and transcriptional profile of organ-specific vine-associated microbiome.

Some recent works focused on GTD-affected vines highlighted the importance of microbial communities influence. In field conditions, plant infectious diseases are rarely due to single host-pathogen interactions. They are the result of simultaneous biotic and abiotic stresses on plants, which induce a sequential

TABLE 1 | Brief description of metabarcoding: tools advantages and disadvantages.

Tool	Bioinformatics tools for metabarcoding		
	Brief description	**Advantages**	**Disadvantages**
QIIME	Pipeline for performing microbiome analysis by exploiting a set of integrated scripts for analyzing raw microbial DNA samples, including taxonomic classification using marker genes.	Allows flexible multi-script pipelines to be constructed. Allows wide statistical analysis with advanced graphical visualizations. Provides compute resources for free.	Command line interface. Installation on local machine may be difficult for non-experts. Not multi-platform.
OBITool	Set of programs specifically designed for analyzing NGS data in a DNA metabarcoding context, designed to target microbial communities from various ecological contexts.	Relies mainly on filtering and sorting algorithms, allowing users to set up flexible data analysis pipelines. It takes into account taxonomic annotations, allowing sorting and filtering of sequence records based on the taxonomy.	Command line interface. Installation on local machine may be difficult for non-experts. Not multi-platform.
PRINSEQ	Projected to trim adapter sequences and low quality ends and to remove the reads containing ambiguous nucleotides and duplicate reads from the sequencing data output, accelerating read data analysis.	User-friendly. Generates complete statistics of data-seq for parameters like sequence length, GC content, quality score and replicates. Capable of treating both single and paired-end reads. Exploitable also for metagenomics and metatranscriptomics data.	Window size needs to be defined by users for the initial trimming step. Limited to pre-processing.
MOTHUR	Principally designed to target the microbial ecology community, it provides an extensible package with functionality accessible through a domain-specific language. It incorporates algorithms from previous tools plus additional features.	Single program for complete analysis with basic visualizations.	Custom command line interface. Incomplete usage of software engineering techniques. Not multi-platform.
DADA2	R package implementing the full amplicon workflow, from filtering to merging of paired-end reads.	Uses a statistical model of amplicon errors to infer sequence variance instead of construct OTUs. Very high accuracy.	Command line interface.

TABLE 2 | Brief description of GRA estimation tools: advantages and disadvantages.

Tool	Bioinformatics tools for Genomes Relative Abundance (GRA) estimation		
	Brief description	**Advantages**	**Disadvantages**
TETRA	Pioneering classifier that uses tetranucleotide-derived z-score correlations to taxonomically classify genomic fragments. Compositional-based.	Provides statistical analysis of tetranucleotide usage patterns in genomic fragments. It works either via a web-service or a stand-alone program.	Accuracy at genus level is reached using long reads (>1 kb). Tends to create multiple clusters for reads originating from highly abundant species when the sample contains multiple species with highly varying levels of abundance.
CompostBin	DNA compositional-based algorithm which adopts a weighted Principal Component Analysis (PCA)-based strategy. Compositional-based.	Reduces the dimensionality of compositional space. Bins raw sequence reads without need for assembly or training.	Accuracy at genus level is reached using long reads (>1 kb). Tends to create multiple clusters for reads originating from highly abundant species when the sample contains multiple species with highly varying levels of abundance.
TACOA	Multi-class taxonomic classifier combining the idea of the k-nearest neighbor with strategies from kernel-based learning. Compositional-based.	Easily installed and run on a desktop computer. Its reference set can be easily updated with newly sequenced genomes.	Accuracy at genus level is reached using long reads (>1 kb).
AbundanceBin	Binning tool, based on the l-tuple content of reads, developed on the assumption that reads are sampled from genomes following a Poisson distribution. Compositional-based.	Capable to return accurate results also when the sequence lengths are very short (∼75 pb).	Binning efficiency decrease in case of samples which tend to have a uniform distribution of species.
MEGAN	Standalone computer program allowing large metagenomic data sets. It uses BLAST or other comparison tools to assign species to each read, and then employs the NCBI taxonomy. Alignment-based.	Allows large data sets to be dissected without the need for assembly or the targeting of specific phylogenetic markers. Provides statistical and graphical output. Computes quantitatively accuracy and specificity.	Uses bit-score of individual hits as the sole parameter for judging significance, thus affecting specificity and accuracy of taxonomic assignments in different scenarios.
GRAMMy	Probabilistic framework developed for GRA. It uses the Mixture Model theory.	Exploitable with mapping, alignment and composition-based tools. Possibility to handle very short reads obtaining accurate results.	Accuracy in estimated abundance decreases in case of closely related microbes whose genomic sequences are highly similar.

TABLE 3 | Brief description of metatranscriptomics tools: advantages and disadvantages.

Tool	Bioinformatics tools for metatranscriptomics		
	Brief description	**Advantages**	**Disadvantages**
HUMAnN2	Pipeline for profiling the presence/absence and activity level of microbial pathways in a community.	Easy to install and extensive documentation and examples. Uses commonly available tools and databases.	Command line interface.
MetaTrans	Pipeline aiming to analyze structure and functions of active microbial communities using the power of multi-threading computers.	Its design facilitates the inclusion of third-party tools in each of its stages. Possibility to perform RNA-Seq analyses addressing both 16S rRNA taxonomy and gene expression.	Installation on local computer may be difficult for non-experts. Require proper local setup on a powerful computer.
COMAN	Web-based tool dedicated to automatically and comprehensively analyzing metatranscriptomic data.	Easy-to-use interface and extensive instructions for non-experts. Processes uploaded raw reads automatically to ultimately achieve functional assignments, which are then exploited to perform further analysis.	Web-based interface not suitable for big analysis.

colonization process of the host tissue (Travadon et al., 2016; Song et al., 2017). Beneficial microbial communities work as a barrier defending against plant pathogens. This reduces the potential of pathogens invasiveness, since a significant fraction of their niche overlaps (Wei et al., 2015).

In *Esca*-affected vines the fungal community structure undergoes a considerable change in comparison with healthy plants (Morales-Cruz et al., 2017). In apparently healthy vines, very low fungal counts were recorded. *Phaeomoniella chlamydospora* and *Phaeoacremonium minimum* (two Esca pathogens) appeared to be relatively more abundant than other taxa. The most variable fungal composition was reported in vines with wood symptoms but no foliar symptoms, with a generalized notable increase in abundance and activity of pathogenic fungal taxa. Furthermore, *P. chlamydospora* and *P. minimum* were reported to be highly predominant in wood together with *Diaporthe ampelina* (causal agents of *Botryosphaeria dieback*). No significant changes were reported for the bacterial community, since the nine most abundant species belonged to the genera *Bacillus* and *Pantoea*. However, a deeper analysis of the bacterial community is required, especially from a functional point of view (Bruez et al., 2015). For instance, the antagonistic activity of two microbial strains of *Bacillus pumilus* and *Paenibacillus* sp. against *P. chlamydospora* was recently tested *in vitro*. These two strains can synthesize volatile compounds with antifungal activity. Furthermore, *B. pumilus* inoculation confers a systemic resistance in grapevine (Haidar et al., 2016). Such experiments indicate that pathogen detection methods aiming to differentiate between the early and late stages of infection should be quantitative (Morales-Cruz et al., 2017). Therefore, the usage of metatranscriptomics tools, in combination with pathway analysis, of healthy and affected plants, might elucidate the functional relationships between microbial communities, leading to the discovery of novel interactions.

The fungal pathogen *Eutypa lata* is predominant in wood tissues of *Eutypa dieback* affected vines, followed by high abundances of *Diplodia seriata* and *Phaeomoniella chlamydospora* (Morales-Cruz et al., 2017). No information is available concerning the bacterial community changes with respect to this disease. Yet, it was shown that in crown gall affected grapevines, changes in bacterial community is site-specific since a shift in composition happened only in graft unions. *Agrobacterium vitis* is the infectious agent causing the disease (Faist et al., 2016). Authors recorded that the difference in microbial community composition was due to nine bacterial species. The most abundant ones were *A. vitis*, *Pseudomonas* sp. and *Enterobacteriaceae* sp. However, it was determined that the induction of this disease by *A. vitis* do not necessarily requires a core microbiome. Morales-Cruz et al. (2017) detected some pathogenic fungi also in asymptomatic samples, especially *P. chlamydospora* and *P. minimum* with a ratio of *1:200* in respect to GTD-affected samples. Therefore, a more in-depth metagenomic analysis is needed in order to elucidate the composition of the microbial community, with a greater effort on pathological strains.

Metatranscriptomic analysis and functional profiling also helps identifying biological processes linked to specific conditions. For example, Morales-Cruz et al. (2017) showed that most virulence-related expressed genes belonged to carbohydrate active enzymes and transporters, followed by genes related to secondary metabolism, cytochrome P450s and peroxidases. In addition, authors demonstrated that it is possible to distinguish the *Esca* pathogenic functional profile from *Eutypa dieback* one. A recent metabolomics experiment also showed indirectly that inoculation of specific endophytes strains causes a shift in grapevine secondary metabolism, and activation of defense pathways. Furthermore, it was confirmed the existence of strain-specific colonization patterns (Lòpez-Fernàndez et al., 2016).

Transcriptomic analysis of grapevine leaves and wood tissues revealed differentially expressed genes linked to latent *Neofusicoccum parvum* grapevine infection (Czemmel et al., 2015). However, this information is still largely incomplete.

In accordance with Busby et al. (2017), research should have different objectives to enable a reasoned, conscious and effective microbial exploitation, taking into account the impact of plant protection products on the quality of production and human health. In this direction, computational tools could be exploited

to detect interactions among genes, pathogens and treatments, leading to a greater insight on their possible use and effects. Recent review on such methods is available in Lotfi Shahreza et al. (2017).

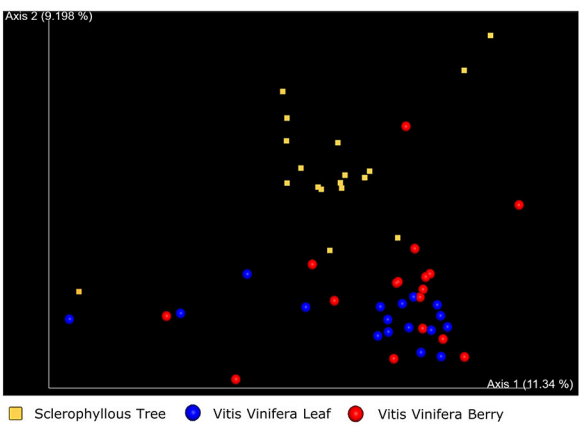
4. A CASE STUDY: METABARCODING ANALYSIS OF GRAPEVINE IN CHILE

In order to show the power of bioinformatics metagenomic analysis in the study of grapevine, we developed a case study from 16S rRNA sequencing data from vineyards and adjacent forest areas in the Chilean territory (SRA Project: SRP110820; Miura et al. 2017). The study analyzed 6 vines from three Chilean geographic areas, and sclerophyllous trees from the adjacent forest area. Vines were sampled from both leaves and fruits. All analysis were replicated 3 times, leading to 54 samples, of which 36 grapevine and 18 controls. 2 samples were discarded due to quality issues. All analysis were conducted using QIIME 2 release 2017.9.

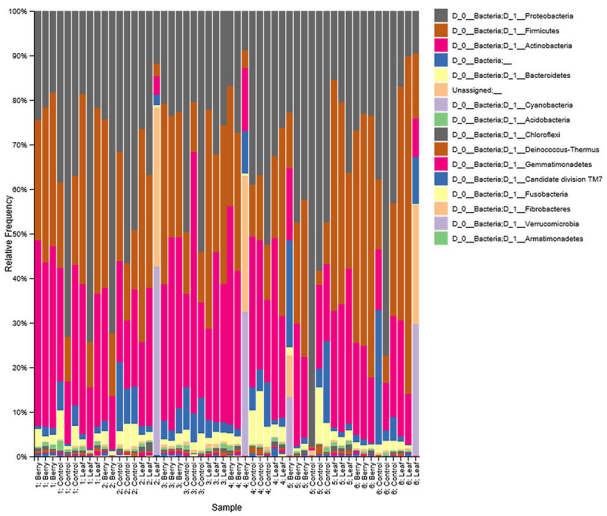
Paired-end Raw Illumina fastq files downloaded from SRA (SRP110820) were demultiplexed, quality filtered, and analyzed using QIIME. Reads containing one or more ambiguous base calls were discarded and truncated to a length of 200 nt. A subsequent filtering phase was performed using Deblur (Amir et al., 2017) to obtain putative error-free sequences from the original data. A phylogenetic tree was therefore built by making use of MAFFT (Katoh and Standley, 2013) and FastTree 2 (Price et al., 2010) allowing computation of subsequent diversity metrics.

Alpha-diversity (within-sample species richness) and beta-diversity (between-sample community dissimilarity) estimates were calculated within QIIME using weighted UniFrac (Lozupone and Knight, 2005) distance between samples for bacterial 16S rRNA reads (evenly sampled at 1,000 reads per sample). Principal coordinates were computed from the resulting distance matrices to compress dimensionality intro 2D principal coordinate analysis (PCoA) plots, enabling visualization of sample relationships (**Figure 1A**). To determine whether sample classifications (host, sample site) contained

A



B



C

Taxonomic Group	Berry	Leaf	Control
Bacteria - Bacteroidetes - Cytophagia - Cytophagales - Cytophagaceae - Hymenobacter	3	5	19
Bacteria - Proteobacteria - Alphaproteobacteria - Rhizobiales - uncultured bacterium	3	2	26,5
Bacteria - Proteobacteria - Alphaproteobacteria - Rhizobiales	1	1	7
Bacteria - Proteobacteria - Alphaproteobacteria - Rhodospirillales - Acetobacteraceae	9	9	38
Bacteria - Proteobacteria	6	5	54

D

Taxonomic Group	Sclerophyllous tree	Vitis vinifera
Bacteria - Bacteroidetes - Cytophagia - Cytophagales - Cytophagaceae - Hymenobacter	19	4
Bacteria - Bacteroidetes	5	1
Bacteria - Proteobacteria - Alphaproteobacteria - Rhizobiales - uncultured bacterium	26,5	3
Bacteria - Proteobacteria - Alphaproteobacteria - Rhizobiales	7	1
Bacteria - Proteobacteria - Alphaproteobacteria - Rhodospirillales - Acetobacteraceae	38	9
Bacteria - Proteobacteria - Deltaproteobacteria	12	2
Bacteria - Proteobacteria	54	5

FIGURE 1 | Results of the metabarcoding analysis of grapevine in Chile. The study analyzes 6 vines from three Chilean geographic areas, and sclerophyllous trees from the adjacent forest area. Vines are sampled from both leaves and fruits. All analysis are replicated 3 times, leading to 54 samples, of which 36 grapevine and 18 controls. 2 samples are discarded due to quality issues. Principal coordinates analysis plot is reported to visualize relationships between samples **(A)**. We also show the relative abundance of operational taxonomic units, assigned using QIIME feature classifier, at phylum level **(B)**, together with the median abundance values obtained for the most significant species in the differential analysis between sampling sites **(C)** and plant species **(D)**.

differences in phylogenetic or species diversity, permutational MANOVA with 1,000 permutations was used to test significant differences between sample groups based on weighted UniFrac. No significant differences could be found in terms of alpha-diversity between hosts ($p = 0.18$) and sample sites ($p = 0.41$), and beta-diversity between sample sites ($p = 0.13$), however a significant difference in beta-diversity could be observed between host species ($p = 0.03$).

OTUs were assigned using QIIME feature classifier, which employs a Naive Bayes classifier to map each sequence to a taxonomy. The classifier was trained on a qiime-compatible Silva database (release 119), which includes sequences from 16S/18S rRNA. Any OTU representing less than 0.001% of the total filtered sequences was removed to avoid inclusion of erroneous reads, leading to inflated estimates of diversity. In **Figure 1B** we report the relative frequency of OTUs for each sample, sorted by site and host. Significant taxonomic differences between sample conditions were tested using ANCOM (Mandal et al., 2015). All results are available in **Figures 1C,D**.

Results shown in **Figure 1B** are consistent with Miura et al.. We are able to retrieve the three main bacterial phyla (Actinobacteria, Firmicutes and Proteobacteria), and relative abundances are consistent. Furthermore, grapevine-related bacterial communities are similar to the phyllosphere of sclerophyllous trees when OTUs clustering is carried out at high bacterial taxonomic levels. This is not surprising since plants are known to show similar pattern at phylum level (Turner et al., 2013). Nonetheless, several differences can be detected at the genus, species or strain level. It is possible to make distinction between bacterial communities living on different plant species. This reflects the finely tuned metabolic adaptations required to live in symbiosis with the host (**Figure 1D**), but also between microbial communities living on different organs, reflecting the adaptations required to live in environment characterized by certain microclimatic conditions (**Figure 1C**).

5. CONCLUSIONS AND PERSPECTIVES

Gaining wider knowledge about plant-microbiota interactions at a molecular scale is an urgent task, since it can lead to the development of new biotechnological approaches, enhancing agriculture productivity and sustainability. NGS technologies together with bioinformatics are fundamental tools in this process. They have the potential to reveal new details concerning interactions between microbial communities and plants with unprecedented resolution. The characterization of microbial communities in grapevines at various conditions using high throughput sequencing technologies will sooner lead to the identification of disease-specific, cultivar-specific and climate-specific grapevine microbiota. Their exploitation in the field of viticulture could lead to the discovery of new applicable microbial strains, and could help us gaining a holistic and more complete view of plant-microbiome interactions at a genetic level. This would allow not only to potentiate antagonisms against phytopathogens and crop yield, but also to positively affect economically important traits, such as flowering time, and quality and flavor of grapes, must and wine.

AUTHOR CONTRIBUTIONS

AP conceived developed and coordinated the research. AP, SA, AF, and RG analyzed the data. SA and GM wrote the paper. All authors read and approved the final version of the manuscript.

FUNDING

Sviluppo Regionale (PO-FESR 2007-2013), Linea di intervento 4.1.1.2. Grant number: CUP G23F11000840004.

REFERENCES

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Adam-Blondon, A., Jaillon, O., Vezzulli, S., Zharkikh, A., Troggio, M., Velasco, R., et al. (2011). "Genome sequence initiatives," in *Genetics, Genomics and Breeding of Grapes*, eds. A.-F. Adam-Blondon, J. M. Martínez-Zapater, C. Kole (Enfield, NH: Science Publishers), 211–234. doi: 10.1201/b10948-10

Adam-Blondon, A.-F., Bernole, A., Faes, G., Lamoureux, D., Pateyron, S., Grando, M., et al. (2005). Construction and characterization of bac libraries from major grapevine cultivars. *Theor. Appl. Genet.* 110, 1363–1371. doi: 10.1007/s00122-005-1924-9

Alaimo, S., Giugno, R., Acunzo, M., Veneziano, D., Ferro, A., and Pulvirenti, A. (2016). Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget* 7:54572. doi: 10.18632/oncotarget.9788

Alaimo, S., Marceca, G. P., Ferro, A., and Pulvirenti, A. (2017). Detecting disease specific pathway substructures through an integrated systems biology approach. *Non-Coding RNA* 3:20. doi: 10.3390/ncrna3020020

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. doi: 10.1128/mBio.00631-16

Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., et al. (2016). Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *MBio* 7:e00631-16.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., and Coissac, E. (2016). obitools: a unix-inspired software package for dna metabarcoding. *Mol. Ecol. Resources* 16, 176–182. doi: 10.1111/1755-0998.12428

Bruez, E., Haidar, R., Alou, M. T., Vallance, J., Bertsch, C., Mazet, F., et al. (2015). Bacteria in a wood fungal disease: characterization of bacterial communities in wood tissues of esca-foliar symptomatic and asymptomatic grapevines. *Front. Microbiol.* 6:1137. doi: 10.3389/fmicb.2015.01137

Busby, P. E., Soman, C., Wagner, M. R., Friesen, M. L., Kremer, J., Bennett, A., et al. (2017). Research priorities for harnessing plant microbiomes in sustainable agriculture. *PLoS Biol.* 15:e2001793. doi: 10.1371/journal.pbio.2001793

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). Dada2: high-resolution sample inference from illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Campisano, A., Antonielli, L., Pancher, M., Yousaf, S., Pindo, M., and Pertot, I. (2014). Bacterial endophytic communities in the grapevine depend on pest management. *PLoS ONE* 9:e112763. doi: 10.1371/journal.pone.0112763

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Chatterji, S., Yamazaki, I., Bai, Z., and Eisen, J. (2008). "Compostbin: a dna composition-based algorithm for binning environmental shotgun reads," in *Research in Computational Molecular Biology*, eds M. Vingron and L. Wong (Singapore: Springer), 17–28.

Czemmel, S., Galarneau, E. R., Travadon, R., McElrone, A. J., Cramer, G. R., and Baumgartner, K. (2015). Genes expressed in grapevine leaves reveal latent wood infection by the fungal pathogen neofusicoccum parvum. *PLoS ONE* 10:e0121828. doi: 10.1371/journal.pone.0121828

De Vleesschauwer, D., and Höfte, M. (2003). Using *serratia plymuthica* to control fungal pathogens of plants. *CAB Rev.* 2, 1–12. doi: 10.1079/PAVSNNR20072046

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., and Nattkemper, T. W. (2009). Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* 10:56. doi: 10.1186/1471-2105-10-56

Edgar, R. C. (2013). Uparse: highly accurate otu sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Elshafie, H. S., Camele, I., Racioppi, R., Scrano, L., Iacobellis, N. S., and Bufo, S. A. (2012). *In vitro* antifungal activity of burkholderia gladioli pv. agaricicola against some phytopathogenic fungi. *Int. J. Mol. Sci.* 13, 16291–16302. doi: 10.3390/ijms131216291

Faist, H., Keller, A., Hentschel, U., and Deeken, R. (2016). Grapevine (*Vitis vinifera*) crown galls host distinct microbiota. *Appl. Environ. Microbiol.* 82, 5542–5552. doi: 10.1128/AEM.01131-16

Fasoli, M., Dal Santo, S., Zenoni, S., Tornielli, G., Farina, L., Zamboni, A., et al. (2012). The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell* 24, 3489–3505. doi: 10.1105/tpc.112.100230

Fortes, A. M., Agudelo-Romero, P., Silva, M. S., Ali, K., Sousa, L., Maltese, F., et al. (2011). Transcript and metabolite analysis in trincadeira cultivar reveals novel information regarding the dynamics of grape ripening. *BMC Plant Biol.* 11:149. doi: 10.1186/1471-2229-11-149

Gramaje, D., and Di Marco, S. (2015). Identifying practices likely to have impacts on grapevine trunk disease infections: a european nursery survey. *Phytopathol. Mediter.* 54:313. doi: 10.14601/Phytopathol_Mediterr-16317

Grimplet, J., Wheatley, M. D., Jouira, H. B., Deluc, L. G., Cramer, G. R., and Cushman, J. C. (2009). Proteomic and selected metabolite analysis of grape berry tissues under well-watered and water-deficit stress conditions. *Proteomics* 9, 2503–2528. doi: 10.1002/pmic.200800158

Guillaumie, S., Fouquet, R., Kappel, C., Camps, C., Terrier, N., Moncomble, D., et al. (2011). Transcriptional analysis of late ripening stages of grapevine berry. *BMC Plant Biol.* 11:165. doi: 10.1186/1471-2229-11-165

Haidar, R., Roudet, J., Bonnard, O., Dufour, M. C., Corio-Costet, M. F., Fert, M., et al. (2016). Screening and modes of action of antagonistic bacteria to control the fungal pathogen phaeomoniella chlamydospora involved in grapevine trunk diseases. *Microbiol. Res.* 192, 172–184. doi: 10.1016/j.micres.2016.07.003

Holland, T. C., Bowen, P., Bogdanoff, C., and Hart, M. M. (2014). How distinct are arbuscular mycorrhizal fungal communities associating with grapevines? *Biol. Fertil. Soils* 50, 667–674. doi: 10.1007/s00374-013-0887-2

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). Megan analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Jackson, R. (1994). Grapevine species and varieties. *Wine Sci. Principl. Appl.* 11–31.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Katoh, K., and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Lamoureux, D., Bernole, A., Le Clainche, I., Tual, S., Thareau, V., Paillard, S., et al. (2006). Anchoring of a large set of markers onto a bac library for the development of a draft physical map of the grapevine genome. *Theor. Appl. Genet.* 113, 344–356. doi: 10.1007/s00122-006-0301-7

Leimena, M. M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E. J., et al. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genom.* 14:530. doi: 10.1186/1471-2164-14-530

Lijavetzky, D., Carbonell-Bejerano, P., Grimplet, J., Bravo, G., Flores, P., Fenoll, J., et al. (2012). Berry flesh and skin ripening features in *Vitis vinifera* as assessed by transcriptional profiling. *PLoS ONE* 7:e39547. doi: 10.1371/journal.pone.0039547

Lòpez-Fernàndez, S., Compant, S., Vrhovsek, U., Bianchedi, P. L., Sessitsch, A., Pertot, I., et al. (2016). Grapevine colonization by endophytic bacteria shifts secondary metabolism and suggests activation of defense pathways. *Plant Soil* 405, 155–175. doi: 10.1007/s11104-015-2631-1

Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J., and Green, J. R. (2017). A review of network-based approaches to drug repositioning. *Brief. Bioinformat.* doi: 10.1093/bib/bbx017. [Epub ahead of print].

Lozupone, C., and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26:27663. doi: 10.3402/mehd.v26.27663

Martinez, X., Pozuelo, M., Pascal, V., Campos, D., Gut, I., Gut, M., et al. (2016). Metatrans: an open-source pipeline for metatranscriptomics. *Sci. Rep.* 6:26447. doi: 10.1038/srep26447

Martins, G., Lauga, B., Miot-Sertier, C., Mercier, A., Lonvaud, A., Soulas, M.-L., et al. (2013). Characterization of epiphytic bacterial communities from grapes, leaves, bark and soil of grapevine plants grown, and their relations. *PLoS ONE* 8:e73013. doi: 10.1371/journal.pone.0073013

McSpadden Gardener, B., and Fravel, D. (2002). Biological control of plant pathogens: research, commercialization, and application in the USA. *Plant Health Prog.* 10, 207–209. doi: 10.1094/PHP-2002-0510-01-RV

Miura, T., Sánchez, R., Castañeda, L. E., Godoy, K., and Barbosa, O. (2017). Is microbial terroir related to geographic distance between vineyards? *Environ. Microbiol. Rep.* 9, 742–749. doi: 10.1111/1758-2229.12589

Morales-Cruz, A., Allenbeck, G., Figueroa-Balderas, R., Ashworth, V. E., Lawrence, D. P., Travadon, R., et al. (2017). Closed-reference metatranscriptomics enables *in planta* profiling of putative virulence activities in the grapevine trunk disease complex. *Mol. Plant Pathol.* doi: 10.1111/mpp.12544. [Epub ahead of print].

Mousa, W. K., and Raizada, M. N. (2013). The diversity of anti-microbial secondary metabolites produced by fungal endophytes: an interdisciplinary perspective. *Front. Microbiol.* 4:65. doi: 10.3389/fmicb.2013.00065

Mullins, M. G., Bouquet, A., and Williams, L. E. (1992). *Biology of the Grapevine*. Cambridge University Press.

Ni, Y., Li, J., and Panagiotou, G. (2016). Coman: a web server for comprehensive metatranscriptomics analysis. *BMC Genom.* 17:622. doi: 10.1186/s12864-016-2964-z

Pancher, M., Ceol, M., Corneo, P. E., Longa, C. M. O., Yousaf, S., Pertot, I., et al. (2012). Fungal endophytic communities in grapevines (*Vitis vinifera* l.) respond to crop management. *Appl. Environ. Microbiol.* 78, 4308–4317. doi: 10.1128/AEM.07655-11

Pieterse, C. M., de Jonge, R., and Berendsen, R. L. (2016). The soil-borne supremacy. *Trends Plant Sci.* 21, 171–173. doi: 10.1016/j.tplants.2016.01.018

Pinto, C., Pinho, D., Sousa, S., Pinheiro, M., Egas, C., and Gomes, A. C. (2014). Unravelling the diversity of grapevine microbiome. *PLoS ONE* 9:e85622. doi: 10.1371/journal.pone.0085622

Polesani, M., Bortesi, L., Ferrarini, A., Zamboni, A., Fasoli, M., Zadra, C., et al. (2010). General and species-specific transcriptional responses to downy mildew infection in a susceptible (*Vitis vinifera*) and a resistant (*V. riparia*) grapevine species. *BMC Genom.* 11:117. doi: 10.1186/1471-2164-11-117

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Pulvirenti, A., Giugno, R., Distefano, R., Pigola, G., Mongiovi, M., Giudice, G., et al. (2015). A knowledge base for *Vitis vinifera* functional analysis. *BMC Syst. Biol.* 9:S5. doi: 10.1186/1752-0509-9-S3-S5

Raspor, P., Miklič-Milek, D., Avbelj, M., and Čadež, N. (2010). Biocontrol of grey mould disease on grape caused by botrytis cinerea with autochthonous wine yeasts. *Food Technol. Biotechnol.* 48, 336–343.

Rolli, E., Marasco, R., Saderi, S., Corretto, E., Mapelli, F., Cherif, A., et al. (2017). Root-associated bacteria promote grapevine growth: from the laboratory to the field. *Plant Soil* 410, 369–382. doi: 10.1007/s11104-016-3019-6

Rouphael, Y., Franken, P., Schneider, C., Schwarz, D., Giovannetti, M., Agnolucci, M., et al. (2015). Arbuscular mycorrhizal fungi act as biostimulants in horticultural crops. *Sci. Horticult.* 196, 91–108. doi: 10.1016/j.scienta.2015.09.002

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Simon, C., and Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* 77, 1153–1161. doi: 10.1128/AEM.02345-10

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., and Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for dna barcoding. *Bioinformatics* 32, 2920–2927. doi: 10.1093/bioinformatics/btw346

Song, Z., Kennedy, P. G., Liew, F. J., and Schilling, J. S. (2017). Fungal endophytes as priority colonizers initiating wood decomposition. *Funct. Ecol.* 31, 407–418. doi: 10.1111/1365-2435.12735

Sreekantan, L., Mathiason, K., Grimplet, J., Schlauch, K., Dickerson, J. A., and Fennell, A. Y. (2010). Differential floral development and gene expression in grapevines during long and short photoperiods suggests a role for floral genes in dormancy transitioning. *Plant Mol. Biol.* 73, 191–205. doi: 10.1007/s11103-010-9611-x

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., et al. (2008). A novel signaling pathway impact analysis. *Bioinformatics* 25, 75–82. doi: 10.1093/bioinformatics/btn577

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. O. (2004). Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics* 5:163. doi: 10.1186/1471-2105-5-163

Tillett, R. L., Ergül, A., Albion, R. L., Schlauch, K. A., Cramer, G. R., and Cushman, J. C. (2011). Identification of tissue-specific, abiotic stress-responsive gene expression patterns in wine grape (*Vitis vinifera* l.) based on curation and mining of large-scale est data sets. *BMC Plant Biol.* 11:86. doi: 10.1186/1471-2229-11-86

Tomkins, J. P., Peterson, D. G., Yang, T.-J., Main, D., Ablett, E., Henry, R. J., et al. (2001). Grape (*Vitis vinifera* l.) bac library construction, preliminary stc analysis, and identification of clones associated with flavonoid and stilbene biosynthesis. *Am. J. Enol. Viticult.* 52, 287–291.

Travadon, R., Lecomte, P., Diarra, B., Lawrence, D. P., Renault, D., Ojeda, H., et al. (2016). Grapevine pruning systems and cultivars influence the diversity of wood-colonizing fungi. *Fungal Ecol.* 24, 82–93. doi: 10.1016/j.funeco.2016.09.003

Trotel-Aziz, P., Couderchet, M., Biagianti, S., and Aziz, A. (2008). Characterization of new bacterial biocontrol agents acinetobacter, bacillus, pantoea and pseudomonas spp. mediating grapevine resistance against botrytis cinerea. *Environ. Exp. Bot* 64, 21–32. doi: 10.1016/j.envexpbot.2007.12.009

Turner, T. R., James, E. K., and Poole, P. S. (2013). The plant microbiome. *Genome Biol.* 14:209. doi: 10.1186/gb-2013-14-6-209

Vandenkoornhuyse, P., Quaiser, A., Duhamel, M., Le Van, A., and Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. *New Phytol.* 206, 1196–1206. doi: 10.1111/nph.13312

Vivier, M. A., and Pretorius, I. S. (2002). Genetically tailored grapevines for the wine industry. *Trends Biotechnol.* 20, 472–478. doi: 10.1016/S0167-7799(02)02058-9

Vivier, M. A., and Pretorius, I. S. (2000). Genetic improvement of grapevine: tailoring grape varieties for the third millennium - a review. *South Afr. J. Enol. Viticult.* 21, 5–26.

Wei, Z., Yang, T., Friman, V.-P., Xu, Y., Shen, Q., and Jousset, A. (2015). Trophic network architecture of root-associated bacterial communities determines pathogen invasion and plant health. *Nat. Commun.* 6:8413. doi: 10.1038/ncomms9413

Wu, Y.-W., and Ye, Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J. Comput. Biol.* 18, 523–534. doi: 10.1089/cmb.2010.0245

Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A., and Sun, F. (2011). Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS ONE* 6:e27992. doi: 10.1371/journal.pone.0027992

Zarraonaindia, I., Owens, S. M., Weisenhorn, P., West, K., Hampton-Marcell, J., Lax, S., et al. (2015). The soil microbiome influences grapevine-associated microbiota. *MBio* 6:e02527-14. doi: 10.1128/mBio.02527-14

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 and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership