USER-FRIENDLY TOOLS APPLIED TO GENETICS OR SYSTEMS BIOLOGY

EDITED BY: Helder Nakaya, Juilee Thakar and Vinicius Maracaja-Coutinho
PUBLISHED IN: Frontiers in Genetics and Frontiers in Physiology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-133-6
DOI 10.3389/978-2-88966-133-6

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

USER-FRIENDLY TOOLS APPLIED TO GENETICS OR SYSTEMS BIOLOGY

Topic Editors:
Helder Nakaya, University of São Paulo, Brazil
Juilee Thakar, University of Rochester, United States
Vinicius Maracaja-Coutinho, University of Chile, Chile

Citation: Nakaya, H., Thakar, J., Maracaja-Coutinho, V., eds. (2020). User-Friendly Tools Applied to Genetics or Systems Biology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-133-6

Table of Contents

frontiers in Genetics

EDITORIAL
published: 09 September 2020
doi: 10.3389/fgene.2020.00985

Editorial: User-Friendly Tools Applied to Genetics or Systems Biology

Helder I. Nakaya[1]*, *Juilee Thakar*[2,3]* and *Vinicius Maracaja-Coutinho*[4]*

[1] Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil, [2] Department of Microbiology & Immunology, University of Rochester, Rochester, NY, United States, [3] Department of Briostatistics & Computational Biology, University of Rochester, Rochester, NY, United States, [4] Advanced Center for Chronic Diseases – ACCDiS, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile

Keywords: user-friendly tools, systems biology, omics analyses, bioinformatics and computational biology, computational tool and servers

Editorial on the Research Topic

User-Friendly Tools Applied to Genetics or Systems Biology

Life scientists now have access to an unprecedented amount of experimental data. A single laboratory can measure the levels of all transcripts, proteins, or metabolites of an organism under different perturbations or can sequence the entire genome of hundreds of individuals or specimens. Systems biology aims to study the behavior and interaction of these molecules, using advanced mathematical models. Modern data-intensive genetics is also often dependent on statistical tools for identifying signals through population-level measurements. However, according to Sydney Brenner "we are drowning in a sea of data and starving for knowledge. Today, biology is more about gathering data than hunting down new ideas." This is partly due to the fact that a substantial number of researchers who are capable of thinking about new insights, are not able to deal with the vast amounts of data generated by modern technologies. This Research Topic aimed to help those researchers interested in analyzing high-throughput data, but lacking knowledge on programming languages or bioinformatics skills. With the collaboration of computer scientists and software developers, this issue brings an interesting collection of user-friendly tools with broad applications in genetics and systems biology.

As the first layer of biological information, the DNA carries the genetic instructions for the fine-tuning functioning of all known organisms. In this context, Sariya et al. performed a benchmarking of reference panels and tools for rare variants imputation in genome-wide association studies (GWAS) in admixed populations. Thus, Sariya et al. study will facilitate the selection of panels, tools, and parameters for rare variant imputation in GWAS. Related to prokaryotic genomes, two user-friendly tools were described with the purpose of performing comparative genomics analyses focused on Bacteria. Gene Tags Assessment by Comparative Genomics (GTACG) performs pan-genome comparative analyses (Santiago et al.) by identifying homologous genes and defining the gene families, followed by the documentation of the core/accessory genome, phylogenetic analysis and data visualization in an easy-to-use graphic interface. PhageWeb is web service for identifying prophage regions and for characterizing bacterial genomes (de Sousa et al.).

The analysis of the transcriptome, i.e., the set of all transcripts of a cell or tissue, provides an overview of the processes and signaling pathways related to diseases and various biological conditions. Four user-friendly computational tools described here (Pipeliner, ABioTrans, Simplicity DiffExpress, and MDP), facilitate the processing and analysis of RNA-seq data. By combining the Anaconda package manager with Nextflow scripting language, Pipeliner enable users to generate modular computational workflows for processing various types of sequencing data, including single-cell expression data (Federico et al.). ABioTrans is a web-browser based

OPEN ACCESS

Edited and reviewed by:
Maximino Aldana, National Autonomous University of Mexico, Mexico

***Correspondence:**
Helder I. Nakaya
hnakaya@usp.br
Juilee Thakar
juilee_thakar@urmc.rochester.edu
Vinicius Maracaja-Coutinho
vinicius.maracaja@uchile.cl

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

Received: *15 July 2020*
Accepted: *05 August 2020*
Published: *09 September 2020*

Citation:
Nakaya HI, Thakar J and Maracaja-Coutinho V (2020) Editorial: User-Friendly Tools Applied to Genetics or Systems Biology. Front. Genet. 11:985. doi: 10.3389/fgene.2020.00985

user interface that allows users to not only directly read RNA-Seq data files deposited in the GEO database, but also to perform dimensionality reduction, differential expression analysis, and gene ontology classifications (Zou et al.). After the raw RNA-seq data is summarized in read-count tables, Simplicity DiffExpress can be used to run differential expression analysis and to determine a bespoke statistical model validation for it (Palu et al.). Often, however, the huge heterogeneity among individuals can impact gene expression analyses. MDP webtool uses a dynamic interface to inspect gene expression data and identify samples that are potential biological outliers (Gonçalves et al.). It is also useful to identify subgroups of patients classified with a particular disease but with different expression profiles or to reveal particularities of distinct illness that are perturbing the expression of genes or pathways.

The integration of the set of transcripts, proteins or metabolites in a particular condition using network approaches allows analysis beyond just one genes/protein. The webCEMiTool provides an easy-to-use environment for identifying gene co-expression modules, followed by their functional characterization through the automatic integration of gene-to-gene or protein-protein interaction networks, gene set enrichment analysis and overrepresentation of pathways or ontologies (Cardozo et al.). The FindTargetsWEB focused on analyzing genome-scale metabolic networks of bacteria in order to identify potential therapeutic targets (Merigueti et al.). It searches for fragile genes available in the network, in which its blockage could impair one or more metabolic functions.

BioNetStat provides a user-friendly environment for the comparison of two or more networks simultaneously, by exploring different topological features available in each network (Jardim et al.). The review from Ramos et al. provides a very interesting and intuitive explanation of the key concepts and terminology behind network biology, as well as a didactic guide on how to perform network analysis using user-friendly tools.

After integration, users can develop or simulate biological models representing the living system of the studied organism in particular conditions of interest. In this context, Afshar et al. generated a model in CellML format for glucose uptake in the epithelial cell of the small intestine (enterocyte). This model structure permits different changes in the components and parameters, facilitating its reuse and customization. Ii et al. developed a tool, named XitoSBML, that helps the users to automatically generate Systems Biology Markup Language (SBML) Level 3 Version 1 spatial model files from microscopic cellular images (Ii et al.). The converted model holds molecular concentrations, locations and biochemical reactions, which can be used by SBML-supported simulators to perform spatial simulations based on the generated model.

Finally, after walking from genomes to systems/networks and the computational modeling of living systems, the user might be interested in store or explore its information in biological databases. In this Research Topic two databases for genetics and systems biology data organization were described. croFGD integrated genomic information and dozens of RNA sequencing data from different tissues and biological conditions of *Catharanthus roseus*, a medicinal plant with pharmacological activities, in order to build a functional genomics database (She et al.). It provides annotations, expression data, and network models (e.g., co-expression, protein-protein interactions, microRNA-target interactions), which can be explored dynamically through a web searchable interface and a set of tools for data analyses specifically for this species. Ma et al. developed a similar database for Moso bamboo (*Phyllostachys edulis*), the most economically valuable bamboo in Asia, called BambooNET.

A fundamental characteristic for a tool to be adopted widely by life scientists is that it should be user-friendly. Even if the application is specific to a small area of knowledge, the software needs to be easy to run by researchers without advanced knowledge in programming or statistical tools. As a trade-off, user-friendly versions generally have fewer parameters and adjustments than versions which are run on scripts or command lines. We hope that the reader will find a useful collection of such tools for genetic or systems biology research, democratizing bioinformatics and computational biology to a broad group of users with lesser computing background.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was financed by FAPESP (2018/14933-2) and ANID FONDAP initiative (grant number 15130011).

Conflict of Interest: 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: 27 November 2018
doi: 10.3389/fgene.2018.00574

Co-expression Gene Network Analysis and Functional Module Identification in Bamboo Growth and Development

Xuelian Ma[1†], Hansheng Zhao[2†], Wenying Xu[1], Qi You[1], Hengyu Yan[1], Zhimin Gao[2] and Zhen Su[1*]*

[1] State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing, China, [2] State Forestry Administration Key Open Laboratory on the Science and Technology of Bamboo and Rattan, Institute of Gene Science for Bamboo and Rattan Resources, International Center for Bamboo and Rattan, Beijing, China

OPEN ACCESS

Edited by:
Vinicius Maracaja-Coutinho, Universidad de Chile, Chile

Reviewed by:
Hamed Bostan, North Carolina State University, United States
Dinesh Kumar, Indian Council of Agricultural Research (ICAR), India

***Correspondence:**
Zhimin Gao
gaozhimin@icbr.ac.cn
Zhen Su
zhensu@cau.edu.cn

[†] 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: 21 July 2018
Accepted: 08 November 2018
Published: 27 November 2018

Citation:
Ma X, Zhao H, Xu W, You Q, Yan H, Gao Z and Su Z (2018) Co-expression Gene Network Analysis and Functional Module Identification in Bamboo Growth and Development. Front. Genet. 9:574. doi: 10.3389/fgene.2018.00574

Bamboo is one of the fastest-growing non-timber forest plants. Moso bamboo (*Phyllostachys edulis*) is the most economically valuable bamboo in Asia, especially in China. With the release of the whole-genome sequence of moso bamboo, there are increasing demands for refined annotation of bamboo genes. Recently, large amounts of bamboo transcriptome data have become available, including data on the multiple growth stages of tissues. It is now feasible for us to construct co-expression networks to improve bamboo gene annotation and reveal the relationships between gene expression and growth traits. We integrated the genome sequence of moso bamboo and 78 transcriptome data sets to build genome-wide global and conditional co-expression networks. We overlaid the gene expression results onto the network with multiple dimensions (different development stages). Through combining the co-expression network, module classification and function enrichment tools, we identified 1,896 functional modules related to bamboo development, which covered functions such as photosynthesis, hormone biosynthesis, signal transduction, and secondary cell wall biosynthesis. Furthermore, an online database (http://bioinformatics.cau.edu.cn/bamboo) was built for searching the moso bamboo co-expression network and module enrichment analysis. Our database also includes *cis*-element analysis, gene set enrichment analysis, and other tools. In summary, we integrated public and in-house bamboo transcriptome data sets and carried out co-expression network analysis and functional module identification. Through data mining, we have yielded some novel insights into the regulation of growth and development. Our established online database might be convenient for the bamboo research community to identify functional genes or modules with important traits.

Keywords: bamboo, gene network analysis, functional module, gene expression views, growth and development

Abbreviations: BR, brassinosteroid; CPD, CONSTITUTIVE PHOTOMORPHOGENESIS AND DWARFISM; CPM, clique percolation method; DET2, DE-ETIOLATED2; FDR, false discovery rate; FPKM, fragments per kilobase of transcript per million mapped reads; GO, Gene Ontology; GSEA, gene set enrichment analysis; ICBR, International Center for Bamboo and Rattan; KEGG, Kyoto Encyclopedia of Genes and Genomes; MR, mutual rank; PCC, Pearson correlation coefficient; ROT3, CYP90C1/ROTUNDIFOLIA; SCW, secondary cell wall; SD, standard deviation; TF, transcription factor.

INTRODUCTION

Bamboo, an important fast-growing non-timber forest plant worldwide, has been an essential forest resource with an annual trade value of >2.5 billion US dollars, and approximately 2.5 billion people depend on it economically (Peng et al., 2013a,b; Zhao et al., 2017). Moso bamboo (*Phyllostachys edulis*, once known as *Phyllostachys heterocycla*) is the most economically valuable bamboo in Asia, especially in China. With the release of the whole-genome sequence of moso bamboo, there are increasing demands for refined annotation of bamboo genes on the whole-genome level. Considering the small proportion of annotated genes in the bamboo genome and high accumulation of data, it is necessary and urgent to conduct big-data mining to yield novel insights into bamboo growth and development.

Generally, genes with coordinated expression across a variety of experimental conditions indicate the presence of functional linkages between genes. Thus, co-expression gene networks can associate these genes of unknown function with biological processes in an intuitive way. An increasing number of studies have supported the versatility of co-expression analysis for inferring and annotating gene functions (D'Haeseleer et al., 2000; Aoki et al., 2007; Usadel et al., 2009; Morenorisueno et al., 2010; Li et al., 2015; Serin et al., 2016). Through data mining tools and algorithms that describe complex co-expression patterns of multiple genes in a pairwise fashion, global co-expression network analyses consider all samples (multiple data sources with independence) together and establish connections between genes based on the collective information available (Bassel et al., 2011). Compared with such a network, the conditional co-expression network aims to enhance our understanding of gene function from a portion of transcriptome data sets that have much in common, such as having the same source and a similar acquisition of raw materials and inferring gene transcriptional regulatory mechanisms in developmental processes based on a series of selected associated samples. In co-expression analysis, gene expression views can help clearly present the tendency of differential gene expression between samples. Consequently, co-expression networks with expression views can be used to associate genes of unknown function with biological processes, to discern gene transcriptional regulatory mechanisms *in vivo* and to prioritize candidate regulatory genes or modules of vital traits.

Based on the de novo sequencing data, together with the full-length complementary DNA and RNA-seq data of moso bamboo, BambooGDB has become the first genome database with comprehensively functional annotation for bamboo (Zhao et al., 2014). It is also an analytical platform composed of comparative genomic analysis, protein-protein interaction networks, pathway analysis and visualization of genomic data. However, it has only 12 RNA-seq data sets in different tissues of moso bamboo, which falls far short of existing RNA-seq data sets and does not meet the needs of researchers. Moreover, there are no analyses of co-expression networks, functional modules, *cis*-elements and gene set enrichment in BambooGDB. ATTED-II (Aoki et al., 2016), a co-expression database for plant species, provides a view of multiple co-expression data sets for nine species (*Arabidopsis*, field mustard, soybean, barrel medic, poplar, tomato, grape, rice and maize). Only two of them are members of the grass family (Poaceae), like bamboo. It is exceedingly necessary to present co-expression networks for bamboo.

Recently, large amounts of transcriptome data have become available on bamboo for the establishment of co-expression gene networks associated with plant growth and development. We collected 52 high-quality genome-wide transcriptome data sets on moso bamboo covering six tissues from the NCBI SRA database (He et al., 2013; Peng et al., 2013a; Huang et al., 2016; Wei et al., 2016; Zhao et al., 2016, 2018). In addition, we have newly produced 26 in-house transcriptome data sets across six tissues of different growth stages from the Genome Atlas of Bamboo and Rattan (GABR). To efficiently extract information from large data sets, we applied *in silico* methods to build genome-wide global and conditional co-expression networks, and further, to identify functional modules for annotating and predicting bamboo gene functions. Furthermore, we constructed the BambooNET database[1] to integrate the high-throughput transcriptome data, co-expression networks, functional modules, etc. BambooNET also included co-expression network analysis, *cis*-element analysis and GSEA tools, which might be an online server for refining annotation of bamboo gene functions.

MATERIALS AND METHODS

Moso Bamboo Samples From ICBR

Twenty-six moso bamboo (*Phyllostachys edulis*) samples of ICBR were collected from six main bamboo-producing areas in China during the spring of 2015, including (1) Yixing, Jiangsu Province (N:31°15′08.41″, E:119°43′42.55″, 212 M); (2) Tianmu Mountain, Zhejiang Province (N:30°19′13.42″, E:119°26′55.21″, 480 M); (3) Xianning, Hubei Province (N:29°81′10.02″, E:114°31′21.12″ 150 M); (4) Taojiang, Hunan Province (N:28°28′39.74″, E:112°11′18.62″, 320 M); (5) Guilin, Guangxi Province (N:28°28′39.74″, E:112°11′18.62″, 216 M) and (6) Chishui, Guizhou Province (N:28°28′15.27″, E:105°59′41.43″, 120 M), which covered rhizome, root, shoot, leaf, sheath, and bud during different development stages. Each mixed sample was collected from the above six areas.

Data Process and Gene Expression Profiling Analysis

The whole-genome sequence of moso bamboo was accessed from the 2013 public version 1 (Peng et al., 2013a) and corresponded to a genome size of ∼2 Gb and 31,987 protein-coding genes. The reads of 78 RNA-seq samples were aligned to the bamboo genome (version 1.0) using TopHat v2.1.1 software (Trapnell et al., 2009). Calculation of FPKM and the identification of differentially expressed genes were performed using Cuffdiff in Cufflinks v2.2.1 software (Trapnell et al., 2010). GO enrichment analysis was performed using the agriGO website (Du et al., 2010).

To determine the minimum threshold of the gene expression value (FPKM) among 78 bamboo samples, the lowest 5% of all

[1]http://bioinformatics.cau.edu.cn/bamboo

gene FPKM values in each RNA-seq sample and the SD of each experimental group were computed. Then, the mathematical formula "threshold = average(5% value) + 3 * SD" (You et al., 2016, 2017) was used to calculate the minimum expression value of each experimental group. The minimum threshold of FPKM was 0.1474.

Co-expression Network Construction Algorithm and Parameters

The PCC represents the co-expression relationship between two genes among the 78 samples. The closer the relationships between the genes were, the higher the PCC scores. MR, an algorithm for calculating the rank of PCC, takes a geometric average of the PCC rank from gene A to gene B and from gene B to gene A. Specifically, when gene A is the third highest co-expressed gene for gene B, the PCC rank of gene A to gene B is 3. Thus, MR ensures more credible co-expression gene pairs would be left out, so the PCC and MR were used to construct a co-expression network.

Pearson correlation coefficient:

$$r_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\left(\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)\left(\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\right)} \quad (1)$$

MR:

$$\mathrm{MR(AB)} = \sqrt{Rank(A\rightarrow B) \times Rank(B\rightarrow A)} \quad (2)$$

Here, we retained co-expressed gene pairs with a single direction rank of PCC ($Rank_{AB}$ *or* $Rank_{BA}$) less than 3 and MR score less than 30 in a co-expression network (Aoki et al., 2016), and these gene pairs were regarded as having positive co-expression relationships when their PCC values were more than zero and negative co-expression relationships when their values were less than zero.

All samples were used to construct global networks, while ICBR samples were used for conditional networks. Following a similar procedure, 65 data sets without the stress treatment were selected to define tissue-preferentially expressed genes, and 10 data sets associated with dehydration and cold treatment were selected for stress-differentially expressed genes.

Modules Identification Algorithm and Parameters

The CPM (Adamcsek et al., 2006) was used to find modules with nodes more densely connected to each other than to nodes outside the group in the bamboo co-expression networks. Parameter selection was based on the number of modules, the coverage rate of genes and the overlap rate of community. Hence, we selected a $k = 6$ clique size, which meant each node had co-expression interactions with at least five nodes in a module (**Supplementary Figure S5**). The functions of modules were predicted by gene set analysis (Yi et al., 2013) through integrating annotations such as GO, gene families (transcription regulators, kinases, and carbohydrate-active enzymes), and KEGG. The TF family and kinase family classifications were collected from iTAK (Yi et al., 2016) and PlantTFDB (Jin et al., 2017). A total of 3,305 TFs and 1,598 kinases were identified. Moreover, non-significant entries were filtered by the Fisher's exact test and multiple hypothesis testing (FDR $\leq$ 0.05). In the end, 1,896 modules containing at least 6 genes each were found in bamboo, covering functions such as metabolism, hormones, development, and transcriptional regulation.

Cis-Element Significance Analysis

The *cis*-element significance test is a statistical algorithm based on *Z* score and *P*-value filtering. When scanned in the 3 kb promoter region of bamboo genes, motifs with a *P*-value less than 0.05 were significantly enriched in a regulatory module (Yu et al., 2014; You et al., 2016).

The *Z* score was calculated as

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

where $\bar{X}$ is the sum value of a motif in the promoters of a list of genes, μ is the mean value of the same motif in 1000(n) random lists of genes with same scale, and σ is the SD of the 1000-mean value based on random selection.

Ortholog Identification in *Arabidopsis*

Bidirectional blast alignments were conducted for the analysis of protein sequences in moso bamboo and *Arabidopsis*. Our criteria for the orthologous search were as follows: the top three hits in each bidirectional blast alignment were selected as the best orthologous pairs; in addition, pairs with an *E*-value less than 1E-25 were regarded as secondary orthologous pairs. **Table 3** lists the results of the orthologous search, including for NST1, SND1 and VND7.

Search and Visualization Platform

The network search function was based on MySQL, Apache and PHP scripts. Cytoscape.js, an open source java script package, can dynamically display the components, construction and variation of the network.

RESULTS

Network Construction and Module Identification

We integrated 78 transcriptome data sets for moso bamboo (*Phyllostachys edulis*), which can be divided into two parts according to the data source: 52 public data sets from NCBI and 26 in-house data sets from ICBR (**Table 1**). The data sets spanned most tissues of bamboo, including leaf, culm (stem), shoot, root, rhizome, bud and panicle as well as stress-treated (dehydration and cold) samples from the public platform. The data sets from ICBR were available for the construction of conditional network, covering different portions from tissue root, shoot, bud, leaf and so forth. Furthermore, each ICBR sample was a mixture from six areas of bamboo production in China. This variety is

TABLE 1 | Details of RNA-seq sample resources.

Tissue	Sample information	Source	Sample number	Reference
Leaf	Transcriptome for photosystems	SRX1035287	3	BMC Plant Biol. 2016;16(1):34 (PMID:26822690)
Leaf	During dehydration and cold stresses	SRS1759772	10	PLoS One.2016;11(11):e0165953 (PMID: 27829056)
Culm (stem)	Transcriptome of developing culms	SRX329521	1	BMC plant biology.2013;13(1): 119 (PMID: 23964682)
Shoot apex (young, 25 cm long)	Shoot apical meristem region	SRS1683502	4	NCBI (2017)
	Young internode region		4	
	Young node region		4	
	Basal mature internode region		4	
	Mature node region		4	
Shoot	Three moso bamboo cultivars(10-15 cm long)	SRP067720	3	NCBI (2016)
Shoot	A thick-wall moso variant and its native wild-type	SRP075216	2	New Phytologist.2016;214(1):81 (PMID: 27859288)
Leaf	Vegetative tissues	ERS123950	2	Nature Genetics.2013;45(4):456 (PMID: 23435089)
Panicle	Panicles at the early stage and flowering stage		4	
Root	Vegetative tissues		2	
Rhizome	Vegetative tissues		2	
Shoot	20-cm-long shoot		1	
Culm	Moso bamboo	SRX342661	2	NCBI (2015)
Root-1	0.1 cm root on shoot	SRP094812	1	Gigascience.2018;doi: 10.1093/gigascience/giy115 (PMID:30202850)
Root-2	0.5 cm root on shoot		1	
Root-3	2 cm root on shoot		1	
Root-4	10 cm root on shoot		1	
Root-5	New root with lateral roots		1	
Root-6	Root on rhizome		1	
Shoot-A1/2/3	Top/middle/lower portion of 0.2 m shoot		3	
Shoot-B1/2/3	Top/middle/lower portion of 1.5 m shoot		3	
Shoot-C1/2/3	Top/middle/lower portion of 3 m shoot		3	
Shoot-D1/2/3	Top/middle/lower portion of 6.7 m shoot		3	
Leaf-1	Blade		1	
Leaf-2	Leaf sheath		1	
Sheath	Sheath sheet		1	
Bud-1/2/3	Bud on top/middle/lower portion of 3 m shoot		3	
Bud-4	Bud on rhizome		1	
Rhizome	Rhizome		1	

beneficial for the study of fast growth and development regulation in bamboo.

A well-developed integrated strategy was used for network construction (You et al., 2016, 2017). First, the raw RNA-seq reads of bamboo samples were mapped to the *moso bamboo* reference genome by TopHat, and then the FPKM of genes were calculated by Cufflinks. Since the read mapping ratios of 7 RNA-seq samples (culm tissue) were too low (<25%), we ultimately retained 78 RNA-seq samples for global co-expression network construction (details of mapping results shown in **Supplementary Table S1**). Through the FPKM value distribution boxplot of all samples (**Supplementary Figure S1**), the minimum threshold FPKM value of 0.1474 among 78 bamboo samples was chosen as a cut-off value to identify whether the gene was expressed according to all genes' FPKM values in each sample. The PCC was adopted as the correlation coefficient between two genes and to measure the co-expression relationship. The lowest 5% PCC (−0.4) and highest 5% PCC values (0.6) were considered thresholds for negative and positive correlation, respectively, in the PCC distribution diagram of all gene pairs (**Supplementary Figure S2A**). The MR

method (Aoki et al., 2016) was widely used in many species such as the model plant *Arabidopsis*. Strict parameters were set to get optimal co-expressed gene pairs in this way. It mainly classified co-expressed gene pairs into three levels: MR top3, MR $\leq$ 5 and 5 $<$ MR $\leq$ 30. As a result, the PCC cut-off of 0.6 and the MR top3 + MR cut-off of $\leq$30 were better for the construction of co-expression network. Finally, there were 302,383 and 185,044 pairs with 31,681 nodes in the positive co-expression network (PCC value $>$ 0) and negative co-expression network (PCC value $<$ 0), respectively. The network was inferred to be scale-free from the distribution of nodes and their linked edges numbers (**Supplementary Figure S2B**).

All transcriptome data sets were used to construct the global networks in this study, while the 26 data sets from ICBR were used for conditional networks as a parallel analysis with the same method (MR) and procedure as global networks. In addition, 65 data sets without stress treatment were selected to define tissue-preferentially expressed genes, and 10 data sets associated with dehydration and cold treatment were selected for stress-differentially expressed genes. We overlaid the gene expression results onto the co-expression network with multiple dimensions (development and stress).

Furthermore, co-expression networks allow modularized analysis of biological processes to discover regulatory genes or modules of vital traits. The CPM (Adamcsek et al., 2006; Li et al., 2014) together with function enrichment tools was applied to classify possible function modules. As a result, 1,896 functional modules containing at least 6 genes each were identified in bamboo, covering functions such as metabolism, hormones, development, and transcriptional regulation.

Gene Network Analysis of Photosynthesis-Related Genes

Photosynthesis provides energy for the fast growth and development of bamboo. It may possess a unique carbon assimilation mechanism, and it would be interesting to study the light-harvesting process in bamboo (Jiang et al., 2012). Additionally, an efficient light-harvesting step is critical for the success of photosynthesis (Cheng and Fleming, 2009; Zhao et al., 2016). We selected three light-harvesting complex (*LHC*) genes of photosystem I and photosystem II in bamboo (**Table 2**), including PH01003036G0080 (orthologous gene of *LHCA1*), PH01001378G0550 (orthologous gene of *CAB1* or *LHCB1.3*) and PH01000242G0150 (orthologous gene of *CAB2* or *LHCB1.1*), and searched their global co-expression networks with a tissue-preferential gene expression view in bamboo. Based on the *LHCB*-related gene expression views among different tissues, three samples for each tissue were used to detect different expression levels, which were quantified by FPKM

TABLE 2 | The genes of light-harvesting complex genes of photosystems I and II in bamboo.

Gene ID	Orthologous in *Arabidopsis*	*E*-value		Annotation
PH01003036G0080	AT3G54890	1.1E-99	LHCA1	Light-harvesting complex I chlorophyll a/b binding protein 1
PH01001974G0230	AT3G54890	1E-26	LHCA1	Light-harvesting complex I chlorophyll a/b binding protein 1
PH01000086G1040	AT3G61470	1E-117	LHCA2	Light-harvesting complex I chlorophyll a/b binding protein 2
PH01000120G1210	AT1G61520	7E-109	LHCA3	Light-harvesting complex I chlorophyll a/b binding protein 3
PH01002466G0350	AT1G61520	1E-115	LHCA3	Light-harvesting complex I chlorophyll a/b binding protein 3
PH01000008G1530	AT3G47470	8E-108	LHCA4	Light-harvesting complex I chlorophyll a/b binding protein 4
PH01000177G0160	AT3G47470	1E-106	LHCA4	Light-harvesting complex I chlorophyll a/b binding protein 4
PH01005293G0040	AT3G47470	1E-107	LHCA4	Light-harvesting complex I chlorophyll a/b binding protein 4
PH01000173G0670	AT1G45474	7.6E-83	LHCA5	Light-harvesting complex I chlorophyll a/b binding protein 5
PH01001378G0550	AT1G29930	2E-132	LHCB1.3	AB140, **CAB1**, CAB140, chlorophyll a/b binding protein1, LHCB1.3, light-harvesting chlorophyll a/b protein1.3
PH01000242G0150	AT1G29920	9E-134	LHCB1.1	AB165, **CAB2**, chlorophyll a/b binding protein2, LHCB1.1, light-harvesting chlorophyll a/b protein1.1
PH01000653G0680	AT1G29910	8E-135	LHCB1.2	AB180, **CAB3**, chlorophyll a/b binding protein3, LHCB1.2, light harvesting chlorophyll a/b binding protein1.2
PH01000625G0360	AT1G15820	3.00E-111	LHCB6	CP24, LHCB6, light harvesting complex photosystem II subunit 6
PH01002452G0070	AT2G34420	1.00E-128	LHB1	Light-harvesting complex II chlorophyll a/b binding protein 1
PH01000046G0840	AT2G34420	2.40E-99	LHB1B2	Light-harvesting complex II chlorophyll a/b binding protein 1
PH01005133G0020	AT2G34430	7.00E-131	LHB1B1	Light-harvesting complex II chlorophyll a/b binding protein 1
PH01000848G0570	AT2G05070	3.00E-120	LHCB2.2	Light-harvesting complex II chlorophyll a/b binding protein 2
PH01000184G0790	AT2G05100	7.00E-120	LHCB2.1	Light-harvesting complex II chlorophyll a/b binding protein 2
PH01000848G0570	AT3G27690	3.00E-120	LHCB2.3	Light-harvesting complex II chlorophyll a/b binding protein 2
PH01000198G0580	AT5G54270	4.00E-135	LHCB3	Light-harvesting complex II chlorophyll a/b binding protein 3
PH01003394G0090	AT5G54270	4.00E-134	LHCB3	Light-harvesting complex II chlorophyll a/b binding protein 3
PH01000198G1100	AT2G40100	2.00E-102	LHCB4	Light-harvesting complex II chlorophyll a/b binding protein 4
PH01001205G0170	AT4G10340	2.00E-112	LHCB5	Light-harvesting complex II chlorophyll a/b binding protein 5
PH01003298G0130	AT4G10340	9.00E-108	LHCB5	Light-harvesting complex II chlorophyll a/b binding protein 5

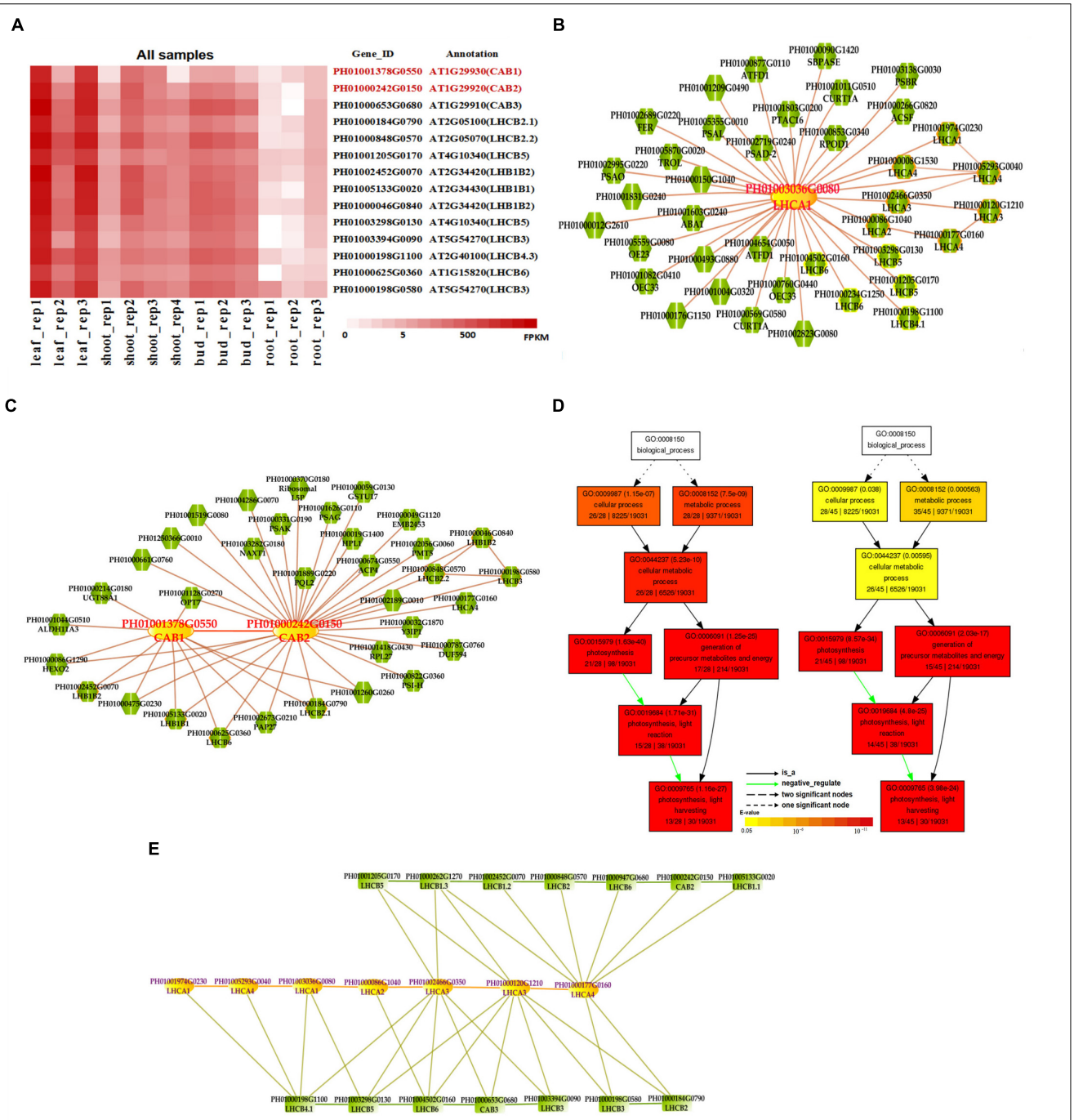


FIGURE 1 | The global co-expression network analysis of bamboo light-harvesting complex genes of photosystem I and photosystem II. **(A)** Heatmap of *LHCB* genes expressed in leaf, shoot, bud, and root samples. **(B)** A global co-expression network of the gene *LHCA1* (PH01003036G0080) in bamboo. The other *LHCA* genes are highlighted with orange borders, while nodes with yellow borders are *LHCB* genes. **(C)** A global co-expression network of *CAB1* (PH01001378G0550, *LHCB1.3*) and *CAB2* (PH01000242G0150, *LHCB1.1*). The other *LHCB* genes are also highlighted with orange borders. **(D)** GO enrichment analysis (left) of all genes from the global co-expression network for *LHCA1* (PH01003036G0080) and GO enrichment analysis (right) of all genes from the global co-expression network for *CAB1* (PH01001378G0550) and *CAB2* (PH01000242G0150) by agriGO. **(E)** The co-expression network between genes *LHCA* and *LHCB* in bamboo.

value (**Figure 1A**). These genes preferentially expressed in leaf compared to other tissues, while they were almost not expressed in root. To validate the robustness and credibility of the networks and further study the possible regulatory mechanisms of *LHCA1* and their co-expressed genes in bamboo, we selected the *LHCA1* gene (PH01003036G0080) as an example to visualize the global

co-expression network (**Figure 1B**). Through GO enrichment analysis of all genes from this network by using agriGO (Du et al., 2010; Tian et al., 2017) (**Figure 1D**), the results showed that these co-expressed genes were strongly associated with the GO terms of photosynthesis and light harvesting, light reaction, and generation of precursor metabolites and energy, which matched the previous findings that the primary function of LHC protein was the absorption of light through chlorophyll excitation and transfer of absorbed energy to photochemical reaction centers (Dolganov et al., 1995; Li et al., 2000; Montané and Kloppstech, 2000; Zhao et al., 2016). A similar result was also obtained in *CAB1*, *CAB2* and their co-expressed genes following the above process (**Figure 1C**). Zhao et al. (2016) found that more copies of *LHC* genes indicated more energy may be required in the fast-growth stage of moso bamboo. *LHCA* and *LHCB* coexist with some other *LHC* genes in these global co-expression networks (**Figure 1B**). From the perspective of only *LHCA* and *LHCB* genes' co-expression (**Figure 1E**), *LHCA* and *LHCB* are intimately linked with each other.

In addition, we also searched the co-expression network of photosynthesis-related genes in the conditional co-expression network. We performed the same procedure for the global network analysis as in the conditional network of *LHC* genes. The gene expression views in the conditional network (**Figure 2A**) showed a similar tendency to those in the global network. Meanwhile, we conducted GO enrichment analysis of all genes from the conditional co-expression network for *LHCA1* (PH01003036G0080), *CAB1* (PH01001378G0550) and *CAB2* (PH01000242G0150) by agriGO (Du et al., 2010; Tian et al., 2017). The GO terms were associated with photosynthesis, light reaction and light harvesting (**Figure 2D**). In addition to overlaps, the conditional co-expression network had some specific genes that were different from the global network (**Figures 2B,C**).

Comparative genomics might help to construct and identify functional modules in bamboo. We made a comparison between the top 300 PCC co-expressed genes in bamboo and in *Arabidopsis* (collected from ATTED-II and AraNet) (**Figures 2E,F**). The co-expression networks of PH01003036G0080 and AT3G54890 (*LHCA1*) showed high similarity, suggesting the reliability of our bamboo co-expression network.

Network Analysis of Genes Related to Brassinosteroid Biosynthetic and Signal Transduction Pathways

Phytohormones are indispensable in plant development and various environment adaption (Lacombe and Achard, 2016). BRs are a group of plant steroidal hormones that play vital roles in almost all aspects of plant growth and development (Du et al., 2017). Several key enzymes in BR biosynthesis pathways have been found in *Arabidopsis*, such as DET2/DWF6, CYP90B1/DWF4, CYP90A1/CPD/DWF3, CYP90C1/ROT3, CYP90D1, and CYP85A2/BR6OX2 (Fujioka et al., 1997; Choe et al., 1998; Yukihisa et al., 2003; Kim et al., 2005; Ohnishi et al., 2012). First, we searched the global network for gene PH01003419G0030 (orthologous gene of *CYP90A1/CPD/DWF3*), PH01000278G0580 (orthologous gene of *CYP85A1/BR6OX1*), PH01001995G0390 (orthologous gene of *CYP85A2/BR6OX2*), and PH01003429G0090 (orthologous gene of *CYP90D1*) (**Figure 3A**). Second, we conducted GSEA analysis of GO, gene family, PlantCyc and KEGG categories for all genes from this global network by using PlantGSEA (Yi et al., 2013) (**Figure 3C**). The GO terms of BR biosynthetic process, BR metabolic process and steroid biosynthetic process were significantly enriched, suggesting that this network corresponds to the BR biosynthetic pathway and the genes from this network may be involved in BR biosynthesis in bamboo. Third, we chose the genes *CYP85A1/BR6OX1* and *CYP85A2/BR6OX2* in bamboo and their top 300 co-expressed genes and compared them with their orthologous genes and their top 300 from ATTED-II in *Arabidopsis* (**Figure 3B**). There were many orthologous gene pairs between them, which could indicate these co-expressed genes were conserved and increased the credibility of predicting BR biosynthetic functional modules in bamboo. We also searched the global network for PH01000234G0890 (orthologous gene of *BAK1*, also known as BRI1-associated receptor kinase), PH01000584G0630 (orthologous gene of *BIN2*, also known as BR-insensitive 2) and their co-expressed genes (**Figure 3D**). With the GO enrichment analysis by agriGO on all the genes from this network, some GO terms were enriched such as the BR-mediated signaling pathway, steroid hormon mediated signaling pathway and responses to steroid hormone stimuli (**Supplementary Figure S4**).

Co-expression Network Analysis of Secondary Cell Wall Biosynthesis

Transcription factors in the NAC family, including VNDs, SNDs, and NSTs, acting as master switches for SCW thickening, play important roles in the SCW formation process, including the deposition of hemicellulose, cellulose and lignin (**Table 3**). MYB46 directly binds to the promoters and activates the transcription of genes involved in lignin and xylan biosynthesis, functioning as a central and direct regulator of the genes involved in the biosynthesis of all three major secondary wall components in *Arabidopsis* (Kim et al., 2013, 2014a). Thus, we selected some key NAC and MYB TFs to study their functions in regulating SCW formation and strong lignified culms in bamboo (**Table 4**). The gene expression profiling of SCW-related NAC family genes was statistically analyzed with the *Z*-score test in ICBR samples. The hierarchical cluster results of these genes demonstrated that *NST/SND* genes were highly expressed in the bamboo shoot compared to other tissues (**Figure 4**). We searched the constructed global and conditional networks with a gene expression view for these clustered *NAC* genes. The networks might indicate the possible regulatory mechanism of the SCW thickening process during bamboo development (**Figure 5**).

In the conditional network, some *NST/SND* genes, such as PH01006140G0010 (*SND3*), PH01001753G0040 (*SND2*), PH01000439G0460 (*NST2*) and PH01000003G1230 (*NST1*), were co-expressed together with *LAC4*, one of the key SCW

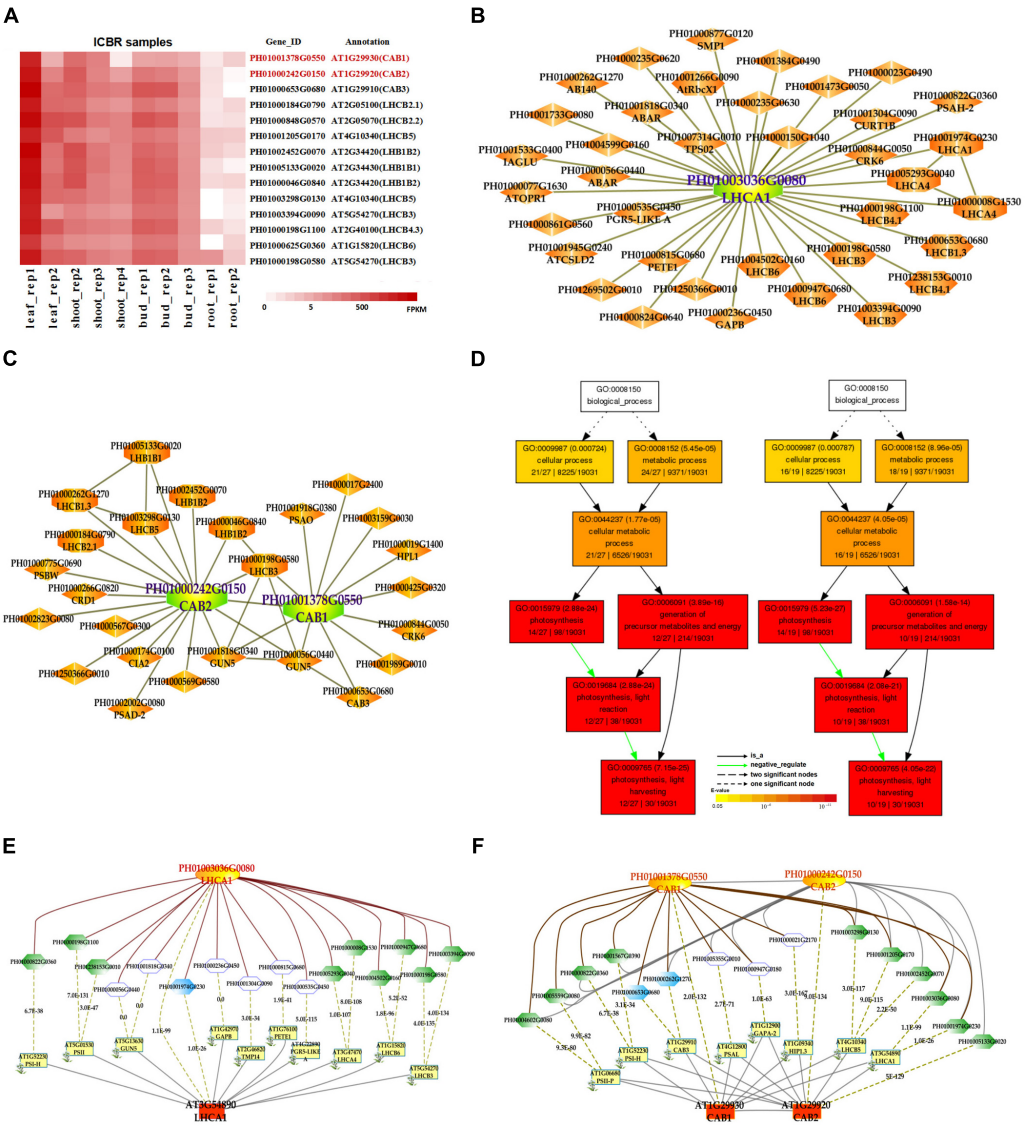


FIGURE 2 | The conditional co-expression network analysis of bamboo light-harvesting complex genes of photosystem I and photosystem II. **(A)** Heatmap of *LHCB* genes expressed in conditional leaf, shoot, bud, and root samples. **(B)** A conditional co-expression network of the gene *LHCA1* (PH01003036G0080) in bamboo. The other *LHCA* genes are highlighted with orange borders, while nodes with yellow borders are *LHCB* genes. **(C)** A conditional co-expression network of *CAB1* (PH01001378G0550, *LHCB1.3*) and *CAB2* (PH01000242G0150, *LHCB1.1*). The other *LHCB* genes are also highlighted with orange borders. **(D)** GO enrichment analysis (left) of all genes from the conditional co-expression network for *LHCA1* (PH01003036G0080) and GO enrichment analysis (right) of all genes from the conditional co-expression network for *CAB1* (PH01001378G0550) and *CAB2* (PH01000242G0150) by agriGO. **(E)** Comparison of conditional co-expression networks between *LHCA1* in bamboo and AT3G54890 (*LHCA1*) in *Arabidopsis*. Dotted lines link orthologous pairs between the two species, and the numbers in the middle of the dotted lines are the *E*-values of the BLAST results. **(F)** Comparison of conditional co-expression networks between *CAB1* and *CAB2* in bamboo and AT1G29930 (*CAB1*) and AT1G29920 (*CAB2*) in *Arabidopsis*. Dotted lines link orthologous pairs between the two species, and the numbers in the middle of the dotted lines are the *E*-values of the BLAST results.

metabolism-related genes. Xu et al. (2014) has discovered that the putative targets of miRNA-397 are several family members of laccase precursors in bamboo, including the *LAC4* gene PH01001798G0410. Additionally, *SND3* was co-expressed with the IRREGULAR XYLEM gene *IRX15* (PH01000227G0920), *AtMYB103* in bamboo (PH01000508G0100), *AtCESA8* in bamboo (PH01000018G0380) and the phenylpropanoid biosynthesis pathway gene *HCT* (PH01000118G1330).

In the global network, *SND3* was co-expressed with *SND2* and *NST1*, *CESA4*, *HAT22*, *PAL1* and an ortholog of *AtMYB83* in bamboo (PH01000006G2680). In *Arabidopsis*, both MYB46 and MYB83 act in the regulation of secondary wall biosynthesis by binding to the promoters of the xylan and lignin biosynthetic genes (Mccarthy et al., 2009; Kim et al., 2014b).

With regard to the global network for PH01000508G0100 (orthologous gene of *AtMYB103*), the co-expressed genes were

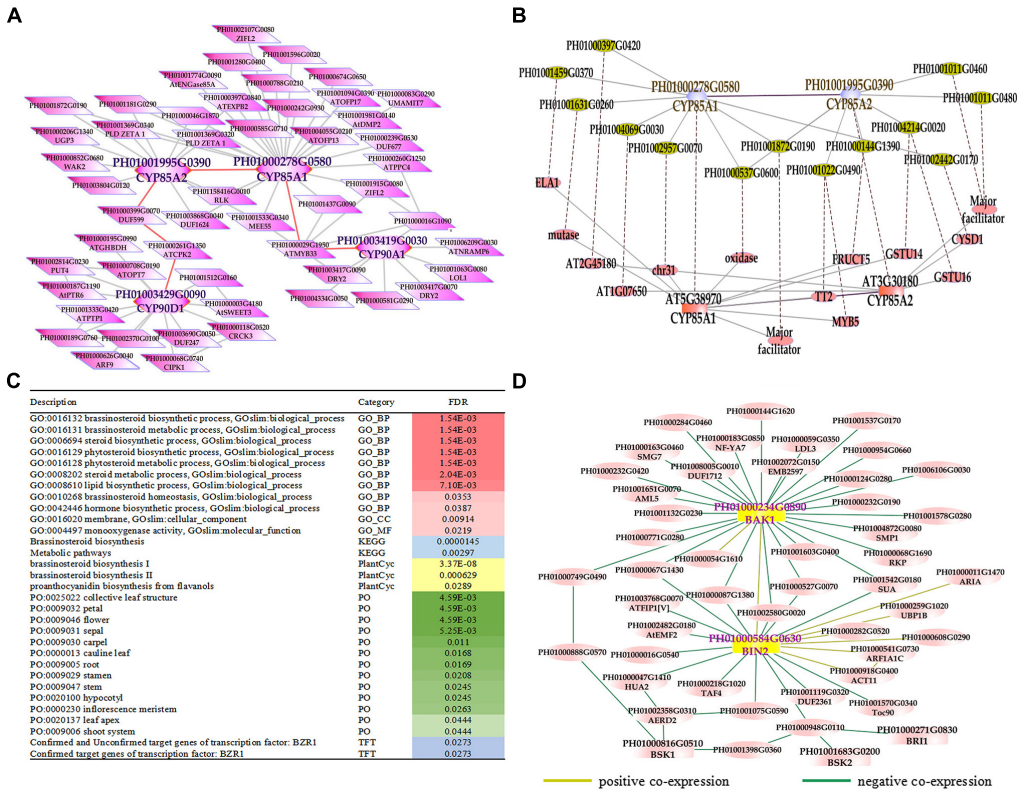


FIGURE 3 | The global co-expression network analysis of bamboo brassinosteroid (BR) biosynthesis genes and signal transduction genes. **(A)** A global co-expression network of BR biosynthesis genes *CYP85A1* (PH01000278G0580), *CYP85A2* (PH01001995G0390), *CYP90A1* (PH01003419G0030) and *CYP90D1* (PH01003429G0090) in bamboo. **(B)** Comparison of global co-expression networks between BRs biosynthesis genes *CYP85A1* and *CYP85A2* in bamboo and AT5G38970 (*CYP85A1*) and AT3G30180 (*CYP85A2*) in *Arabidopsis*. Dotted lines link orthologous pairs between the two species. **(C)** GSEA analysis of all genes from **Figure 3A**. The results shown in this table list the description, category and FDR value. The red bar on the right represents the FDR value of the enriched GO terms. The darker the red color becomes, the lower the FDR value is. Other colored bars, including blue, yellow, green, and purple bars, represent KEGG pathways, Plantcyc, PO terms and TFT, respectively. **(D)** A global co-expression network of BR signal transduction genes *BAK1* (PH01000234G0890) and *BIN2* (PH01000584G0630) in bamboo. A yellow line between two nodes indicates a positive co-expression relationship, and a green line between two nodes indicates a negative co-expression relationship.

mainly *HCT, ATCESA7, ATCESA8, CESA4*, and some *NAC* genes *SND2, SND3*. There were a few specific co-expressed genes, the *NAC* gene *NST2* and another ortholog of *AtMYB103* in bamboo (PH01000462G0290). Meanwhile, we supplied a visualization of global networks for some genes related to the phenylpropanoid biosynthesis pathway, such as PH01001164G0160 (orthologous gene of *HCT*), PH01001569G0030 (orthologous gene of *HCT*) and PH01000009G1900 (orthologous gene of *C4H*) (**Figure 5**). These genes were also co-expressed with some genes related to phenylpropanoid biosynthesis pathways, including *CCR1, CCR4, 4CL1, 4CL2,* and *CYP98A*, whose promoter regions share a *cis*-acting motif called 'AC element' that is recognized by MYB58 and MYB63 in *Arabidopsis* (Zhou et al., 2009). Through motif analysis of 3 kb of these bamboo genes' promoter regions, the 'AC element' was found to be significantly enriched.

To increase the reliability of networks in bamboo, we further made a comparison between the top 300 PCC co-expressed genes of *SND3* and *MYB103* in Arabidopsis (from ATTED-II) and those in bamboo. Plenty of orthologous gene pairs in *SND3* network comparison could be grouped into several sections, such as *LAC* genes, *MYB* genes, zinc finger genes, *IRX* family genes and other *NAC* genes. Generally, our co-expression network analysis, together with the *cis*-element and GO enrichment analyses, efficiently identified components and recapitulated a regulatory module of the SCW biosynthetic process.

A Combination of Several Functional Regulatory Modules Related to Bamboo Development

The function modules contained nodes that were more densely connected to each other than to nodes outside the group in bamboo co-expression networks. We identified an important functional module related to photosynthesis by co-expression network analysis, and the function of this module was predicted to associate with photosynthesis and light harvesting (FDR: 2.00E-8) by GSEA (**Figure 6**). We also identified functional modules related to BR biosynthetic pathways (FDR: 1.83E-3) and diterpenoid biosynthetic pathways (FDR: 6.18E-3) based on a similar approach (**Figure 6**). In addition, three

TABLE 3 | Information of NAC family in bamboo.

Gene ID	Subfamily	Orthologous in *Arabidopsis*	*E*-value
PH01000439G0460	NST2, ANAC066	AT3G61910	3.4E-80
PH01001896G0060	SND1, NST3	AT1G32770	2.9E-80
PH01000003G1230	NST1, EMB2301	AT2G46770	8.3E-91
PH01000352G0610	NST2, ANAC066	AT3G61910	1.7E-79
PH01000298G0850	SND3, ANAC010	AT1G28470	4.90E-89
PH01001753G0040	SND2, ANAC073	AT4G28500	3E-103
PH01006140G0010	SND3, ANAC010	AT1G28470	9.2E-88
PH01000046G0160	SND2, ANAC073	AT4G28500	1E-91
PH01000059G0340	VND2	AT4G36160	4.3E-87
PH01000001G1600	VND4	AT1G12260	2E-98
PH01000044G0380	VND1	AT2G18060	1.5E-74
PH01000877G0160	VND7	AT1G71930	9.2E-48
PH01004291G0080	VND7	AT1G71930	9.80E-47
PH01003084G0080	VND4	AT1G12260	1.50E-77
PH01000845G0490	VND7	AT1G71930	6E-79
PH01000083G0130	VND5	AT1G62700	3.40E-63

TABLE 4 | Information of MYB family in bamboo.

Gene ID	Subfamily	Orthologous in *Arabidopsis*	*E*-value
PH01002276G0160	ATMYB80	AT5G56110	6.8E-54
PH01000041G2150	ATMYB80	AT5G56110	2E-53
PH01000198G1320	ATMYB80	AT5G56110	4.4E-53
PH01000060G0800	MYB85	AT4G22680	2.30E-73
PH01000427G0040	MYB42	AT4G12350	6E-67
PH01001430G0250	MYB85	AT4G22680	1.7E-66
PH01003093G0130	MYB85	AT4G22680	5.1E-70
PH01128678G0010	MYB69	AT4G33450	3.5E-49
PH01002104G0150	MYB52	AT1G17950	4E-53
PH01002184G0220	MYB63	AT1G79180	2.4E-42
PH01000030G0050	MYB63	AT1G79180	2.1E-62
PH01000386G0660	MYB58	AT1G16490	3.2E-54
PH01001133G0430	MYB54	AT1G73410	6E-54
PH01000006G2680	MYB46	AT5G12870	1.5E-54
PH01000008G3080	MYB20	AT1G66230	5.5E-73
PH01005828G0060	MYB43	AT5G16600	4.10E-76
PH01000847G0490	MYB43	AT5G16600	7.7E-69
PH01000569G0800	MYB20	AT1G66230	4.2E-68
PH01001342G0270	MYB20	AT1G66230	3.3E-70
PH01000462G0290	AtMYB103	AT1G63910	9E-71
PH01000508G0100	AtMYB103	AT1G63910	5.4E-69

regulatory modules were found to possibly participate in phenylpropanoid biosynthetic pathways. For example, one functional module was significantly related to phenylpropanoid biosynthesis (FDR: 1.07E-7) and flavonoid biosynthesis (FDR: 0.02), including PH01000009G1900 (orthologous gene of *C4H*), PH01001044G0220 (orthologous gene of *CCR1*), and PH01001444G0130 (orthologous gene of *CCR1*).

We further combined several regulatory gene modules that were identified and conducted module analysis of functions related to fast growth of bamboo culms (**Figure 6**). All these modules were related to bamboo growth and development, such as photosynthesis, BR biosynthesis, and phenylpropanoid biosynthesis. Among the different modules related to phenylpropanoid biosynthesis, the connected node (PH01001164G0160, orthologous gene of *HCT*) could play a vital role in the regulatory pathway based on its co-expression network analysis. The connections between functional modules could represent crosslinks between different modules related to the similar function or different pathways. Thus, modules with nodes connected to other modules were selected and displayed in the database to help to further study their key functions. Accordingly, the combination of these functional modules displayed a series of possible key genes from hormone signals to culm development, mimicking the dynamic regulatory process in bamboo and highlighting the connections between these nodes in regulatory modules during growth stages.

Online Co-expression Network Database for Moso Bamboo

Here, we developed the BambooNET database, a platform with co-expression network analysis, *cis*-element analysis and GSEA tools and provided an online server for gene functional module analyses in multi-dimensional co-expression networks for moso bamboo (*Phyllostachys edulis*), which will help to refine annotation of bamboo gene functions. In this database, different categories of the co-expression network can be selected to visualize using the Cytoscape web tool, including the global network and the conditional network, which includes a search function for either a single gene or a list of genes. Notably, there are three main analysis options in the co-expression network platform: positive relationship, negative relationship and predicted protein-protein interaction relationship. In the tissue-preferential analysis, there are eight tissues, namely, the shoot, root, culm (stem), leaf, panicle, bud, rhizome and sheath. In addition, the gene expression changes of a certain sub-network among different tissues can be clearly observed. The stress-differential analysis displays not only the differences in gene expression between 2 and 8 h under dehydration and cold stress but also the fold changes of gene expression after each stress treatment. In the view of the network, the nodes in red or blue represent up- or down-regulated genes in leaves after a stress treatment, respectively. Moreover, some tools in this platform are available for the gene lists from the selected specific network to analyze, annotate and identify some functional modules, besides gene set analysis (Yi et al., 2013) and UCSC Genome Browser (Speir et al., 2016), such as BLAST search, keyword search and module enrichment analysis, comprising co-expression network and miRNA-target network. The website can be accessed at http://bioinformatics.cau.edu.cn/bamboo.

DISCUSSION

In this study, we constructed a genome-level co-expression network containing more than 90% of predicted genes and

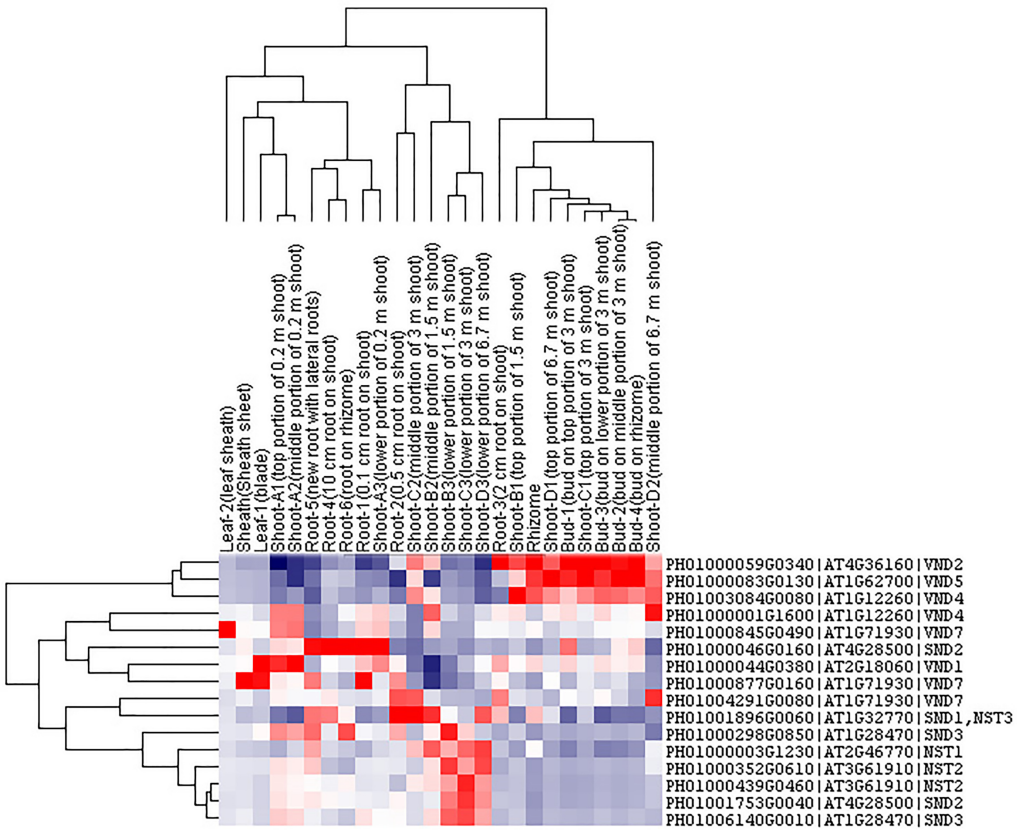


FIGURE 4 | Cluster analysis of *SND*, *NST* and *VND* genes from ICBR samples of bamboo. The gene names are indicated on the right, while the tissue types are shown above each column. High, average and low levels of expression in a specific tissue are indicated by red, white, and blue, respectively.

487,427 positive and negative edges with existing transcriptome data on bamboo. The samples cover most of the development stages of bamboo growth and development, such as the root, leaf, culm (stem), and shoot. In addition, a network-based platform, covering global, conditional and predicted protein-protein interaction networks, has been built successfully to refine the annotation of bamboo genes or modules with functions related to bamboo growth and development. Through the data mining system, networks of various aspects have combined with several functional analysis tools, including ortholog annotation, gene family classification, *cis*-element analysis and GO analysis, to evaluate the reliability of the predictions.

Although the whole-genome sequence of moso bamboo has been released, the genome annotation is still far from complete. For these sparsely annotated genes, compared with the study of single-gene identifications, modules are a valuable resource for predicting gene function. Combined with genes from co-expression networks, it would be interesting to identify modules associated with the biological process of bamboo growth and development. The potential functional modules related to phytohormones can display the module functions for BR biosynthetic pathways (FDR: 1.83E-3) and diterpenoid biosynthetic pathways (FDR: 6.18E-3) by GSEA analysis of KEGG pathways (**Figure 6**). These tightly linked genes within the one module may have related key biological functions in the process of fast growth in bamboo and can be used for genetic improvement and molecular regulatory mechanisms of moso bamboo. There is great potential for producing a large number of mutant traits of target genes related to bamboo growth and development using the clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (CRISPR/Cas9)-based genome-editing systems. Natural variants of these genes in different bamboo species may also be favorable for genetic improvement of traits in crops. Compared with the model plant *Arabidopsis*, the similar functions of orthologous genes have validated the credibility of the above network analysis and module implications based on the network comparison between *Arabidopsis* and bamboo (**Figure 3B**).

The co-expression networks with different samples are usually different. An ideal method should be able to incorporate global networks and conditional networks for different samples. Compared with all samples, there is much less diversity and far more comparability between ICBR samples (**Supplementary Figure S3**). Moreover, all ICBR samples should be classified as vegetative tissues, which may have parts specifically related to

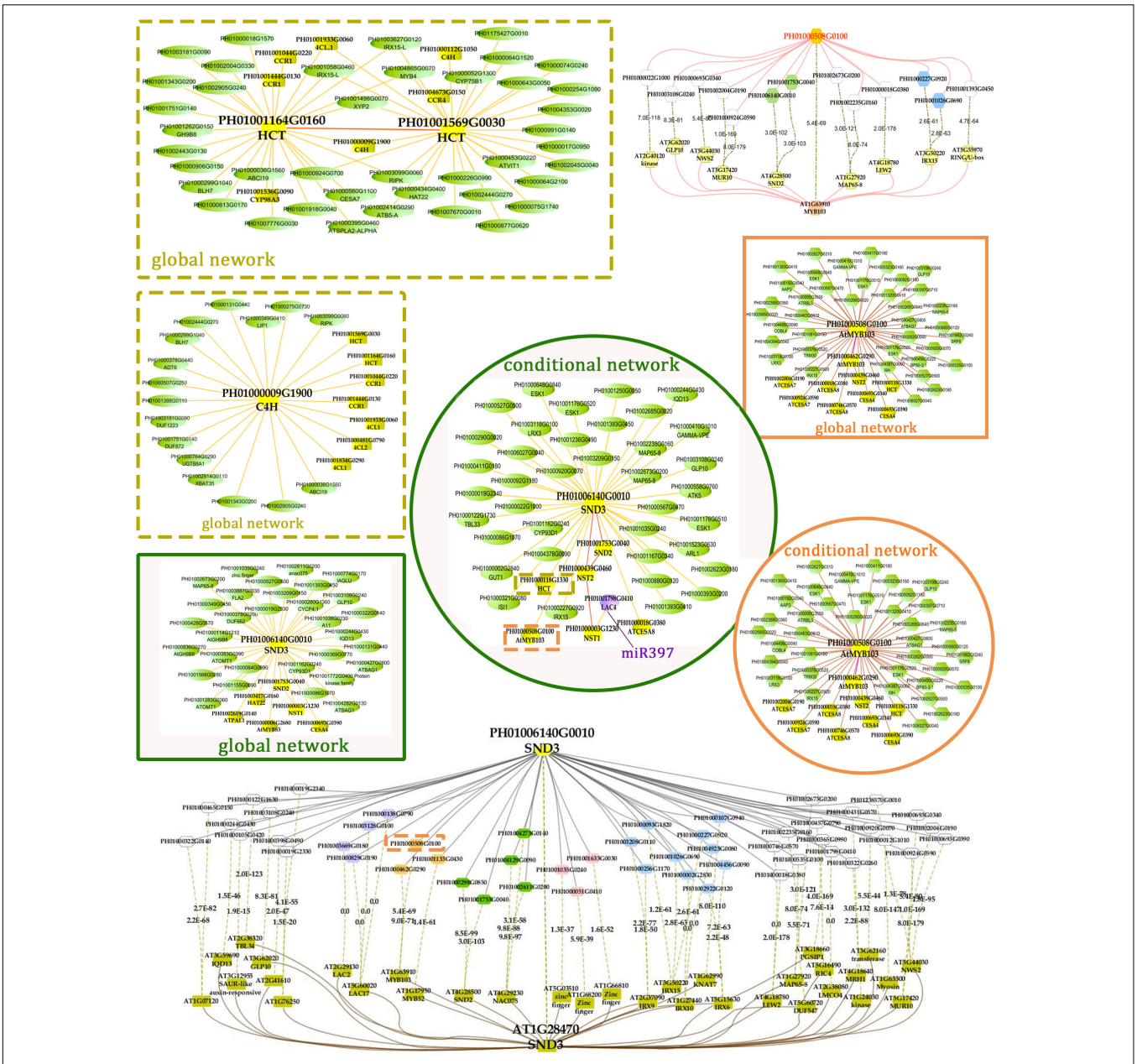


FIGURE 5 | A co-regulatory network for *NSTs/SNDs*-related genes in bamboo. PH01006140G0010 (*SND3*) and PH01000508G0100 (orthologous to *AtMYB103*) were used to present regulatory networks with all samples (global network) and partial samples (conditional network). The rounded boxes represent conditional networks, while the rectangular boxes represent global networks. The networks for the genes PH01006140G0010 (*SND3*) and PH01000508G0100 (orthologous to *AtMYB103*) are in the boxes with green solid borders and orange solid borders, respectively. The networks for *HCT* (PH01001164G0160 and PH01001569G0030) and *C4H* (PH01000009G1900) are in the boxes with yellow-dotted borders. The comparison views of conditional co-expression networks are also shown. The gray edges link to *SND3* in bamboo, and brown edges link to AT1G28470 (*SND3*) in *Arabidopsis*. The red edges link the *AtMYB103* (AT1G63910) gene and orthologs of *AtMYB103* in bamboo (PH01000508G0100). Dotted lines link orthologous pairs between the two species, and the numbers in the middle of the dotted lines are the *E*-values of the BLAST results. The *NAC, MYB*, zinc finger, *IRX* and *LAC* genes are highlighted with green, orange, pink, blue, and purple nodes, respectively. In the conditional co-expression network for PH01006140G0010 (*SND3*), the *LAC4* gene PH01001798G0410 is highlighted in purple, which is a miR397 target gene.

fast growth and the development of shoot and culm. For the co-expression networks for *LHC* genes (**Figures 2B,C**), the genes between global and conditional networks have their differences and overlaps. Therefore, the conditional co-expression network together with the global network can be complementary and then imitate the potential *LHC* genes' regulatory mechanism of fast-growth stage in bamboo. Specifically, we also investigated whether network modules are associated with specific tissue types and are enriched for specific biological process analysis by agriGO based on cluster analysis of *SND*, *NST* and *VND*

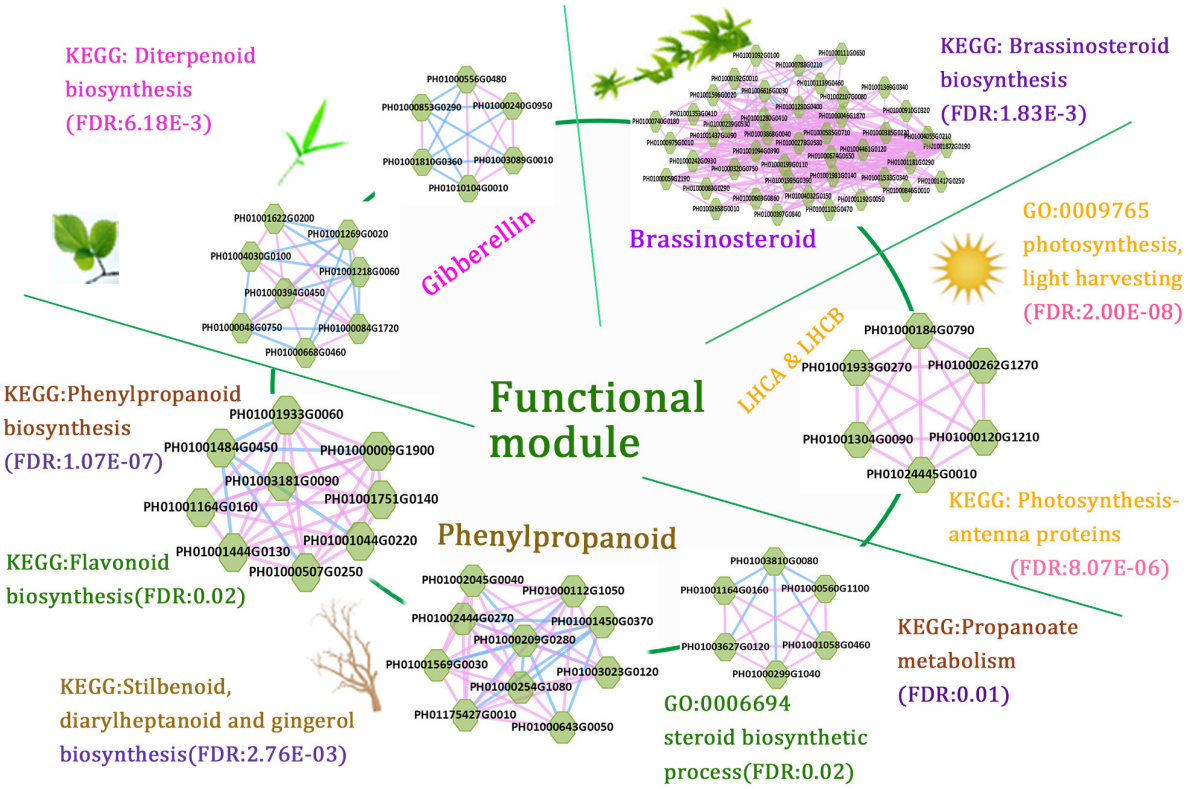


FIGURE 6 | A combination of functional regulatory modules related to bamboo development through module prediction. The functions of the modules were predicted through integrating annotations such as GO, gene families (transcription regulators, kinases, and carbohydrate-active enzymes), and KEGG, and non-significant entries were filtered by the Fisher test and multiple hypothesis testing (FDR $\leq$ 0.05). The different modules with similar functions are in the same section.

genes between conditional samples (from ICBR) and global samples (all source) of bamboo. These genes are preferentially expressed within shoot tissues relative to all other tissue types in conditional samples (**Figure 4**), which would be essential in the growth of bamboo, especially the tissue shoot. For example, one regulatory module with the gene *SND3* for SCW thickening was identified based on the conditional co-expression networks, which might indicate that these genes can fulfill their function in shoot development stages. Meanwhile, the global networks provided additional genes for further exploration of shoot tissue development.

Although BambooGDB (Zhao et al., 2014) has been integrated high-throughput sequencing data and provided researchers worldwide with a central genomic resource and an extensible analysis platform for bamboo genome, it is still necessary to build an online database for refining gene annotation and discovering novel gene functions. Through Cytoscape, our online bamboo co-expression database displays the multi-dimensional network structure and module enrichment for clear visualization and convenient analysis. Based on the co-expression network, the strategy for functional module prediction and refined gene function annotation is general and effective, so more regulatory modules could be identified by the same strategy based on a detailed biological focus or event, such as fast growth. We successfully identified 1,896 functional modules through the CPM method (Adamcsek et al., 2006), which can be searched and studied through the module enrichment in the database. These findings make it more convenient to understand the molecular regulatory mechanisms of bamboo's vital developmental traits, extremely its fast growth, which can help to dissect the molecular biological processes of bamboo. In addition, the unannotated genes establish connections to their co-expressed genes and can be refined by functional module enrichment analysis to further study unknown functions with biological processes and discern gene transcriptional regulatory mechanisms *in vivo* with the help of gene expression view in different tissues. With our multi-dimensional co-expression network, more than 90% of unannotated bamboo genes might be predicted potential functions.

However, the results might be unsatisfactory owing to the lack of complete data sets on all kinds of tissues in bamboo development stages. We believe that the detection of function modules will become much more efficient with more comprehensive transcriptome data sets of moso bamboo. Furthermore, our online bamboo co-expression database will be improved to facilitate data visualization. We will further incorporate faster and more efficient tools, such as JBrowse (Buels et al., 2016), which are very convenient for genomic track data visualization. Finally, we expect to functionally characterize

modules and to investigate how to alter modules to drive developmental changes across all developmental stages and how genes in these modules act in biological pathways.

CONCLUSION

Here, multi-dimensional bamboo samples and comparable computing measurements have been used to build a co-expression network to refine the annotation of bamboo genes or functional modules with important agronomic traits, such as growth processes. Meanwhile, module functional enrichment analysis tools, such as gene family classification, *cis*-element analysis and GO analysis, have been used to evaluate the reliability of the predictions. Based on the gene expression analysis and conditional network, the strategy for functional module prediction and refined gene function annotation is general and effective. Thus, more regulatory modules could be identified by the same strategy based on a detailed biological focus or event, such as fast growth. Therefore, this approach will improve our understanding of the molecular regulatory mechanisms underlying vital agronomic traits, such as growth and development. We hope that more transcriptome data will improve the network analysis for functional module identification and reduce biases or mistakes caused by its current limitations, increasing our understanding of bamboo growth and development.

AUTHOR CONTRIBUTIONS

ZS, WX, and ZG designed the project. XM, HZ, WX, and HY performed the research. XM, QY, and WX analyzed the data and conducted the bioinformatics analysis. XM, HZ, WX, ZG, and ZS wrote the article.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (31771467 and 31371291).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00574/full#supplementary-material

REFERENCES

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023. doi: 10.1093/bioinformatics/btl039

Aoki, K., Ogata, Y., and Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.* 48, 381–390. doi: 10.1093/pcp/pcm013

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. (2016). ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiol.* 57:e5. doi: 10.1093/pcp/pcv165

Bassel, G. W., Glaab, E., Marquez, J., Holdsworth, M. J., and Bacardit, J. (2011). Functional network construction in arabidopsis using rule-based machine learning on large-scale data sets. *Plant Cell* 23, 3101–3116. doi: 10.1105/tpc.111.088153

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17:66. doi: 10.1186/s13059-016-0924-1

Cheng, Y. C., and Fleming, G. R. (2009). Dynamics of light harvesting in photosynthesis. *Annu. Rev. Phys. Chem.* 60, 241–262. doi: 10.1146/annurev.physchem.040808.090259

Choe, S., Dilkes, B. P., Fujioka, S., Takatsuto, S., Sakurai, A., and Feldmann, K. A. (1998). The DWF4 gene of *Arabidopsis* encodes a cytochrome P450 that mediates multiple 22alpha-hydroxylation steps in brassinosteroid biosynthesis. *Plant Cell* 10, 231–243. doi: 10.1105/tpc.10.2.231

D'Haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726. doi: 10.1093/bioinformatics/16.8.707

Dolganov, N. A., Bhaya, D., and Grossman, A. R. (1995). Cyanobacterial protein with similarity to the chlorophyll a/b binding proteins of higher plants: evolution and regulation. *Proc. Natl. Acad. Sci. U.S.A.* 92, 636–640. doi: 10.1073/pnas.92.2.636

Du, J., Zhao, B., Sun, X., Sun, M., Zhang, D., Zhang, S., et al. (2017). Identification and characterization of multiple intermediate alleles of the key genes regulating brassinosteroid biosynthesis pathways. *Front. Plant Sci.* 7:1893. doi: 10.3389/fpls.2016.01893

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 38, W64–W70. doi: 10.1093/nar/gkq310

Fujioka, S., Li, J., Choi, Y. H., Seto, H., Takatsuto, S., Noguchi, T., et al. (1997). The Arabidopsis deetiolated2 mutant is blocked early in brassinosteroid biosynthesis. *Plant Cell* 9, 1951–1962. doi: 10.1105/tpc.9.11.1951

He, C., Cui, K., Zhang, J., Duan, A., and Zeng, Y. (2013). Next-generation sequencing-based mRNA and microRNA expression profiling analysis revealed pathways involved in the rapid growth of developing culms in moso bamboo. *BMC Plant Biol.* 13:119. doi: 10.1186/1471-2229-13-119

Huang, Z., Zhong, X. J., He, J., Jin, S. H., Guo, H. D., Yu, X. F., et al. (2016). Genome-wide identification, characterization, and stress-responsive expression profiling of genes encoding lea (late embryogenesis abundant) proteins in moso bamboo (*Phyllostachys edulis*). *PLoS One* 11:e0165953. doi: 10.1371/journal.pone.0165953

Jiang, Z. H., Peng, Z. H., Gao, Z. M., Liu, C., and Yang, C. H. (2012). Characterization of different isoforms of the light-harvesting chlorophyll a/b complexes of photosystem II in bamboo. *Photosynthetica* 50, 129–138. doi: 10.1007/s11099-012-0009-7

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Kim, G. T., Fujioka, S., Kozuka, T., Tax, F. E., Takatsuto, S., Yoshida, S., et al. (2005). CYP90C1 and CYP90D1 are involved in different steps in the brassinosteroid biosynthesis pathway in *Arabidopsis thaliana*. *Plant J. Cell Mol. Biol.* 41, 710–721. doi: 10.1111/j.1365-313X.2004.02330.x

Kim, W. C., Kim, J. Y., Ko, J. H., Kang, H., and Han, K. H. (2014a). Identification of direct targets of transcription factor MYB46 provides insights into the transcriptional regulation of secondary wall biosynthesis. *Plant Mol. Biol.* 85, 589–599. doi: 10.1007/s11103-014-0205-x

Kim, W. C., Reca, I. B., Kim, Y. S., Park, S., Thomashow, M. F., Keegstra, K., et al. (2014b). Transcription factors that directly regulate the expression of CSLA9 encoding mannan synthase in *Arabidopsis thaliana*. *Plant Mol. Biol.* 84, 577–587. doi: 10.1007/s11103-013-0154-9

Kim, W. C., Ko, J. H., Kim, J. Y., Kim, J., Bae, H. J., and Han, K. H. (2013). MYB46 directly regulates the gene expression of secondary wall-associated cellulose

synthases in Arabidopsis. *Plant J.* 73, 26–36. doi: 10.1111/j.1365-313x.2012.05124.x

Lacombe, B., and Achard, P. (2016). Long-distance transport of phytohormones through the plant vascular system. *Curr. Opin. Plant Biol.* 34, 1–8. doi: 10.1016/j.pbi.2016.06.007

Li, J., Wang, X., and Cui, Y. (2014). Uncovering the overlapping community structure of complex networks by maximal cliques. *Physica A* 415, 398–406. doi: 10.1016/j.physa.2014.08.025

Li, X. P., Björkman, O., Shih, C., Grossman, A. R., Rosenquist, M., Jansson, S., et al. (2000). A pigment-binding protein essential for regulation of photosyntheticlight harvesting. *Nature* 403, 391–395. doi: 10.1038/35000131

Li, Y., Pearl, S. A., and Jackson, S. A. (2015). Gene networks in plant biology: approaches in reconstruction and analysis. *Trends Plant Sci.* 20, 664–675. doi: 10.1016/j.tplants.2015.06.013

Mccarthy, R. L., Zhong, R., and Ye, Z. H. (2009). MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in arabidopsis. *Plant Cell Physiol.* 50, 1950–1964. doi: 10.1093/pcp/pcp139

Montané, M. H., and Kloppstech, K. (2000). The family of light-harvesting-related proteins (LHCs, ELIPs, HLIPs): was the harvesting of light their primary function? *Gene* 258, 1–8. doi: 10.1016/S0378-1119(00)00413-3

Morenorisueno, M. A., Busch, W., and Benfey, P. N. (2010). Omics meet networks - using systems approaches to infer regulatory networks in plants. *Curr. Opin. Plant Biol.* 13, 126–131. doi: 10.1016/j.pbi.2009.11.005

Ohnishi, T., Godza, B., Watanabe, B., Fujioka, S., Hategan, L., Ide, K., et al. (2012). CYP90A1/CPD, a brassinosteroid biosynthetic cytochrome P450 of arabidopsis, catalyzes C-3 oxidation. *J. Biol. Chem.* 287, 31551–31560. doi: 10.1074/jbc.M112.392720

Peng, Z., Lu, Y., Li, L., Zhao, Q., Feng, Q., Gao, Z., et al. (2013a). The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nat. Genet.* 45, 456–461. doi: 10.1038/ng.2569

Peng, Z., Zhang, C., Zhang, Y., Hu, T., Mu, S., Li, X., et al. (2013b). Transcriptome sequencing and analysis of the fast growing shoots of moso bamboo (*Phyllostachys edulis*). *PLoS One* 8:e78944. doi: 10.1371/journal.pone.0078944

Serin, E. A. R., Harm, N., Hilhorst, H. W. M., and Wilco, L. (2016). Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* 7:444. doi: 10.3389/fpls.2016.00444

Speir, M. L., Zweig, A. S., Rosenbloom, K. R., Raney, B. J., Paten, B., Nejad, P., et al. (2016). The UCSC genome browser database: 2016 update. *Nucleic Acids Res.* 44, D717–D725. doi: 10.1093/nar/gkv1275

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/nar/gkx382

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., et al. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ.* 32, 1633–1651. doi: 10.1111/j.1365-3040.2009.02040.x

Wei, Q., Jiao, C., Guo, L., Ding, Y., Cao, J., Feng, J., et al. (2016). Exploring key cellular processes and candidate genes regulating the primary thickening growth of Moso underground shoots. *New Phytol.* 214, 81–96. doi: 10.1111/nph.14284

Xu, P., Mohorianu, I., Yang, L., Zhao, H., Gao, Z., and Dalmay, T. (2014). Small RNA profile in moso bamboo root and leaf obtained by high definition adapters. *PLoS One* 9:e103590. doi: 10.1371/journal.pone.0103590

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–W103. doi: 10.1093/nar/gkt281

Yi, Z., Chen, J., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors,transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

You, Q., Xu, W., Zhang, K., Zhang, L., Yi, X., Yao, D., et al. (2017). ccNET: database of co-expression networks with functional modules for diploid and polyploid *Gossypium*. *Nucleic Acids Res.* 45, D1090–D1099. doi: 10.1093/nar/gkw910

You, Q., Zhang, L., Yi, X., Zhang, K., Yao, D., Zhang, X., et al. (2016). Co-expression network analyses identify functional modules associated with development and stress response in *Gossypium arboreum*. *Sci. Rep.* 6:38436. doi: 10.1038/srep38436

Yu, J., Zhang, Z., Wei, J., Yi, L., Xu, W., and Su, Z. (2014). SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways. *BMC Genomics* 15:271. doi: 10.1186/1471-2164-15-271

Yukihisa, S., Hideki, G., Ayako, N., Suguru, T., Shozo, F., and Shigeo, Y. (2003). Organ-specific expression of brassinosteroid-biosynthetic genes and distribution of endogenous brassinosteroids in Arabidopsis. *Plant Physiol.* 131, 287–297. doi: 10.1104/pp.013029

Zhao, H., Gao, Z., Wang, L., Wang, J., Wang, S., Fei, B., et al. (2018). Chromosome-level reference genome and alternative splicing atlas of moso bamboo (*Phyllostachys edulis*). *Gigascience* 7:giy115. doi: 10.1093/gigascience/giy115

Zhao, H., Lou, Y., Sun, H., Li, L., Wang, L., Dong, L., et al. (2016). Transcriptome and comparative gene expression analysis of *Phyllostachys edulis* in response to high light. *BMC Plant Biol.* 16:34. doi: 10.1186/s12870-016-0720-9

Zhao, H., Peng, Z., Fei, B., Li, L., Hu, T., Gao, Z., et al. (2014). BambooGDB: a bamboo genome database with functional annotation and an analysis platform. *Database* 2014:bau006. doi: 10.1093/database/bau006

Zhao, H., Zhao, S., International Network for Bamboo and Rattan, Fei, B., Liu, H., Yang, H., et al. (2017). Announcing the genome atlas of bamboo and rattan (GABR) project: promoting research in evolution and in economically and ecologically beneficial plants. *Gigascience* 6, 1–7. doi: 10.1093/gigascience/gix046

Zhou, J., Lee, C., Zhong, R., and Ye, Z. H. (2009). MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. *Plant Cell* 21, 248–266. doi: 10.1105/tpc.108.063321

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: 18 December 2018
doi: 10.3389/fgene.2018.00644

PhageWeb – Web Interface for Rapid Identification and Characterization of Prophages in Bacterial Genomes

*Ailton Lopes de Sousa[1], Dener Maués[2], Amália Lobato[1], Edian F. Franco[1], Kenny Pinheiro[1], Fabrício Araújo[1], Yan Pantoja[1], Artur Luiz da Costa da Silva[1], Jefferson Morais[2] and Rommel T. J. Ramos[1]**

[1] Institute of Biological Sciences, Federal University of Para, Belém, Brazil, [2] Institute of Exact and Natural Sciences, Federal University of Para, Belém, Brazil

This study developed a computational tool with a graphical interface and a web-service that allows the identification of phage regions through homology search and gene clustering. It uses G+C content variation evaluation and tRNA prediction sites as evidence to reinforce the presence of prophages in indeterminate regions. Also, it performs the functional characterization of the prophages regions through data integration of biological databases. The performance of PhageWeb was compared to other available tools (PHASTER, Prophinder, and PhiSpy) using Sensitivity (Sn) and Positive Predictive Value (PPV) tests. As a reference for the tests, more than 80 manually annotated genomes were used. In the PhageWeb analysis, the Sn index was 86.1% and the PPV was approximately 87%, while the second best tool presented Sn and PPV values of 83.3 and 86.5%, respectively. These numbers allowed us to observe a greater precision in the regions identified by PhageWeb while compared to other prediction tools submitted to the same tests. Additionally, PhageWeb was much faster than the other computational alternatives, decreasing the processing time to approximately one-ninth of the time required by the second best software. PhageWeb is freely available at http://computationalbiology.ufpa.br/phageweb.

Keywords: phage, prophage, clustering, web interface, web service, characterization, bacterial genome

OPEN ACCESS

Edited by:
Helder Nakaya,
University of São Paulo, Brazil

Reviewed by:
Luciane Schons da Fonseca,
Massachusetts Institute of Technology, United States
Yu Xue,
Huazhong University of Science and Technology, China

****Correspondence:***
Rommel T. J. Ramos
rommelthiago@gmail.com

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

Received: *21 August 2018*
Accepted: *27 November 2018*
Published: *18 December 2018*

Citation:
Sousa AL, Maués D, Lobato A, Franco EF, Pinheiro K, Araújo F, Pantoja Y, Costa da Silva AL, Morais J and Ramos RTJ (2018) PhageWeb – Web Interface for Rapid Identification and Characterization of Prophages in Bacterial Genomes. Front. Genet. 9:644. doi: 10.3389/fgene.2018.00644

INTRODUCTION

Phages are the most abundant organisms on earth (Rohwer, 2003), inhabiting various environments and they are able to infect various bacterial species. Phages are also an important factor in bacterial evolution through horizontal gene transfer (Ochman et al., 2000) because they allow the insertion of extrinsic genetic material that can provide new characteristics to their hosts, such as antibiotic resistance, virulence factors, operons or even genomic islands (Bernheim and Sorek, 2018). These characteristics are present in cases of diphtheria (Brüssow et al., 2004), cholera (Kim et al., 2010), and food poisoning by enterohaemorrhagic *Escherichia coli* (Tozzoli et al., 2014). Moreover, phages have biotechnological applications as cloning in phage display (Winter et al., 1994), diagnosis of infections by phagotyping (Haq et al., 2012; Schofield et al., 2012), vehicles for vaccine delivery (Jafari and Abediankenari, 2015) and phage therapy as an alternative to antibiotics (Levin and Bull, 2004). Phages also play an ecological role, helping recycle nutrients, and increasing photosynthesis in the oceans (Mann et al., 2003; Sullivan et al., 2003). These organisms have two life cycles: lytic

and lysogenic. During the lytic cycle, after the successful integration in the bacterial genome, phages can perform incision and excision, or remain dormant in the genome. They are called prophages. Depending on the size of the region and the success of the insertion, the prophage may remain complete and/or become cryptic (Canchaya et al., 2003; Brüssow et al., 2004) by decay, where the remains of its genetic material can provide the host genes that benefit its survival.

Prophages can be considered a cluster of phage-like genes (Zhou et al., 2011). Computational approaches, such as clustering algorithms are used to determine if these genes are close enough to each other to constitute a prophage region (Lima-Mendez et al., 2008; Zhou et al., 2011). Moreover, an important factor for the identification of prophages is the integration of the phages into specific insertion sites, such as in the bacterial genome tRNA genes (Delesalle et al., 2016). Thus, insertions in these genes indicate extrinsic genetic material, although phages do not use these sites exclusively. In addition, G+C content has been a feature used to confirm horizontal gene transfer, the presence of genomic islands and, generally, the identification of mobile genetic elements (Langille et al., 2010). In such regions, the G+C content may be quite distinct compared to the rest of the organism's genome, and this feature is commonly used to confirm, *in silico*, the presence of horizontal gene transfer – HGT (Eng et al., 2011).

Many bacterial genomes available in public databases contain phage DNA integrated into their chromosome and phage DNA, in some cases, can make up 10–20% of the bacterial genome (Casjens, 2003). Due to the reduced cost of sequencing of complete bacterial genomes and the high costs for detection of prophages by bench methodologies (Metzker, 2010), new in silico tools for prophage detection in sequenced genomes (Lima-Mendez et al., 2008; Zhou et al., 2011; Akhter et al., 2012) and for prediction of DNA phage sequences in metagenomic data (Amgarten et al., 2018) have been developed. These computational tools generally use an approach that identifies sets of encoding protein genes according to some similarity to known phage genes. However, some of these tools present hindrances, such as the absence of a graphical interface, slow processing and a lack of a broader methodology for finding prophages in bacterial genomes (Srividhya et al., 2007).

Thus, this work presents PhageWeb, a tool to identify prophages in bacterial genomes that considers the similarity of gene sequences against a phage database, using indicators such as alteration of G+C content and, additionally, the presence of tRNA flanking the region which can be used as an evidence of insertion site (Campbell, 2003). These parameters allow analysis of each of the regions through functional characterization with fast processing.

MATERIALS AND METHODS

Pipeline

PhageWeb receives bacterial genomic sequences in GenBank or EMBL format, or the NCBI's Accession Number of the bacterial genome as input for analysis. After, it uses the DIAMOND tool (Buchfink et al., 2015) to identify phage-homologous regions in bacterial genomes based on its own database (updated by the application itself), generating a data table that is integrated into the pipeline. The user can change the parameters to refine their analyses: MinPts (minimum number of phage proteins in a region) and the alignment identity against the phage database. Once the input data have been submitted, homology search and gene clustering step select prophage candidate regions. After G+C content and tRNA sites are identified and the characterization of the predictive sequences is performed. Finally, a phage gene conservation analysis optional is performed to indicate the possible integrity of the predicted regions, based on percentual of elements genic. If in a given region identified by PhageWeb there is an index for example of 80% or more of genes belonging to a given phage, it considers a potentially conserved region; but if the region has an index of less than 80%, it will be considered no conserved. The percentage value is optionally assigned by the user at the beginning of each analysis. The pipeline of PhageWeb is shown in **Figure 1**.

Graphical Elements

The interactive graphics for prophage regions in this application were encoded using the JavaScript component of the AngularPlasmid component[1] – a DNA plasmid visualization component developed using Google's AngularJS framework. AngularPlasmid provides an implementation that creates plasmid maps that are easy to use on the web. Instead of client-side JavaScript coding or other server-side programming languages, AngularPlasmid provides easy-to-use HTML markup, making generation as easy as creating a web page.

Phage Database

The PhageWeb database consists of a collection of prophages sequences reported in several public databases. Two sources of data collection were used: the genome database of the National Center for Biotechnology Information (NCBI) database[2] and the European Bioinformatics Institute (EBI) database[3]. The latter has an interactive environment for collecting and sharing information related to phage genomics. This way, the identified sequences were stored in a database developed in MySQL and incorporated into the application. All nucleotide sequences (FASTA and annotated files), as well as the database, are available in the tool, which is updated weekly.

Controlled Dataset

Eighty-four complete bacterial genomes that have predicted regions and manually annotated prophages (Casjens, 2003) were collected to be used to verify and quantify processing time, accuracy and performance of PhageWeb in relation to other software.

[1]http://angularplasmid.vixis.com

[2]http://www.ncbi.nlm.nih.gov

[3]http://www.ebi.ac.uk/genomes/phage.html

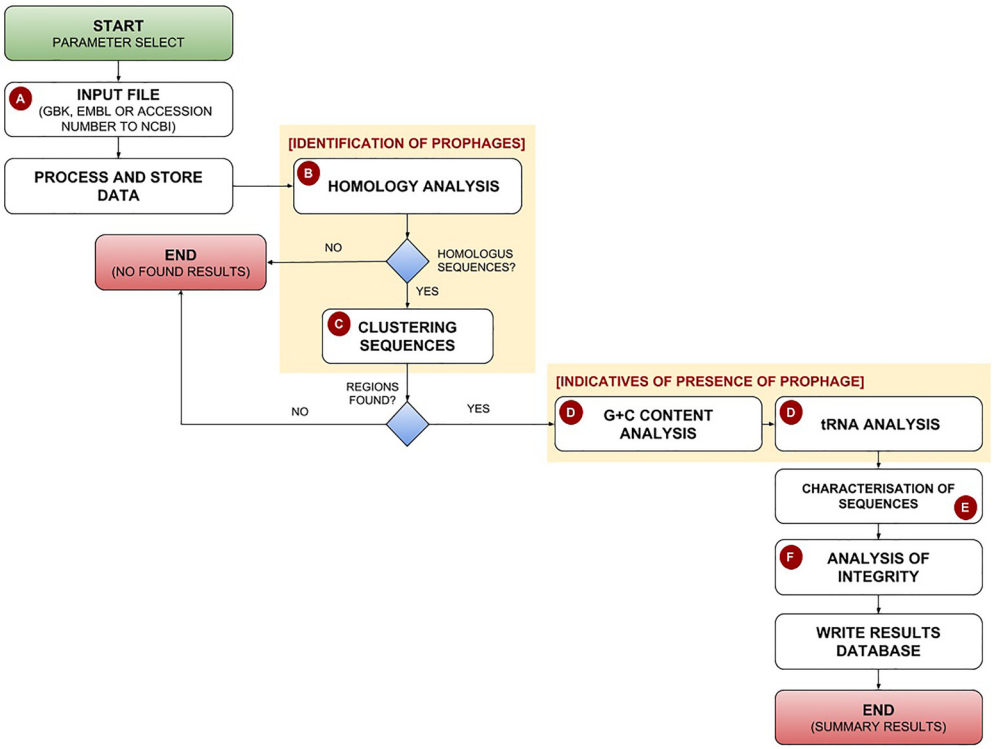


FIGURE 1 | Pipeline for identification and characterization of prophages by the PhageWeb computational tool. **(A)** The pipeline receives as input the parameters (Alignment Identity and MinPts – minimum number of phage proteins in a region) and the annotation file (GenBank, Embl or the NCBI's Accession Number) of the target genome which will be evaluated; **(B)** The homology search of the coding sequences in a local database based on the publically available sequences annotated as phage obtained from NCBI, which is automatically updated once per week; **(C)** After identifying the homology sequences, a clustering analysis based on the distance of the elements is performed: part of phage region to be evaluated on the amount of prophage there; **(D)** This optional step is useful to know if the identified region has more features which can be an evidence of prophage: G+C content (due to be possible variation of G+C content on the flanks of the prophage) and the presence of tRNAs on the flank; **(E)** Use of web services to connect biological databases to perform the functional characterization of the identified sequences; **(F)** Verification of the probable integrity of the prophage based on the composition of genes of each identified phage.

Criteria for Identification of Prophage Regions

Clustering Algorithm

The controlled dataset (Casjens, 2003) was used to identify prophage regions by clustering known phage sequences, based on the coordinates in the genome of the homologous genes (Zhou et al., 2011). Three density-based clustering algorithms were evaluated - DBSCAN, OPTICS, and HDBSCAN – to identify the prophage candidate and to be implemented in PhageWeb. For the performance evaluation of the algorithms, four cluster evaluation metrics were used: Silhouette (Rousseeuw, 1987), Dunn (Dunn, 1974), Davies–Bouldin (Davies and Bouldin, 1979), and the Density-Based Clustering Validation index – DBCV (Moulavi et al., 2014).

G+C Content

To increase the precision in the identification of prophages, a method based on DNA composition (Eng et al., 2011) was used, where a sliding window of 1000 bp moves through the entire target genome to be analyzed. The sliding window divides the genome into several smaller sets (regions), and each region can be evaluated according to its G+C content (Lu and Leong, 2016). Previous studies (Eng et al., 2011) proposed the evaluation of HGT by G+C content of the genes inserted in these regions. This way, PhageWeb proposes to classify a specific region as a prophage if at least 80% of the genes show percent G+C above the mean plus one standard deviation or show percent GC below the mean minus one standard deviation.

Regions tRNA

Phages generally integrate into specific insertion sites. Among them, the tRNA genes of the bacterial genome (Campbell, 2003; Delesalle et al., 2016). Those sites can be used as an indication of the presence of external genetic material insertion, although phages don't use only these places as the target for integration.

Web Services

The functional characterization of the prophage regions is performed by integrating the results obtained in the PhageWeb identification step and public databases like UniProt, NCBI, InterPro, KEGG, Pfam and Gene Ontology through the UniProt public API by Web Service. After the integration, results can be

processed and displayed in charts and tables to simplify analysis and understanding of results.

Software

PhageWeb was developed to be a graphical interface for the rapid identification and characterization of prophages in bacterial genomes, using PHP combined with Python and Perl programming languages, besides the Bootstrap Framework. The PhageWeb tool implements an algorithm that combines similarity searches, using analysis and implementation of clustering algorithms in high density for the identification of regions in bacterial genomes. The software is available for use at: http://www.computationalbiology.ufpa.br/phageweb, and it is compatible with Mozilla Firefox 55.0.3, Opera 38.0.2 and Google Chrome 61.0. Additionally, an Application Programming Interface (API) was created to allow the external execution and, consequently, facilitating the integration of the application with other software. The API and usage instructions are available at: https://github.com/phagewebufpa/API.

Tools Comparison

Three tools available to predict phages sequences on genomes were evaluated: Prophinder (Lima-Mendez et al., 2008), PHASTER (Arndt et al., 2016), and PhiSpy (Akhter et al., 2012).

Prophinder is one of the first web tools for prophage detection. It uses coding sequences (CDS) that are similar to those found in ACLAME database using BLAST. Based on the annotation of the ACLAME database, Prophinder selects the genes with the best correspondence to a potential prophage. PHASTER is also a web tool developed to identify phages inside bacterial genomes. Like Prophinder, it also uses homology search for prediction. PHASTER is an upgraded version of the Phast (Zhou et al., 2011) program and accepts DNA sequences data as well as annotated data in GenBank format as input. In general, PHASTER stands out for its ability to provide quality annotations with the prophage's characteristics and to distinguish between intact and incomplete prophage. PhiSpy, however, differs from the others due to its ability to identify prophage regions that does not have any similarity to known target genes: it is not based on homology search in their predictions. PhiSpy phage detection algorithm was developed based on seven phage distinguishing characteristics: length of the protein, the direction of the transcription chain, A+T inclination and conventional G+C, the abundance of unique phage words, insertion point and similarity of phage proteins. Regarding the parameters, PHASTER, Prophinder, and PhiSpy were used with default parameter values. To compare the performance results of the computational tools, the values of Sensitivity and Positive Predictive Value will be used as evaluation metrics.

TABLE 1 | Performance Evaluation of Clustering algorithms in the identification of prophage regions, based on the metrics Silhouette, Dunn, Davies-Bouldin (DB), and Density-Based Clustering Validation index (DBCV).

Algorithms	Cluster	Silhouette	DBCV	Dunn	DB
Dbscan	151	0.47	−0.73323973	0.0006	0.553
Optic	168	0.54	−0.677653797	0.003	**0.51**
Hdbscan	**186**	**0.86**	**0.285253761**	**0.087**	1.2

Silhouette – Refers to a method of interpretation and validation of data consistency within clusters; Dunn – A metric for evaluating clustering algorithms, and its purpose is to identify clusters of compact clusters, with a small variation among cluster members; Davies-Bouldin – Is a metric to validate how well the cluster was made using quantities and characteristics inherent to the data set; DBCV – This is a relative validation index for arbitrarily density-based clusters. The highlighted results (underscores) represent the algorithm mean value with the best performance in the identification and formation of clusters of the prophages according to the metrics.

TABLE 2 | Comparative analysis of values obtained for Sn (Sensitivity) and PPV (Positive Predictive Value) between computational tools.

	Phaster	Prophinder	PhiSpy	PhageWeb
Sn	83.33%	81.02%	52.78%	86.11%
PPV	86.54%	77.43%	88.37%	87.32%

The complete data this analysis can be observed in the ***Supplementary Information*** *section.*

TABLE 3 | Comparison of functionalities and features of phage prediction tools.

Resource	Phaster	Prophinder	PhiSpy	PhageWeb
Using graphical interface	Yes	Yes	No	Yes
Homology analyses	Yes	Yes	Yes	Yes
Analyses of tRNA sites	Yes	No	No	Yes
G+C content analysis	No	No	No	Yes
Results exportation	Yes	Yes	No	Yes
Circular genome view	Yes	No	No	Yes
Characterization of sequences	Yes	No	No	Yes
Alignment details	Yes	No	No	Yes
Support for biological databases integration	No	No	No	Yes
Output types	Text, graphics	Text, graphics	Text only	Text, graphics
Run time (seconds)	∼365	∼1890	∼5547	∼22

Sensitivity and Positive Predictive Value

The performance of PhageWeb against other platforms was evaluated using Sensitivity (Sn), representing the proportion of individuals or elements with the positive classification that yielded a positive result for a particular test, and using the Positive Predictive Value (PPV), which describes the number of true positives. Sn is obtained by: (reference prophages detected/total reference prophages) and PPV is obtained by: (reference prophages detected/(reference prophages detected + non-reference prophages detected). The alignment identity settings can be adjusted by the user of the PhageWeb, however, performance tests were based on the alignment identity set at: 80%.

TABLE 4 | Prophage regions identified by computational tools for the genome of *Lactococcus lactis* subsp. *lactis* Il1403 (NC 002662) compared to that of the lineage that was manually curated annotation.

Prophage	Reference coordinates	Phaster	Prophinder	PhiSpy	PhageWeb
Region 1	35516–49727	28461–56371	35516–49727	28818–56368	35516–72698
Region 2	447236–483244	443651–484066	451007–483244	447083–484064	447236–483552
Region 3	502723–513742	502338–520485	502723–511542	–	502723–517314
Region 4	1036642–1071558	1033815–1079175	1036642–1071558	1036482–1113152	1036642–1159446
Region 5	1414112–1456949	1414112–1457046	1439215–1446438	1415361–1457456	1415811–1456949
Region 6	2013685–2025635	1997701–2028023	2011426–2025635	–	2013685–2024681
–	False positives	–	–	633126–658623	–

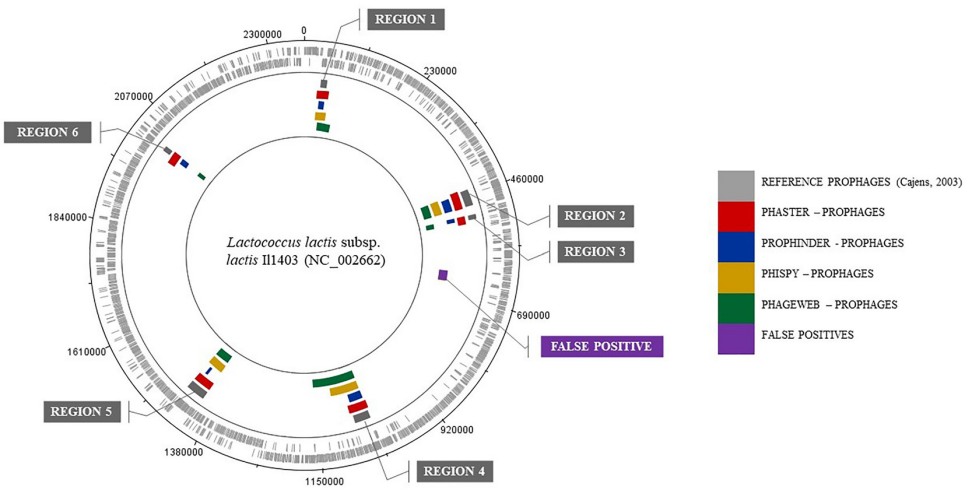


FIGURE 2 | Graphical representation of the results from analyzing the genome of *Lactococcus lactis* subsp. *lactis* Il1403 (NC 002662) by BRIG software.

RESULTS

Clustering

The reference dataset had already identified and annotated prophage regions in each genome, which had several regions of prophages. With the aid of density algorithms (Zhou et al., 2011), we identified the amount of candidate according to the reference data. The algorithm that presented the best performance in the cluster identification was HDBSCAN, followed by OPTICS; the first algorithm gave the best results in the cluster evaluation metrics. For the performance evaluation of the algorithms, four cluster evaluation metrics were used: Silhouette (Rousseeuw, 1987), Dunn (Dunn, 1974), Davies–Bouldin (Davies and Bouldin, 1979), and Density-Based Clustering Validation index – DBCV (Moulavi et al., 2014). **Table 1** shows the number of clusters identified by each algorithm and the average based on each of the four cluster-evaluation metrics. The HDBSCAN algorithm was selected to be used in our tool due to its best performance for identification of prophage in the genome.

Performance Evaluation

The comparison between PHASTER, Prophinder, PhiSpy, and PhageWeb, showed that PhageWeb was superior regarding the identification of prophages in Sensitivity (Sn) and presented positive predictive value (PPV) with the second best result compared to the other applications. For the analyzed dataset, PhageWeb reached 86.1% sensitivity and 87.3% PPV, and it is estimated that, based on the mean runtime for each analyzed genome, PhageWeb had its processing time reduced in the prediction of prophages by one-ninth of the time compared to the other tools (**Table 3**). The results of Sn and PPV for the dataset used can be observed in **Table 2**, that shows a comparison of the values.

Considering the features and performance of phage identification tools, PhageWeb presents the similar features as the others, however, allowing for more complete analysis with detailing of alignment and functional characterization of the sequences: use of G+C content evidences and tRNA regions to improve the reliability of the results and shorter execution time. Runtime values were obtained experimentally from dataset bacterial genomes. A comparative analysis of the resources available for these tools can be observed in **Table 3**. The tests performed for the collection of this resource information were performed obeying the same standard of analysis for all the tools: same input data and only features shared by all the tools were used.

In addition, they are presented to exemplify the results obtained for a prediction of prophages for the genome of *Lactococcus lactis* subsp. *lactis* Il1403 (NC_002662). **Table 4**

shows the results where the coordinates (beginning and end) of the prophage regions in the reference genome are presented, along with the results from the prediction tools. The graphical representation of this analysis through software BRIG (Alikhan et al., 2011) is shown in **Figure 2**.

CONCLUSION

Despite the efficiency of existing tools for bacterial phage analysis genomes, PhageWeb presents an efficient alternative for the identification of prophages. It has high accuracy in the prediction of these organisms as well as in the evaluation of the features and simplicity of use. It also has a graphical interface that allows better interaction and flexibility to manipulate and export the resulting data. In addition, the possibility of performing other analyzes, such as GO and metabolic pathways in the same environment, simplifies the data analysis process, reducing considerably the effort applied in the interaction with biological databases.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the PhageWeb – Dataset (http://computationalbiology.ufpa.br/phageweb/dataset/).

AUTHOR CONTRIBUTIONS

RR and AS conceived the idea of the program and together with DM, KP, EF, FA, and YP developed the tool computational. AL, AC, and JM evaluated the biological and computational information, defined the databases to be integrated and functions to be inserted. All authors reviewed the manuscript.

FUNDING

This work has been supported by the CNPq (Conselho Nacional de Pesquisa Científica) grant #421528/2016-8 and #304711/2015-2, CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), and PROPESP/UFPA (Pró-Reitoria de Pesquisa e Pós Graduação/Universidade Federal do Pará).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00644/full#supplementary-material

REFERENCES

Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). Phispy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 40:e126. doi: 10.1093/nar/gks406

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). Blast ring image generator (brig): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Amgarten, D., Braga, L. P. P., da Silva, A. M., and Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* 9:304. doi: 10.3389/fgene.2018.00304

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). Phaster: a better, faster version of the phast phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387

Bernheim, A., and Sorek, R. (2018). Viruses cooperate to defeat bacteria. *Nature* 559, 482–484. doi: 10.1038/d41586-018-05762-1

Brüssow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68, 560–602. doi: 10.1128/MMBR.68.3.560-602.2004

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Campbell, A. (2003). Prophage insertion sites. *Res. Microbiol.* 154, 277–282. doi: 10.1016/S0923-2508(03)00071-8

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage genomics. *Microbiol. Mol. Biol. Rev.* 67, 238–276. doi: 10.1128/MMBR.67.2.238-276.2003

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? molecular microbi- ology. *Nucleic Acids Res.* 49, 277–300.

Davies, D., and Bouldin, D. (1979). A cluster separation measure. *JIEEE Trans. Pattern Anal. Mach. Intell.* 1, 224–227. doi: 10.1109/TPAMI.1979.4766909

Delesalle, V. A., Tanke, N. T., Vill, A. C., and Krukonis, G. P. (2016). Testing hypotheses for the presence of trna genes in mycobacteriophage genomes. *Bacteriophage* 3:121. doi: 10.1080/21597081.2016.1219441

Dunn, J. (1974). Well-separated clusters and optimal fuzzy partitions. *J. Cybern* 4, 95–104. doi: 10.1080/01969727408546059

Eng, C., Thibessard, A., Danielsen, M., Rasmussen, T. B., Mari, J. F., and Leblond, P. (2011). In silico prediction of horizontal gene transfer in *Streptococcus thermophilus*. *Arch. Microbiol.* 193, 287–297. doi: 10.1007/s00203-010-0671-8

Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S., and Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: a review. *Virol. J.* 9:9. doi: 10.1186/1743-422X-9-9

Jafari, N., and Abediankenari, S. (2015). Phage particles as vaccine delivery vehicles: concepts, applications and prospects. *Asian Pac. J. Cancer Prev.* 16, 8019–8029. doi: 10.7314/APJCP.2015.16.18.8019

Kim, E. J., Lee, C. H., Nair, G. B., and Kim, D. W. (2010). Whole-genome sequence comparisons reveal the evolution of vibrio cholerae o1. *Trends Microbiol.* 23, 479–489. doi: 10.1016/j.tim.2015.03.010

Langille, M. G., Hsiao, W. W., and Brinkman, F. S. (2010). Detecting genomic islands using bioinformatics approaches. *Nat. Rev. Microbiol.* 8:373. doi: 10.1038/nrmicro2350

Levin, B. R., and Bull, J. J. (2004). Population and evolutionary dynamics of phage therapy. *Nat. Rev. Microbiol.* 2, 166–173. doi: 10.1038/nrmicro822

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* 24, 863–865. doi: 10.1093/bioinformatics/btn043

Lu, B., and Leong, H. W. (2016). Computational methods for predicting genomic islands in microbial genomes. *Comput. Struct. Biotechnol. J.* 14, 200–206. doi: 10.1016/j.csbj.2016.05.001

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Bacterial photosynthesis genes in a virus. *Nature* 423, 741–741. doi: 10.1038/424741a

Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nat. Rev. Genet.* 11:31. doi: 10.1038/nrg2626

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., and Sander, J. (2014). "Density-based clustering validation," in *Proceedings of the 2014 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611973440.96

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299. doi: 10.1038/35012500

Rohwer, F. (2003). Global phage diversity. *Cell* 113, 53–65. doi: 10.1016/S0092-8674(03)00276-9

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Schofield, D. A., Sharp, N. J., and Westwater, C. (2012). Phage-based platforms for the clinical detection of human bacterial pathogens. *Bacteriophage* 2, 105–121. doi: 10.4161/bact.19274

Srividhya, K. V., Alaguraj, V., Poornima, G., Kumar, D., Singh, G. P., Raghavenderan, L., et al. (2007). Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS One* 2:1193. doi: 10.1371/journal.pone.0001193

Sullivan, M. B., Waterbury, J. B., and Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium prochlorococcus. *Nature* 423, 1047–1051. doi: 10.1038/nature01929

Tozzoli, R., Grande, L., Michelacci, V., Ranieri, P., Maugliani, A., Caprioli, A., et al. (2014). Shiga toxin-converting phages and the emergence of new pathogenic *Escherichia coli*: a world in motion. *Front. Cell. Infect. Microbiol.* 4:80. doi: 10.3389/fcimb.2014.00080

Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). Making antibodies by phage display technology. *Annu. Rev. Immunol.* 12, 433–455. doi: 10.1146/annurev.iy.12.040194.002245

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. H. (2011). Phast: a fast phage search tool. *Nucleic Acids Res.* 39, 347–352. doi: 10.1093/nar/gkr485

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

BRIEF RESEARCH REPORT
published: 06 March 2019
doi: 10.3389/fgene.2019.00146

webCEMiTool: Co-expression Modular Analysis Made Easy

*Lucas E. Cardozo[1], Pedro S. T. Russo[1], Bruno Gomes-Correia[2], Mariana Araujo-Pereira[1], Gonzalo Sepúlveda-Hermosilla[3], Vinicius Maracaja-Coutinho[2] and Helder I. Nakaya[1]**

[1]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil, [2]Advanced Center for Chronic Diseases–ACCDiS, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile, [3]Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile

Co-expression analysis has been widely used to elucidate the functional architecture of genes under different biological processes. Such analysis, however, requires substantial knowledge about programming languages and/or bioinformatics skills. We present webCEMiTool,[1] a unique online tool that performs comprehensive modular analyses in a fully automated manner. The webCEMiTool not only identifies co-expression gene modules but also performs several functional analyses on them. In addition, webCEMiTool integrates transcriptomic data with interactome information (i.e., protein-protein interactions) and identifies potential hubs on each network. The tool generates user-friendly html reports that allow users to search for specific genes in each module, as well as check if a module contains genes overrepresented in specific pathways or altered in a specific sample phenotype. We used webCEMiTool to perform a modular analysis of single-cell RNA-seq data of human cells infected with either Zika virus or dengue virus.

Keywords: co-expression analysis, systems biology, transcriptomics, web tool, data integration

OPEN ACCESS

Edited by:
Akira Funahashi, Keio University, Japan

Reviewed by:
Takahiro G. Yamada, Keio University, Japan
Marco Vanoni, Università degli Studi di Milano Bicocca, Italy

****Correspondence:***
Helder I. Nakaya
hnakaya@usp.br

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

Received: *29 August 2018*
Accepted: *12 February 2019*
Published: *06 March 2019*

Citation:
Cardozo LE, Russo PST, Gomes-Correia B, Araujo-Pereira M, Sepúlveda-Hermosilla G, Maracaja-Coutinho V and Nakaya HI (2019) webCEMiTool: Co-expression Modular Analysis Made Easy. Front. Genet. 10:146. doi: 10.3389/fgene.2019.00146

INTRODUCTION

Cellular processes are driven by multiple interacting molecules whose activity level must be dynamically regulated (Kitano, 2002). As a result, genes belonging to the same signaling and metabolic pathway or sharing similar functions will tend to be co-expressed across conditions (Wang et al., 2016). Co-expression gene module analysis creates networks comprising sets of genes (i.e., modules) whose expression is highly correlated. Such analysis was applied to reveal functional modules related to infectious (Janova et al., 2015), inflammatory (Beins et al., 2016), and neurological (Voineagu et al., 2011) diseases, as well as several types of cancer (Sharma et al., 2017).

Weighted gene co-expression network analysis (WGCNA) is a widely used method to identify co-expressed gene modules (Zhang and Horvath, 2005). In order to run WGCNA, however, users are required to be familiar to programming environments, as well as to manually select parameters. These features prevent researchers with insufficient knowledge of R to identify gene modules from transcriptome data sets.

[1]https://cemitool.sysbio.tools/

Based on our Bioconductor R package named CEMiTool (Russo et al., 2018), we developed a user-friendly web-based application that allows scientists with no background in bioinformatics to perform comprehensive co-expression network analysis.

MATERIALS AND METHODS

The web interface of webCEMiTool was developed to allow users to quickly generate comprehensive analyses without the need of installing any specific program or internet browser. The only requirement for running the modular analysis is a data set containing the expression levels of all genes in samples under different biological conditions (herein defined as "classes"). There is no defined range number of samples but our previous study suggests a minimum of 15 samples per data set (Russo et al., 2018). Although it was primarily designed for transcriptome data (i.e., RNA-seq or microarrays), it can also be potentially used for identifying modules of proteins, cytokines, and even metabolites. webCEMiTool will then automatically select the input genes and identify the co-expression modules. Each module contains a set of genes whose expression follows a similar pattern.

We implemented, within webCEMiTool, a feature that assesses the activity of gene modules on each class of samples. For this, the users only have to provide a sample annotation tab-delimited text file that informs the class of each sample. A "profile plot" showing the median level of individual genes within the module is then displayed in the "Results" section of the tool (**Figure 1A**).

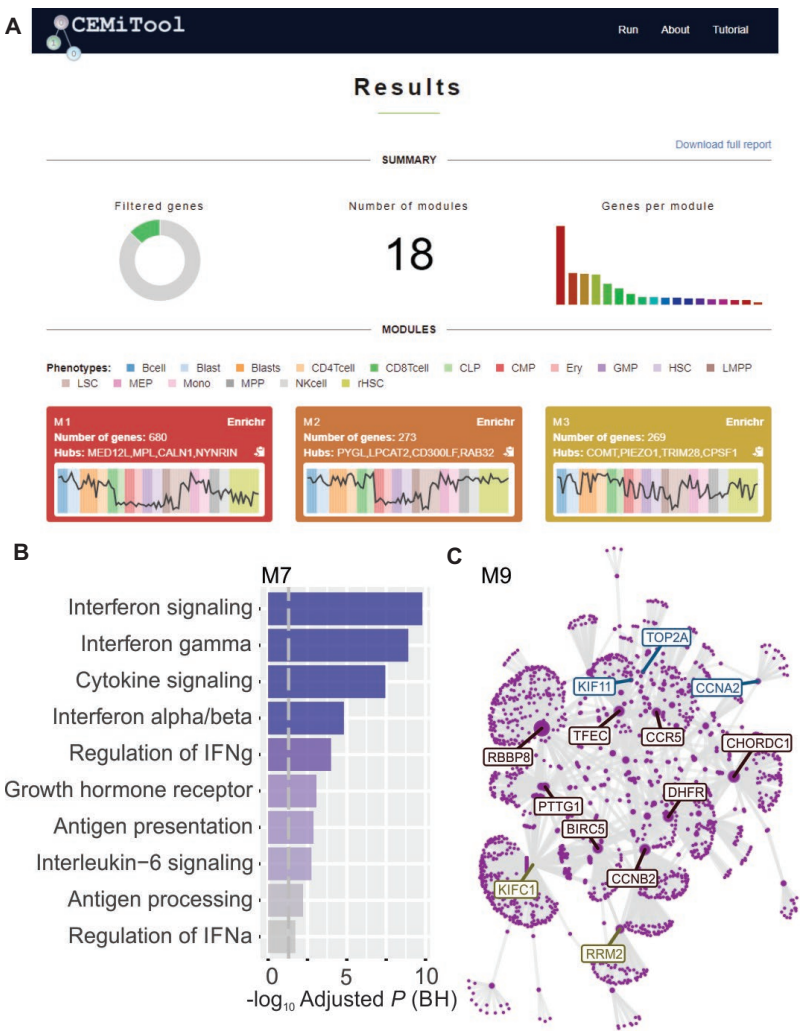


FIGURE 1 | webCEMiTool overview. **(A)** webCEMiTool results summary – The donut chart represents the proportion of selected genes by the unsupervised filter. The front page also displays the number of modules obtained, as well as a bar chart depicting the number of genes in each module. Module profile plots illustrate the median expression activity of genes from the modules across each sample. The colors represent the different sample classes. **(B)** Overrepresentation analysis – This depicts the –log10 adjusted p-value (Benjamini-Hochberg) of the enriched pathways in a module (pathways defined by user-inputted .gmt file). **(C)** Gene network of a module – The top most connected genes (hubs) are labeled and colored based on whether they were originally present in the module (blue), or inserted from a user-inputted interaction file (red), or both (green).

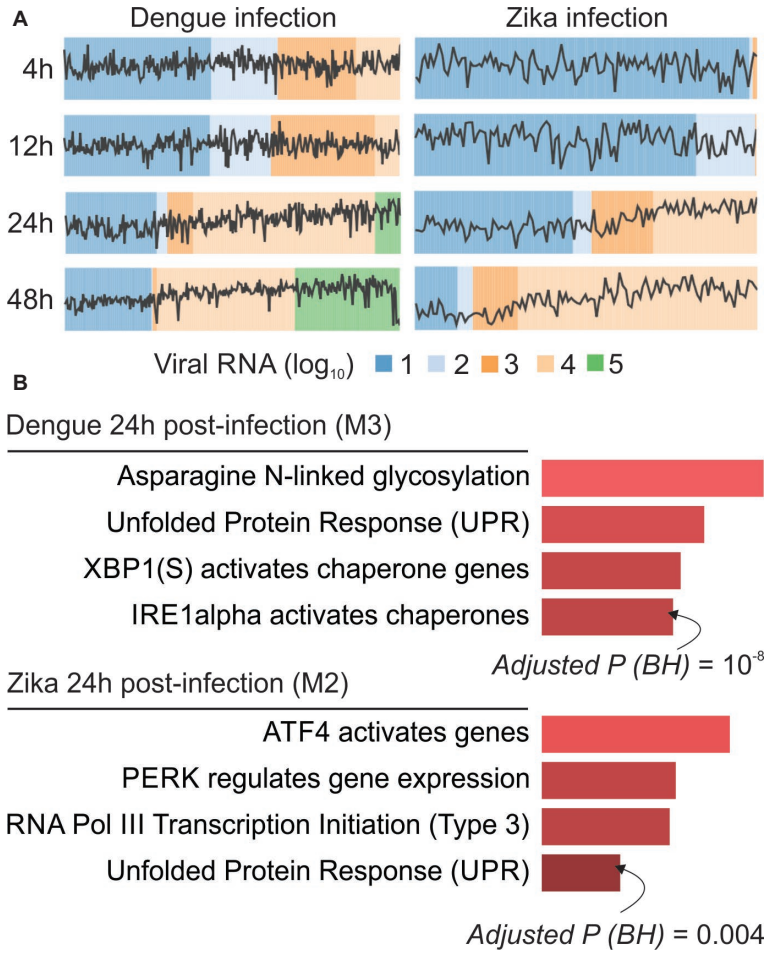


FIGURE 2 | webCEMiTool applied to single-cell RNA-seq data. **(A)** Profile plot of co-expressed gene modules. We selected one representative module for each time point post-dengue virus infection (left) or post-Zika virus infection (right). The black line represents the median expression activity of genes from the modules across each sample. The colors represent the different amount of virus RNA within the cell. **(B)** Overrepresentation analysis of selected modules at 24 h post-virus infection. The bar graphs were adapted from the Enrichr webtool linked to webCEMiTool. The bars are proportional to the –log10 adjusted p-value (Benjamini-Hochberg) of the enriched pathways in a module.

To enable functional analysis, the users can also check if the gene modules are associated with specific signaling or metabolic pathways (**Figure 1B**). These pathways can easily be extracted from databases, such as KEGG, Reactome, and MySigDB. Finally, users can integrate the results with interactome data (i.e., protein-protein interactions, transcription factors and their transcribed genes, or even miRNAs and their target genes). This feature enables users to identify critical regulators of modules (**Figure 1C**), providing valuable insights for experimental validation or potential targets for drugs. Additional details on how to obtain the optional files can be found in the "Tutorial" page of the website.[2]

To demonstrate that our method is robust, we performed an unprecedented large-scale modular analysis with over 1,000 publicly available RNA-seq and microarray data sets and new RNA-seq data of patients infected with Leishmania using the CEMiTool R package version (Russo et al., 2018). Although webCEMiTool and the package have distinct visualization features and are based on different platforms, the core co-expression functionality is essentially the same. The online tool we are describing here is built to enable easy access to gene modular analyses for non-programming researchers, while the R library version is geared towards users with greater knowledge of the R programming language. Additionally, the results dashboard is composed of interactive charts that facilitate interpretation. Moreover, taking advantage of the rising ecosystem of bioinformatics web services, our tool establishes an interface with the Enrichr platform (Chen et al., 2013), enabling a richer experience for our users.

RESULTS

We demonstrated that webCEMiTool can be applied to analyze expression data at the single cell level. Publicly available viscRNA-Seq data (virus-including single cell RNA-Seq) were obtained from NCBI GEO database (accession number GSE110496) and used as input for the analysis. The data refer

[2]https://cemitool.sysbio.tools/tutorial

to the transcriptome of individual human hepatoma (Huh7) cells, which were infected with either dengue virus (DENV) or Zika virus (ZIKV), using multiplicity of infection (MOI) 0, 1, or 10 (Zanini et al., 2018). Cells collected on four different time points (4, 12, 24 and 48 h after infection) were then sorted for single cell transcriptomic analysis with an adapted Smart-seq2 protocol (Zanini et al., 2018). The DENV data set comprises 933 infected cells (MOI = 1 or 10) and 303 controls (MOI = 0), while the ZIKV data set is composed of 488 infected cells (MOI = 1) and 403 controls. Before submitting the analysis to the webCEMiTool platform, both data sets were log10 transformed and genes that were not expressed in more than 80% of the samples were removed. The data sets were then split by virus and by time point and used as input ("Expression file" field) to webCEMiTool. In addition to the gene expression data, we also provided to webCEMiTool the sample phenotypes (i.e., viral loads) and Reactome gene sets.

Our webCEMiTool analyses generated an average of six modules per time point in DENV infection and more than eight modules per time point in ZIKV infection. We have selected one module per time point as a representative of our findings (**Figure 2A**). It is clear that at 24 and 48 h post-infection, the expression activity of representative modules increases according to the viral load (**Figure 2A**). We next performed the pathway enrichment analysis of the representative modules at 24 h post-infection using the webCEMiTool link for Enrichr (**Figure 2B**). These findings not only corroborate what was described in the original publication (Zanini et al., 2018) but also provide new insights about the physiopathology of dengue and Zika virus infections.

DISCUSSION

Although few similar web-based applications were developed to perform co-expression gene analysis (Tzfadia et al., 2016; Desai et al., 2017), these tools do not provide comparable results to webCEMiTool. One such application is GeNET (Desai et al., 2017). This webtool was designed to facilitate gene co-expression analyses and provides enrichment analysis and gene-to-gene networks. However, it only performs these analyses for three organisms (*R. capsulatus*, *M. tuberculosis*, and *O. sativa*). Another example is CoExpNetViz (Tzfadia et al., 2016), a webtool designed for the visualization and construction of gene networks. Similar to GeNET, CoExpNetViz is somewhat limited with respect to the organisms as it is stated to be primarily designed for plant transcriptomes. The webCEMiTool aims to provide co-expression analyses for any organism. Moreover, although CoExpNetViz is presented as a web-based application, its results are returned to users as a compressed folder containing a README.txt file with instructions on how to visualize their results on the Cytoscape app. The users have then to manually insert into Cytoscape the several different output files provided by the tool. These additional steps can also make the process error-prone and possibly daunting to users unfamiliar with Cytoscape. The webCEMiTool offers much more convenient browser-displayed results.

We also showed that webCEMitool is able to analyze single-cell RNA-seq data faster and efficiently. Our results returned relevant information about the biological processes involved with dengue and Zika virus infection. All this analysis were performed in an automated and practical manner, with no need for the user to have deep understanding on the internal processing of gene co-expression data analysis.

AUTHOR CONTRIBUTIONS

LC, PR, BG-C, and MA-P performed the analyses. LC, GS-H, and VM-C developed the webtool. HN conceived the tool and supervised the work. All authors help in the writing of the paper.

FUNDING

This work was supported by grants from FAPESP (2012/19278-6, 2013/08216-2, 2017/05762-7, 2018/10748-6); CNPq (313662/2017-7); FONDECYT-CONICYT (11161020); and PAI-CONICYT (PAI79170021). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

Beins, E., Ulas, T., Ternes, S., Neumann, H., Schultze, J., and Zimmer, A. (2016). Characterization of inflammatory markers and transcriptome profiles of differentially activated embryonic stem cell-derived microglia. *Glia* 64, 1007–1020. doi: 10.1002/glia.22979

Chen, E., Tan, C., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.* 14:128. doi: 10.1186/1471-2105-14-128

Desai, A., Razeghin, M., Meruvia-Pastor, O., and Peña-Castillo, L. (2017). GeNET: a web application to explore and share Gene Co-expression network analysis data. *PeerJ.* 5:e3678. doi: 10.7717/peerj.3678

Janova, H., Böttcher, C., Holtman, I., Regen, T., van Rossum, D., Götz, A., et al. (2015). CD14 is a key organizer of microglial responses to CNS infection and injury. *Glia* 64, 635–649. doi: 10.1002/glia.22955

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Russo, P., Ferreira, G., Cardozo, L., Bürger, M., Arias-Carrasco, R., Maruyama, S., et al. (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinf.* 19:56. doi: 10.1186/s12859-018-2053-1

Sharma, A., Cinti, C., and Capobianco, E. (2017). Multitype network-guided target controllability in phenotypically characterized osteosarcoma: role of tumor microenvironment. *Front. Immunol.* 8:928. doi: 10.3389/fimmu.2017.00918

Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. (2016). CoExpNetViz: comparative co-expression networks construction and visualization tool. *Front. Plant Sci.* 6:1194. doi: 10.3389/fpls.2015.01194

Voineagu, I., Wang, X., Johnston, P., Lowe, J., Tian, Y., Horvath, S., et al. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384. doi: 10.1038/nature10110

Wang, J., Xia, S., Arand, B., Zhu, H., Machiraju, R., Huang, K., et al. (2016). Single-cell co-expression analysis reveals distinct functional modules, co-regulation mechanisms and clinical outcomes. *PLoS Comput. Biol.* 12:e1004892. doi: 10.1371/journal.pcbi.1004892

Zanini, F., Szu-Yuan, P., Bekerman, E., Einav, S., and Quake, S. R. (2018). Single-cell transcriptional dynamics of flavivirus infection. *Elife* 7:e32942. doi: 10.7554/eLife.32942

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128

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: 22 March 2019
doi: 10.3389/fgene.2019.00238

croFGD: *Catharanthus roseus* Functional Genomics Database

Jiajie She†, Hengyu Yan†, Jiaotong Yang, Wenying Xu and Zhen Su**

State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing, China

Catharanthus roseus is a medicinal plant, which can produce monoterpene indole alkaloid (MIA) metabolites with biological activity and is rich in vinblastine and vincristine. With release of the scaffolded genome sequence of *C. roseus*, it is necessary to annotate gene functions on the whole-genome level. Recently, 53 RNA-seq datasets are available in public with different tissues (flower, root, leaf, seedling, and shoot) and different treatments (MeJA, PnWB infection and yeast elicitor). We used in-house data process pipeline with the combination of PCC and MR algorithms to construct a co-expression network exploring multi-dimensional gene expression (global, tissue preferential, and treat response) through multi-layered approaches. In the meanwhile, we added miRNA-target pairs, predicted PPI pairs into the network and provided several tools such as gene set enrichment analysis, functional module enrichment analysis, and motif analysis for functional prediction of the co-expression genes. Finally, we have constructed an online croFGD database (http://bioinformatics.cau.edu.cn/croFGD/). We hope croFGD can help the communities to study the *C. roseus* functional genomics and make novel discoveries about key genes involved in some important biological processes.

Keywords: *Catharanthus roseus*, co-expression network, functional module, gene function, monoterpene indole alkaloid

OPEN ACCESS

Edited by:
Vinicius Maracaja-Coutinho, Universidad de Chile, Chile

Reviewed by:
Vishal Acharya, Institute of Himalayan Bioresource Technology (CSIR), India
Dinesh Kumar, Indian Council of Agricultural Research (ICAR), India

****Correspondence:***
Wenying Xu
x_wenying@yahoo.com
Zhen Su
zhensu@cau.edu.cn

† 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: *16 October 2018*
Accepted: *04 March 2019*
Published: *22 March 2019*

Citation:
She J, Yan H, Yang J, Xu W and Su Z (2019) croFGD: Catharanthus roseus Functional Genomics Database. Front. Genet. 10:238. doi: 10.3389/fgene.2019.00238

INTRODUCTION

Catharanthus roseus, a model plant of the Apocynaceae family, is best known for production of the bis-indole monoterpene indole alkaloids (MIAs). There are four important MIAs, vinblastine and vincristine used in the clinic as anti-cancer agents (Aslam et al., 2010), catharanthine which can reduce blood sugar content (Pan et al., 2012), and vindoline. MIAs belong to a class of terpenoid indole alkaloids (TIAs). Some TIAs exhibit strong pharmacological activities, whose production has beneficial effects on human health (Almagro et al., 2015). The biosynthesis of TIAs is regulated by several key transcription factors (TFs), such as ORCA3, ORCA2, WRKY, MYC, ZCT1, and BIS, which can enhance alkaloid production (Van Der Fits and Memelink, 2000; Suttipanta et al., 2011; Zhang et al., 2011; Li et al., 2013; Van Moerkercke et al., 2015; Rizvi et al., 2016). In addition to these key TFs, some hormones and transporters are essential for the regulation of TIA biosynthesis in *C. roseus* (Liu et al., 2017). Some external signals such as elicitor and jasmonate (JA) can regulate the activities of several TFs involved in TIA biosynthesis (Memelink and Gantet, 2007). Although much progress

has been made in the field of TIAs, functions of some key genes and enzymes associated with the regulation of TIA biosynthesis are still unknown, which makes it difficult to understand the whole process. Notably, the release of the scaffolded genome sequence of *C. roseus* (Kellner et al., 2015), makes it possible to refine functional annotations of genes by integrating multidimensional data and existing methods.

The integration of biological information through gene expression profiling analysis can benefit to elucidating gene function (Noordewier and Warren, 2001). Transcriptomic datasets can be used to establish the gene expression profiles, which can provide some useful information for inferring gene regulatory relationship (Newton and Wernisch, 2014). Transcriptome analysis reveals that some genes involved in TIA biosynthesis are differentially expressed in leaf and root tissues, which can help understand specialized metabolic pathways in *C. roseus* (Verma et al., 2014). Integrated transcriptome and metabolome analysis can establish connections between genes and specialized metabolites, which can identify many genes involved in TIA synthesis and elucidate particular biological pathways (Rischer et al., 2006). Basing on transcriptomic datasets, the network construction can provide important biological knowledge, especially for digging out possible gene functions (Rhee and Mutwil, 2014).

Currently, there has been a plenty of transcriptomic datasets available on the public platform, which lay the foundation for the research in *C. roseus*. By considering all collected transcriptomic samples available together, co-expression network is applied to predicting gene functions on a large scale (Ma et al., 2014). Co-expression network analysis can mimic some important regulatory mechanism *in vivo* and thus discover key regulatory genes or functional modules. van Dam et al. (2017) excavated disease-related functional modules and annotated core genes based on co-expression network analysis. Considering that genes within a specialized metabolite pathway may form tight associations with each other in co-expression network, the method for connecting genes to specialized metabolic pathways in plant is effective, which can identify novel genes associated with specialized metabolic pathways (Wisecaver et al., 2017). Co-expression network analysis identified two missing enzymes, PAS and DPAS, necessary for vinblastine biosynthesis in *C. roseus*, which is important for understanding many other bioactive alkaloids (Caputi et al., 2018).

A growing number of studies have supported the utility of co-expression network analysis for inferring and annotating gene function, and excavating core genes involved in specific biological process. PlaNet used Heuristic Cluster Chiseling Algorithm (HCCA) to construct whole-genome co-expression networks for *Arabidopsis* and six important plant crop species (Mutwil et al., 2011). AraNet presented co-functional gene network for *Arabidopsis* and generated functional predictions for 27 non-model plant species using an orthologous-based projection (Lee et al., 2015). ATTED-II provided 16 co-expression platforms for nine plant species through combining the Pearson correlation coefficient (PCC) and mutual rank (MR) algorithm (Aoki et al., 2016). Our lab have published several functional genomics databases with co-expression network for plant species (Yu et al., 2014; You et al., 2015, 2016; Zhang et al., 2015; Tian et al., 2016; Ma et al., 2018). Besides, ccNET provided comparative gene functional analyses at a multi-dimensional network and epigenome level across diploid and polyploid *Gossypium* species based on the co-expression network (You et al., 2017). With the combination of transcriptomic and epigenomic data, MCENet provided global and conditional networks to help identify maize functional genes or modules associated with agronomic traits (Tian et al., 2018).

Here, we constructed a functional genomics database for *C. roseus* (croFGD). It provided three types of co-expression network, which allowed user to perform network search and analysis from a multi-dimensional perspective. Functional annotation information and several analysis tools were provided for functional prediction of the co-expression genes. Basing on co-expression network, we identified some functional modules which could be applied to the discovery of vital genes associated with agronomic traits. The integration of co-expression network analysis and functional module identification can be used to improve *C. roseus* gene function annotation and helpful for the functional genomics research. Besides, it can promote the research for the synthesis, metabolism of active substances and drug development.

MATERIALS AND METHODS

Transcriptomic Data Source

There were 53 samples in *Catharanthus roseus* collected from the NCBI Sequence Read Archive (SRA), which covered different tissues (root, hairy root, shoot, stem, leaf, flower, seedling, and callus) and different treatments, such as methyl jasmonate (MeJA), peanut witches' broom (PnWB) infection and yeast elicitor (**Supplementary Table S1**).

Data Processing and Gene Expression Profiling Analysis

The *C. roseus* genome had a size of ∼500 Mb, and 33,829 protein-coding genes. All transcriptomic datasets were subjected to quality control using FastQC software (v0.10.1) (Brown et al., 2017). Those datasets with mapping rate <50% were filtered out. The sequence reads were mapped to the *C. roseus* reference genome (ASM94934v1) (Kellner et al., 2015) using Tophat (v2.0.10) software (Trapnell et al., 2009) with default parameters. Cufflinks (v2.2.1) (Trapnell et al., 2010) was used to calculate the FPKM (fragments per kilobase of transcript per million mapped reads) values with default parameters. And differentially expressed genes was calculated by Cuffdiff (v2.2.1) (Trapnell et al., 2013).

Co-expression Network Construction

Pearson correlation coefficient is used to calculate correlation coefficient between two genes. MR represents high credible co-expression gene pairs after ranking the PCC. PCC is calculated based on the formula below. The more similar the expression pattern in samples between genes is, the higher the PCC score

might be. MR is an algorithm basing on PCC, which takes a geometric average of the PCC rank from gene A to gene B and from gene B to gene A.

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$MR(AB) = \sqrt{(Rank(A \rightarrow B) \times Rank(B \rightarrow A))}$$

X or Y represents the FPKM value, and n represents the number of samples. MR ensures those co-expression gene pairs with low credibility will be filtered out, so the PCC and MR are combined to construct co-expression network. Here, all samples were used for the construction of global co-expression network. Among all samples, 44 samples without treatment were used to construct tissue-preferential network, and 32 samples with treatment and corresponding control were used to construct the treat-response network.

Functional Module Identification and Parameter Selection

The Clique Percolation Method (CPM) (Adamcsek et al., 2006) was used to identify modules with nodes densely connected to each other in three types of co-expression networks, including global network, tissue-preferential network and treat-response network. Parameter selection was based on module number, module overlap rate and gene coverage rate. Here, we selected the k = 5 clique size for global co-expression network, which meant each module had at least five nodes and each node had co-expression relationship with each other (**Supplementary Figure S2**). In fact, one functional module could be regarded as a small network. Similarly, we selected the k = 6 clique size for tissue-preferential network and treat-response network. The functions of the modules were annotated through gene set enrichment analysis (GSEA) (Yi et al., 2013), including GO terms, gene families, plantCyc and KEGG pathways.

The Identification of Orthologous Genes in *Arabidopsis*

Bidirectional blast alignments were conducted for the analysis of protein sequences between *C. roseus* and *Arabidopsis*. Our criteria for the identification of orthologous gene pairs were as follows: the top three hits in each bidirectional blast alignment were selected as the best orthologous pairs; in addition, orthologous pairs with an e-value less than 1E-25 were regarded as the second level.

The Classification of Gene Family

Five main gene families, including TFs and regulator factors (TRs), carbohydrate-active enzymes, kinase, ubiquitin and cytochrome P450, were classified to improve limited functional annotation. TF/TRs and kinase family were identified mainly by iTAK tool (Zheng et al., 2016) based on the rule in PlnTFDB (Pérez-Rodríguez et al., 2009) and PlantsP Kinase Classification (Tchieu et al., 2003), respectively. The carbohydrate-active enzymes (CAZy) family (Lombard et al., 2014) was predicted through the method of orthologous search based on *Arabidopsis thaliana*. The enzymes were classified into six groups: glycoside hydrolases (GH), glycosyltransferase (GT), polysaccharide lyases (PL), carbohydrate esterase (CE), auxiliary activities (AA) and carbohydrate-binding modules (CBM). Ubiquitin family was identified through Hidden Markov Model (HMM) search based on models from UUCD (Gao et al., 2013). And cytochrome P450 family was predicted by orthologous relationship with *Arabidopsis* and the candidates were confirmed with ID of PF00067 by Pfam (Finn et al., 2014) search.

Z-Score for Motif Analysis

Motif (*cis*-element) analysis tool is developed to identify significant motifs in one sequence or in the promoter region of interested gene list and thus predict possible functions. Z-score is a statistical measurement of the distance in standard deviations of a sample, which can act as a normalization method to eliminate the difference caused by background for different samples. So far, it is widely applied to calculating the *cis*-element significance (Endo et al., 2014).

The Z-score is calculated as:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$\bar{X}$ represents sum value of a motif in the promoter of one gene list. μ represents mean value of the same motif in 1,000 random gene lists with same scale. σ represents standard deviation of the 1,000 mean value based on random selection.

Plant Materials and Growth Conditions

C. roseus seeds were planted in small pots and kept moistened until the seeds had germinated, and allowed to grow until they had three to five leaves, then transferred to a greenhouse (16 h light/8 h darkness, 28/25°C). For MeJA treatment, 100 μM MeJA was sprayed evenly on leaves and stem of well-growth plants. In order to prevent MeJA decomposition, leaves and stem with treatment and corresponding control were under darkness. After treatment for 6 and 24 h, the leaves and stem were harvested, immediately frozen in liquid nitrogen, and then stored at −80°C for use. Control samples were also harvested. Three biologically repeated samples were harvested.

RNA Isolation and Quantitative Real Time RT-PCR

About 100 mg of tissue was ground in liquid nitrogen before isolation of the RNA. Total RNA was isolated using TRIZOL® reagent (Invitrogen, Carlsbad, CA, United States) and purified using Qiagen RNeasy columns (Qiagen, Hilden, Germany). Reverse transcription was performed using Moloney murine leukemia virus (M-MLV; Invitrogen). We heated 10 μL samples containing 2 μg of total RNA, and 20 pmol of random hexamers (Invitrogen) at 70°C for 2 min to denature the RNA and then chilled the samples on ice for 2 min. We added reaction buffer and M-MLV to a total volume of 20 μL containing 500 μM dNTPs, 50 mM Tris-HCl (PH 8.3), 75 mM KCl, 3 mM $MgCl_2$, 5 mM dithiothreitol, 200 units of M-MLV and 20 pmol random

hexamers. The samples were then heated at 42°C for 1.5 h. The cDNA samples were diluted to 2 ng/μL for real time RT-PCR analysis.

For quantitative real-time RT-PCR, triplicate quantitative assays were performed on 1 μL of each cDNA dilution using the SYBR Green Master Mix with an ABI 7900 sequence detection system according to the manufacture's protocol (Applied Biosystems). The gene-specific primers were designed using PRIMER3[1]. The amplification of 18S rRNA was used as an internal control to normalize all data (forward primer, 5′-CGGCTACCACATCCAAGGAA-3′; reverse primer, 5′-TGTCACTACCT CCCCGTGTCA-3′). Gene-specific primers were listed in **Supplementary Table S2**. The relative quantification method ($\Delta\Delta$CT) was used to evaluate quantitative variation between replicates examined.

CONSTRUCTION AND CONTENT

Database Construction

The database was constructed under the LAMP (Linux + Apache + Mysql + PHP) environment. It mainly contains three parts: (I) functional annotation, which includes gene family, KEGG pathway and miRNA detailed information, etc.; (II) network and module, including co-expression network search and analysis, network comparison and module search; (III) some analysis tools, mainly including *cis*-element enrichment analysis, GSEA, functional module enrichment analysis and UCSC Genome Browser visualization (**Figure 1**).

Functional Annotation

We obtained the functional annotation information in *C. roseus* from the Dryad Digital Repository (Kellner et al., 2015). Among 33,829 protein-coding genes, 14,527 genes were annotated with 4,734 GO terms by blast2GO (Conesa and Gotz, 2008). 5,571 enzymes involved in 213 metabolism pathways were annotated by GhostKOALA (Kanehisa et al., 2016) from KEGG database. We mapped *C. roseus* protein sequences against CathaCyc (Van Moerkercke et al., 2013) using the BLASTP program and 2,421 enzymes involved in 513 metabolism pathways were annotated. Then we predicted 36,882 orthologous pairs between *C. roseus* and *Arabidopsis* through bidirectional blast alignment. There were a total of 1,035 plant motifs collected from the Plant *Cis*-acting Regulatory DNA Elements (PLACE) database (Higo et al., 1999), PlantCARE database (Rombauts et al., 1999), AthaMap database (Steffens, 2004) and literatures. Furthermore, we adopted the inparanoid algorithm (Sonnhammer and Östlund, 2015) and predicted 9,377 protein–protein interaction (PPI) pairs in *C. roseus* from over 18,000 experimentally validated PPI pairs in *Arabidopsis* integrated from several databases, such as BIOGRID (Chatr-Aryamontri et al., 2017), IntAct (Orchard et al., 2014) and related literature (Lumba et al., 2014). We also collected 227 miRNA sequence information derived from a literature (Shen et al., 2017), and then mapped these miRNA sequences against the whole-genome sequence using the GMAP program (Wu and Watanabe, 2005). Furthermore, 143 miRNA targets were identified by psRNATarget (Dai and Zhao, 2011). The miRNA detailed information mainly included location, sequence and structure, miRNA target and expression profiles in seedling after MeJA treatment (**Supplementary Figure S3**). Furthermore, we conducted the gene family classification and finally predicted 88 TFs/TRs families with 1,702 genes, 21 ubiquitin families with 1,192 genes, 98 cytochrome P450 families with 191 genes, 85 kinase families with 778 genes and 96 CAZy families with 1,505 genes (**Table 1**).

Co-expression Network and Functional Module

A well-developed strategy with the integration of PCC and MR algorithm was widely applied to the construction of co-expression network (You et al., 2016, 2017; Obayashi et al., 2018; Tian et al., 2018). We used the 240 BP terms of GO associated with >4 and <20 genes to evaluate the networks. To get optimal gene pairs and evaluate the credibility of co-expression network, we selected different PCC thresholds of PCC > 0.7, PCC > 0.8, PCC > 0.9 and different MR thresholds of MR top3 + MR ≤ 30, MR top3 + MR ≤ 50, MR top3 + MR ≤ 100 to predict gene functions basing on selected GO terms and generated receiver operating characteristic (ROC) curves (**Supplementary Figure S1**). The larger the area under the curve (AUC) value of co-expression network is, the higher the credibility of the network will be. Finally, we selected the thresholds of PCC > 0.7 and MR top3 + MR ≤ 30 to filter out those co-expression gene pairs with low credibility to construct co-expression network. In total, there were 30,096, 29,808 and 30,541 nodes in global network, tissue-preferential network and treat-response network with gene expression view, which covered 88.9%, 88.1%, and 90.3% of genes in *C. roseus*, respectively (**Table 1**). All networks were visualized by Cytoscape 2.8 (Smoot et al., 2011).

Then we overlaid the gene expression value onto the co-expression network to identify whether genes in the network were expressed or not based on the minimum threshold FPKM value. To determine the minimum threshold of the gene expression value (FPKM) among all *C. roseus* samples (detailed mapping results are shown in **Supplementary Table S3**), the lowest 5% of all gene FPKM values in each sample and the standard deviation (SD) of each experimental group were computed. The mathematical formula "threshold = average (5% value) + 3 * SD" (You et al., 2016, 2017) was used to calculate the minimum expression value of each experimental group. The minimum threshold of FPKM was 0.094. We identified differential expressed genes between treatment and control samples by the cutoff: $|\log_2 FC| \geq 1$ and p-value ≤ 0.05. Tissue-preferential analysis in different tissues (root, hairy root, shoot, stem, leaf, flower, seedling, and callus) and treat-response analysis under three types of treatments (MeJA, PnWB infection and yeast elicitor) among five tissues (root, shoot, flower, callus, and hairy root) were supplied for the co-expression network analysis. Meanwhile, predicted miRNA target and PPI pairs were integrated into the network, and further analysis was provided for all members in the network,

[1]http://frodo.wi.mit.edu/primer3/input.htm

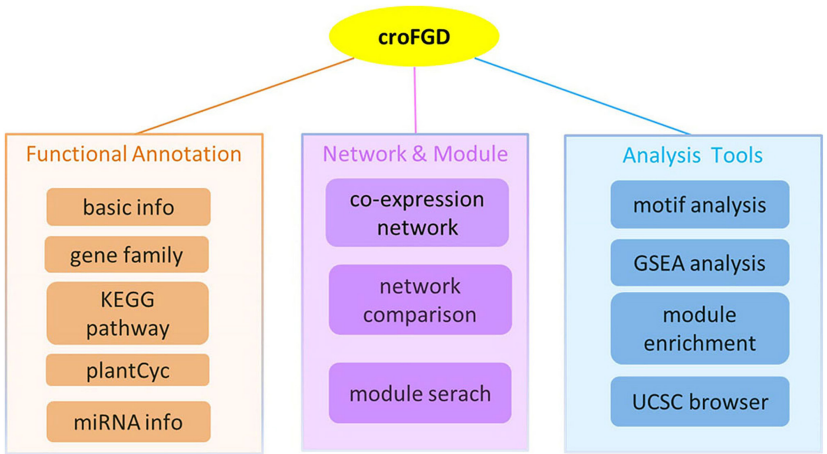


FIGURE 1 | Database architecture. The croFGD database is divided into three main pieces: network & module, functional annotation, and analysis tools. The line with different colors indicates different pieces.

such as gene expression profiling analysis, GSEA, and *cis*-element analysis.

Furthermore, co-expression network could be used to perform modularized analysis and excavation for the discovery of agronomic trait-related vital gene and functional module. The CPM proposed to detect the overlapping communities in the complex network (Palla et al., 2005; Li et al., 2014), provided certain practicability for the discovery of key gene and module. Finally, we applied the algorithm and predicted 2,310, 1,849, and 2,177 functional modules in global network, tissue-preferential network and treat-response network in *C. roseus*, respectively (**Table 1**). The functions of these modules were annotated through GSEA (Yi et al., 2013). The entries which were not significant were filtered out by Fisher's tests and multiple test correction method (FDR $\leq$ 0.05). These functional modules covered diverse functions such as vindoline and vinblastine biosynthesis, jasmonic acid biosynthesis, pathogen resistance and hormone response, etc.

Analysis Tools

Gene Set Enrichment Analysis

Gene set enrichment analysis (Yi et al., 2013) is a powerful method for the functional annotation of interested gene list by computing the overlaps with well-defined background gene sets. Some categories of gene sets, such as GO terms, gene families, plantCyc and KEGG pathways, miRNA targets and functional modules identified from three types of network, were used as background gene sets. The significantly enriched gene set with FDRs $\leq$ 0.05 would be displayed on the GSEA result page.

Functional Module Enrichment Analysis

The tool was used to identify some functional modules from interested gene list especially in the network. The previously annotated miRNA target modules and functional modules identified from three types of network were used as background functional modules. The modules with FDRs $\leq$ 0.05 would be regarded as significantly enriched and the enrichment analysis result page included module annotation, module source, overlap gene number, and FDR value.

TABLE 1 | Data collection and statistics in croFGD.

Database content	Number	Source	Reference
GO terms (genes)	55,505 (14,527)	Blast2GO tool	Conesa and Gotz, 2008
KEGG pathway (genes)	213 (5,571)	GhostKOALA tool	Kanehisa et al., 2016
PlantCyc (genes)	513 (2,421)	Blastp prediction	–
Cis-elements (motifs)	1,035	Database and literature collection	–
Orthologous pairs in *Arabidopsis* (genes)	36,882 (14,719)	Blast alignment	–
Transcription factor and regulators (members)	88 (1,702)	iTAK prediction	Zheng et al., 2016
Kinases (members)	85 (778)		
Carbohydrate-active enzymes (members)	96 (1,505)	Blast alignment	Lombard et al., 2014
Ubiquitin (members)	21 (1,192)	Blast alignment	Zhou et al., 2018
Cytochrome P450 (members)	98 (191)	the cytochrome p450 homepage	Nelson, 2009
Co-expression network nodes (%)	30,096 (88.9%)	PCC and MR	Aoki et al., 2016
Tissue-preferential network nodes (%)	29,808 (88.1%)		
Treat-response network nodes (%)	30,541 (90.3%)		
Protein–protein interaction pairs	9,377	InParanoid algorithm	Sonnhammer and Östlund, 2015
miRNA target modules	143	psRNAtarget prediction	Dai and Zhao, 2011
Function modules from global network (nodes)	2,310 (10,757)	CFinder tool	Adamcsek et al., 2006
Function modules from tissue-preferential network (nodes)	1,849 (12,090)		
Function modules from treat-response network (nodes)	2,177 (12,073)		

Cis-Element Enrichment Analysis

Cis-element (motif), a short conserved sequence, can be recognized by some TFs to regulate the expression levels of downstream genes. The tool was developed to identify motifs in a set of gene promoters and thus predict the function of gene set. The *cis*-element significance test is an algorithm using statistical method based on Z-score and *p*-value filtering (Yu et al., 2014) that can identify significant *cis*-regulatory elements in the promoter region of one gene. The promoter region was set as 3 kb in *C. roseus*. When scanned in the 3 kb promoter region of *C. roseus* genes, motifs with p-value ≤ 0.05 were significantly enriched on account of the frequency of motif occurrence.

Other Tools Supported in croFGD

A quick search, UCSC Genome Browser (Speir et al., 2016) visualization and a manual were provided for users. The search page mainly included gene detail search, gene function search, functional module search and orthologous search. The orthologous search allowed user to input one gene list in *Arabidopsis* to search for corresponding *C. roseus* genes.

FUNCTION APPLICATION

Comprehensive Exploration for the Function of 16OMT Gene

CRO_T004356 (*16OMT*), o-methyltransferase family member, which was reported to be involved in the biosynthesis of TIAs (Pandey et al., 2016; Yamamoto et al., 2016). Taking *16OMT* gene as an example, we explored possible function of the gene through the database. By gene detail search, we found that the gene: (I) was annotated with alkaloid biosynthetic process (GO: 0009821) and myricetin 3′-*O*-methyltransferase activity (GO: 0033799), etc.; (II) had two pfam domains: "Dimerisation (PF08100)" and "Methyltransf_2 (PF00891)" domains; (III) was mainly involved in vindoline and vinblastine biosynthesis; (IV) was relatively high in expression in leaf tissue (**Figure 2A**). We conducted network analysis for three types of co-expression network of *16OMT* gene including tissue-preferential network (**Figure 2B**), global network (**Figure 2C**) and treat-response network (**Figure 2D**). GSEA results for global network genes indicated that these genes might be involved in phenylpropanoid biosynthesis, vindoline and vinblastine biosynthesis. Network comparison results suggested that it was relatively conservative between global network and tissue-preferential network (**Figure 2E**), and there were great differences between global network and treat-response network (**Figure 2F**). Through module search, the gene in the module (**Figure 2G**) might be involved in vindoline and vinblastine biosynthesis, alkaloid biosynthetic process, and protein phosphorylation, *etc*. Therefore, *16OMT* gene might have diverse function in several biological processes like *hos1* gene (MacGregor and Penfield, 2015). The expression heatmaps of all genes in the module were included (**Figure 2H**). UCSC genome browser visualization (**Figure 2I**) indicated that most RNA-seq peaks were enriched in the genic region. Furthermore, stilbenoid, diarylheptanoid, and gingerol biosynthesis pathway was shown (**Figure 2J**).

Co-expression Network Analysis for CPR Gene

CPR, NADPH–cytochrome P450 reductase, which is essential for the activation of cytochrome P450 enzymes, is critical for the biosynthesis of MIAs (Parage et al., 2016). The detailed information of all genes in the global network of *CPR* gene (**Figure 3A**) was listed in **Supplementary Table S4**. In the *CPR* network, some genes (*GES*, *7DLH*, *GOR*, *HDS*, *G8H*, *ISY*, *MCS*, *HDR*, *7DLGT* and *IO*) were involved in MIA biosynthesis pathway (Chebbi et al., 2014; Kumar et al., 2015). These genes were labeled with bold in the MIA biosynthesis pathway (**Figure 3C**). Through GO enrichment analysis (Tian et al., 2017) for all genes in the *CPR* network, the significantly enriched GO terms were associated with terpene biosynthetic process, and isoprenoid biosynthetic process (**Figure 3B**), which were related to MIA biosynthesis (Geu-Flores et al., 2012; Dugé de Bernonville et al., 2015). Through module enrichment analysis for all genes in *CPR* network, three genes (*CYP76C*, *CRO_T015823*, and *CRO_T014922*) in significantly enriched functional modules might be involved in brassinosteroid (BR) biosynthesis, gibberellic acid (GA) response and indole alkaloid biosynthesis, respectively (**Figure 3D**). Therefore, in addition to MIA biosynthesis, *CYP76C* and *CRO_T015823* also played important role in plant growth and development. Besides, *CRO_T014922* might also be involved in MIA biosynthesis together with other genes (*CRO_T019924*, *CRO_T030883*, *CRO_T015465,* and *CRO_T025273*) in the module (**Figure 3D**). Thus, in addition to the function of network, co-expressed genes might be involved in some other functions. Furthermore, co-expression analysis can be combined with module enrichment analysis to predict gene function effectively.

Network Comparison Between Global Network and Tissue-Preferential Network of JAZ1 Gene

JAZ1, a jasmonate-zim-domain protein, was discovered as repressors of jasmonate signaling, which was involved in TIA biosynthesis (Pan et al., 2018). We conducted network comparison between global network and tissue-preferential network of *JAZ1* (**Figure 4A**). The information of co-expressed genes in global network and tissue-preferential network was shown in **Supplementary Table S5**. We found that the two networks displayed different network structure. There were nine overlapped genes including *JAZ1* gene between two networks. Fifteen unique genes (including *TIFY*, *CYP94C,* and *JAZ3*) appeared in global network, while sixteen unique genes including *MYB15* appeared in tissue-preferential network. GSEA results for the genes in global network of *JAZ1* indicated that some gene sets were significantly enriched, such as jasmonic acid biosynthesis, alpha-linolenic acid metabolism, steroid biosynthesis and plant hormone signal transduction (Menke et al., 1999; Koo et al., 2014; Patra et al., 2018). GSEA results for the genes in tissue-preferential network of *JAZ1* illustrated that some gene sets were significantly enriched, such as jasmonic acid biosynthetic process, 12-oxophytodienoate reductase activity, NADPH dehydrogenase

FIGURE 2 | Comprehensive explorations for the function of *16OMT* (*CRO_T004356*) gene. **(A)** The detailed information of *16OMT* gene in *C. roseus*. Three types of co-expression network, including tissue-preferential network **(B)**, global network **(C)** and treat-response network **(D)**. In these networks, the node with yellow color represents the gene submitted initially, and the nodes with green color represent co-expressed genes; the edge with pink color links two genes with positive co-expression relationship; the edge with blue color links two genes with negative co-expression relationship. **(E)** Network comparison between global network and tissue-preferential network. The nodes with yellow color represent overlap genes between two networks, and the nodes with green and dark green color stand for unique genes in two networks, respectively. **(F)** Network comparison between global network and treat-response network. **(G)** The "CFinderADM000741" module. **(H)** Expression heatmaps of genes in "CFinderADM000741" module. **(I)** UCSC genome browser visualization. **(J)** Stilbenoid, diarylheptanoid, and gingerol biosynthesis pathway.

activity, triglyceride lipase activity and oxylipin biosynthetic process (**Figure 4B**) (Tani et al., 2008; Wallström et al., 2014; Wang et al., 2018). Based on the structure and function of the two networks of *JAZ1* gene, there were some conservation and differences between two networks. In *Arabidopsis*, cytochrome p450 family member *CYP94C1* and *CYP94B3* played important role in the regulation of jasmonate response (Niu et al., 2011; Heitz et al., 2012; Koo et al., 2014). In *Gossypium hirsutum*, *GhJAZ2* regulated the jasmonic acid signaling pathway by interacting with the R2R3-MYB transcription factor GhMYB25 (Hu et al., 2016). It needed further study whether the two genes *CYP94C* and *MYB15* coexpressed with *JAZ1* in two networks had

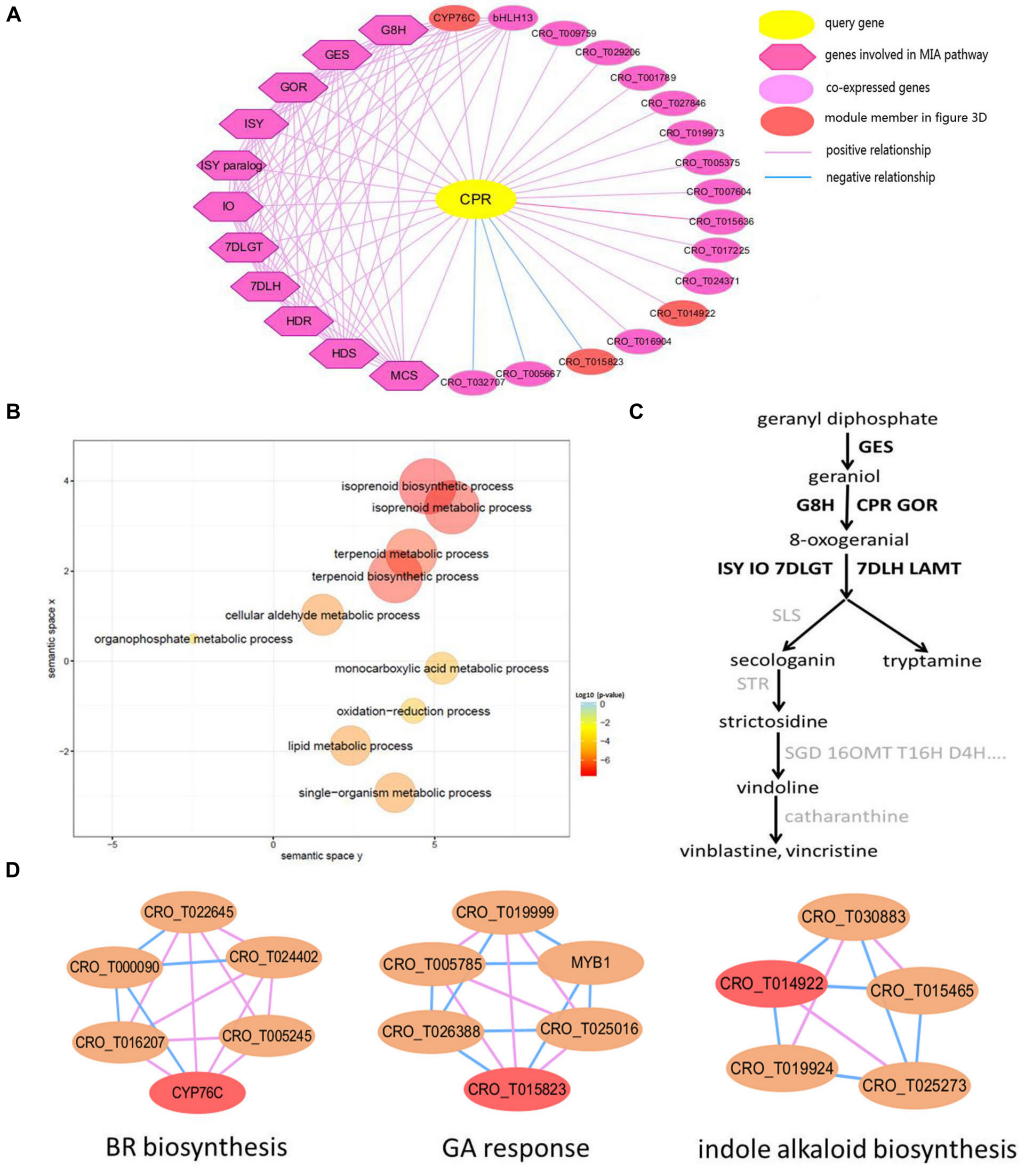


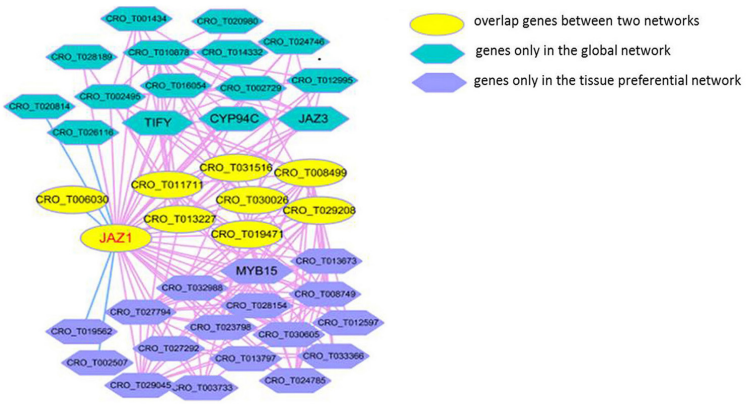
FIGURE 3 | The global network of *CPR* (*CRO_T031702*) gene involved in MIA pathway. **(A)** Global network of *CPR* gene. The query gene *CPR* is highlighted by yellow, the blue line represents negative co-expression relationship between two genes, while the pink line represents positive co-expression relationship. The dark purple diamond represents several genes involved in the MIA biosynthesis pathway, such as *GES*, *7DLH*, *GOR*, *HDS*, *G8H*, *ISY*, *MCS*, *HDR*, *7DLGT*, and *IO*, which are co-expressed with *CPR* gene in the network, and the light purple circular represents other genes co-expressed with query gene. **(B)** Scatter plot of GO enrichment analysis results for all genes in *CPR* network. **(C)** The simplified MIA pathway. The bold represents the gene in *CPR* co-expression network. **(D)** Several functional modules related to genes in *CPR* network. The red node represents genes in *CPR* network.

similar function in *C. roseus* as in *Arabidopsis* and *Gossypium hirsutum*, respectively. These results indicated that network comparison is an effective approach to analyze gene function from the perspective of different networks.

Treat-Response Network With Expression View After MeJA Treatment

In *JAZ1* network with expression view after MeJA treatment in different tissues (shoot, root, hairy root, and seedling) (**Figure 5**), most genes had significant change in expression, such as *JAZ1*, *JAZ3*, *CYP94C*, *MYB*, *MYB15*, and *TIFY*. Detailed information for up and down-regulated genes in these networks was shown (**Supplementary Table S6**). In *C. roseus*, JAZ proteins could repress MYC2 and BIS1 to respond to JA signaling and then modulate MIA biosynthesis (Patra et al., 2018). In rice, enhanced expression of cytochrome p450 family member CYP94C2b could alleviate the jasmonate response and enhanced salt tolerance (Kurotani et al., 2015). In *Arabidopsis*, AtMYB44 could repress JA-mediated defense by

A



B

Category	Description	G	T	FDR(G)	FDR(T)
TF family	**transcription_related, transcription factor: Tify**			5.88E-07	9.56E-03
plantCyc	**jasmonic acid biosynthesis**			4.03E-05	0.038
KEGG pathway	**alpha-Linolenic acid metabolism**			2.95E-06	0.035
KEGG pathway	**steroid biosynthesis**			-	0.035
KEGG pathway	sesquiterpenoid and triterpenoid biosynthesis			-	0.035
KEGG pathway	**plant hormone signal transduction**			0.025	-
P450 family	**CYP94C**			0.012	-
kinase family	PPC:1.15.1: receptor like cytoplasmic kinase I			0.014	-
GO:0003987	acetate-CoA ligase activity			0.049	-
GO:0046423	allene-oxide cyclase activity			0.049	-
GO:0047077	photinus-luciferin 4-monooxygenase (ATP-hydrolyzing) activity			0.049	-
GO:0003714	transcription corepressor activity			0.049	-
GO:0052634	C-19 gibberellin 2-beta-dioxygenase activity			0.049	-
GO:0004765	shikimate kinase activity			0.049	-
GO:0004146	dihydrofolate reductase activity			0.049	0.013
GO:0016629	**12-oxophytodienoate reductase activity**			0.049	0.013
GO:0003824	catalytic activity			0.049	0.013
GO:0003959	**NADPH dehydrogenase activity**			0.049	0.013
GO:0004806	**triglyceride lipase activity**			-	0.013
GO:0009695	**jasmonic acid biosynthetic process**			-	0.027
GO:0048583	regulation of response to stimulus			-	0.027
GO:0031408	**oxylipin biosynthetic process**			-	0.027
GO:0010181	FMN binding			-	0.048

FIGURE 4 | The network comparison between the global network and tissue-preferential network. **(A)** The comparison between the global network and tissue-preferential network of *JAZ1* (*CRO_T006982*). In the network comparison, the nodes with yellow color represents overlap genes between two networks, and the nodes with dark blue color represents genes only in tissue specific network, while the nodes with sky blue color represents genes only in global network. **(B)** GSEA results for two networks of *JAZ1*. The "G" represents global network, and the "T" represents tissue-preferential network.

activating the expression of *WRKY70* at transcriptional level (Shim et al., 2013). *PvTIFY10C* and *GsTIFY10* gene acted as a repressor in the JA signaling pathway in *Phaseolus vulgaris* and *Glycine soja* (Zhu et al., 2011; Aparicio-Fabre et al., 2013), respectively. We conferred that *CYP94C*, *MYB*, *MYB15*, and *TIFY* co-expressed with *JAZ1* might act as JA-response candidate genes in *C. roseus*. Furthermore, *CRO_T012104* (anthranilate synthase beta subunit), *CRO_T013473* (protein of unknown function), *CRO_T010878* (alpha/beta-hydrolases superfamily protein), *CRO_T002729* (allene oxide cyclase), and *CRO_T002624* (tryptophan biosynthesis) almost up-regulated under those five conditions, might also act as JA-response candidate genes. Taking treat-response network of *JAZ1* gene as an example, we selected six genes (*JAZ1*, *TIFY*, *MYB*, *CRO_T012104*, *CRO_T024124*, and *CRO_T002729*) for the real time RT-PCR validation (**Supplementary Figure S4**). These genes were up-regulated after MeJA treatment in shoot tissues and might act as JA-response genes. The qRT-PCR results indicated that these genes acted as JA-response genes in shoot tissues. This not only validated the accuracy of the predicted results, but also demonstrated the reliability of the network. Thus, treat-response network with expression view can clear display the dynamic change of gene expression in a network. Therefore, the co-expression network with multi-dimensional analysis can benefit to analyzing regulatory mechanisms in *C. roseus* development and stress response.

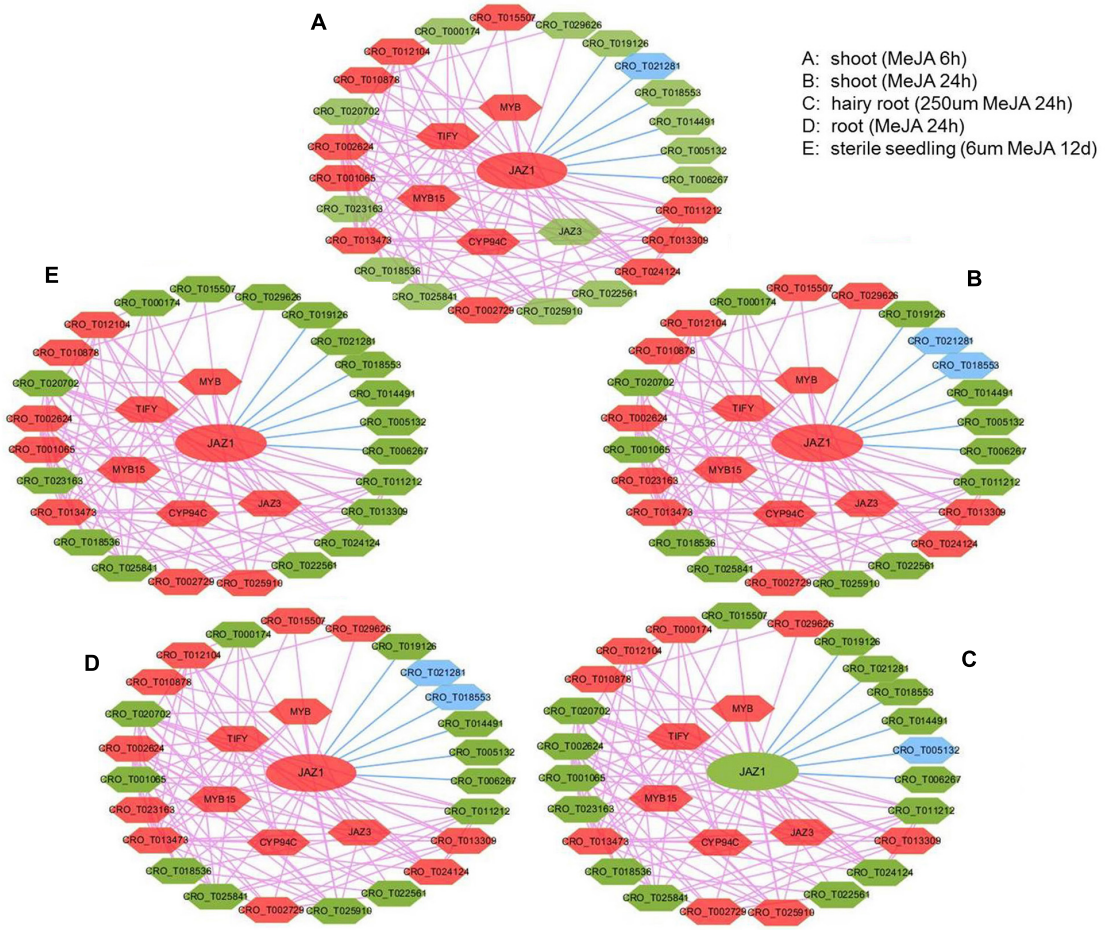


FIGURE 5 | The expression view of *JAZ1* network after MeJA treatment in different tissues (shoot, root, hairy root, and seedling). The *JAZ1* network after MeJA treatment 6 h in shoot **(A)**, MeJA treatment 24 h in shoot **(B)**, 250 μM MeJA treatment 24 h in hairy root **(C)**, MeJA treatment 24 h in root **(D)**, 6 μm MeJA treatment 12 days in sterile seedling **(E)**. The red hexagon represents up-regulated genes, the blue hexagon represents down-regulated genes, and the green hexagon represents genes with no significant change in expression.

DISCUSSION

Our croFGD database aims to provide an online database server for the annotation and prediction of gene function. We constructed global network, tissue-preferential network and treat-response network with expression view, which covered almost 90% of gene in *C. roseus* and identified more than 6,000 functional modules. The annotation of these functional modules covered vindoline and vinblastine biosynthesis, jasmonic acid biosynthesis, hormone response and pathogen resistance, etc. The network analysis strategy, functional module annotation and integrated method could improve and refine gene function annotation from diverse perspectives to some extent. For some crops, it could be applied to excavate important functional module related to agronomic traits, which would be beneficial for genetic breeding.

Through some analysis tools supported in croFGD, we can excavate key genes involved in some important biological processes and predict gene function. In comprehensive exploration for the function of *16OMT* (**Figure 2**), we found that the gene might have complex function, like *hos1* gene (MacGregor and Penfield, 2015). In global network of *CPR*, some genes were involved in MIA biosynthesis, such as *GES*, *7DLH*, *GOR*, *G8H*, *ISY*, and *7DLGT* (**Figure 3A**). The integration of co-expression network analysis and module enrichment analysis can be benefit to predicting gene function effectively and refining gene annotation. Basing on network comparison between two networks of *JAZ1*, there were certain similarities and differences whether in the structure or in the function of two networks (**Figure 4**). In addition, function of two genes *CYP94C* and *MYB15* needed further research. In treat-response network of *JAZ1* gene with expression view after MeJA treatment in different tissues, we identified several possible JA-response candidate genes (**Figure 5** and **Supplementary Table S6**), which was experimentally validated by real time RT-PCR (**Supplementary Figure S4**). These results would be beneficial to understanding some molecular regulatory mechanisms in *C. roseus*, such as MIA biosynthesis and jasmonic acid biosynthesis, *etc*.

Comparative co-expression network analysis between species is an effective approach to predict gene function and improve functional annotation (Pathania et al., 2016). We conducted network comparison for gene list with PCC ranks in the top 300 between *C. roseus* and *Arabidopsis* (obtained from ATTED-II) (**Supplementary Figure S5**). High similarity between co-expression network of *JAZ1* in *C. roseus* and *AT1G19180* (*JAZ1*) in *Arabidopsis* not only demonstrated the reliability of co-expression network, but also illustrated the conservation of *JAZ1* gene function between these two species.

Based on co-expression network with multi-dimensional level, predicting functional module and refining gene function is an effective strategy, which can be used to identify more key genes and regulatory modules when we focus on a detailed biological process. Interestingly, co-expression network is highly associated with the regulation of epigenetic modification, such as DNA methylation (El-Sharkawy et al., 2015) and H3K4me3 (Farris et al., 2015), which can be integrated to understand detailed molecular mechanism, such as the biosynthesis of specific metabolites. There is a certain correlation between co-expression network and metabolic network, the integration of which can be used to predict key enzyme-coding genes and metabolites (Chen et al., 2013), and contribute to better understanding of the molecular mechanisms related to plant metabolic pathway (Rischer et al., 2006; Coneva et al., 2014).

Notably, there are additional limitations and possible improvements for croFGD database. Firstly, the release of the chromosome-level genome of *C. roseus* in the future, will greatly promote the research on functional genomics. Secondly, more RNA-seq samples of other tissues and treatments could be integrated into the co-expression network construction on the transcriptomic level, which will be beneficial to excavate gene function and improve the whole genome annotation in *C. roseus*. Thirdly, epigenomic data, such as ChIP-seq and DNase-seq data, can be integrated to improve the annotation of *cis*-elements and predict gene function. Furthermore, more accurate data, such as gene families, new type of non-coding RNAs, KEGG pathway and GO terms, needs to be integrated, too. Our croFGD database will be updated regularly, and we hope the database can help the community study the functional genomics and yield novel insights into the molecular regulatory mechanisms.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/sra.

AUTHOR CONTRIBUTIONS

JS performed gene functional annotation, functional module identification, and database construction. HY performed data collection and the co-expression network construction. JY gave advice a lot about the web server. WX gave advice for the application of the co-expression network and some key functional module identification. ZS and WX supervised the project. All authors read and approved the final manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (Grant Nos. 31771467 and 31571360).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00238/full#supplementary-material

REFERENCES

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023. doi: 10.1093/bioinformatics/btl039

Almagro, L., Fernández-Pérez, F., and Pedreño, M. A. (2015). Indole alkaloids from *Catharanthus roseus*: bioproduction and their effect on human health. *Molecules* 20, 2973–3000. doi: 10.3390/molecules20022973

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. (2016). ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiol.* 57:e5. doi: 10.1093/pcp/pcv165

Aparicio-Fabre, R., Guillén, G., Loredo, M., Arellano, J., Valdés-López, O., Ramírez, M., et al. (2013). Common bean (*Phaseolus vulgaris* L.) PvTIFY orchestrates global changes in transcript profile response to jasmonate and phosphorus deficiency. *BMC Plant Biol.* 13:26. doi: 10.1186/1471-2229-13-26

Aslam, J., Khan, S. H., Siddiqui, Z. H., Fatima, Z., Maqsood, M., Bhat, M. A., et al. (2010). *Catharanthus roseus* (L.) G. Don. an important drug: it's applications and production. *Int. J. Compr. Pharm.* [Epub ahead of print].

Brown, J., Pirrung, M., and Mccue, L. A. (2017). FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* doi: 10.1093/bioinformatics/btx373 [Epub ahead of print].

Caputi, L., Franke, J., Farrow, S. C., Chung, K., Payne, R. M. E., Nguyen, T. D., et al. (2018). Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* 360, 1235–1239. doi: 10.1126/science.aat4100

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45, D369–D379. doi: 10.1093/nar/gkw1102

Chebbi, M., Ginis, O., Courdavault, V., Glévarec, G., Lanoue, A., Clastre, M., et al. (2014). ZCT1 and ZCT2 transcription factors repress the activity of a gene promoter from the methyl erythritol phosphate pathway in Madagascar periwinkle cells. *J. Plant Physiol.* 171, 1510–1513. doi: 10.1016/j.jplph.2014.07.004

Chen, J., Ma, M., Shen, N., Xi, J. J., and Tian, W. (2013). Integration of cancer gene co-expression network and metabolic network to uncover potential cancer drug targets. *J. Proteome Res.* 12, 2354–2364. doi: 10.1021/pr400162t

Conesa, A., and Gotz, S. (2008). Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008:619832. doi: 10.1155/2008/619832

Coneva, V., Simopoulos, C., Casaretto, J. A., El-kereamy, A., Guevara, D. R., Cohn, J., et al. (2014). Metabolic and co-expression network-based analyses associated with nitrate response in rice. *BMC Genomics* 15:1056. doi: 10.1186/1471-2164-15-1056

Dai, X., and Zhao, P. X. (2011). PsRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* 39, W155–W159. doi: 10.1093/nar/gkr319

Dugé de Bernonville, T., Foureau, E., Parage, C., Lanoue, A., Clastre, M., Londono, M. A., et al. (2015). Characterization of a second secologanin synthase isoform producing both secologanin and secoxyloganin allows enhanced de novo assembly of a *Catharanthus roseus* transcriptome. *BMC Genomics* 16:619. doi: 10.1186/s12864-015-1678-y

El-Sharkawy, I., Liang, D., and Xu, K. (2015). Transcriptome analysis of an apple (Malus × domestica) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. *J. Exp. Bot.* 66, 7359–7376. doi: 10.1093/jxb/erv433

Endo, M., Shimizu, H., Nohales, M. A., Araki, T., and Kay, S. A. (2014). Tissue-specific clocks in Arabidopsis show asymmetric coupling. *Nature* 515, 419–422. doi: 10.1038/nature13919

Farris, S. P., Harris, R. A., and Ponomarev, I. (2015). Epigenetic modulation of brain gene networks for cocaine and alcohol abuse. *Front. Neurosci.* 9:176. doi: 10.3389/fnins.2015.00176

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D30. doi: 10.1093/nar/gkt1223

Gao, T., Liu, Z., Wang, Y., Cheng, H., Yang, Q., Guo, A., et al. (2013). UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. *Nucleic Acids Res.* 4, D445–D451. doi: 10.1093/nar/gks1103

Geu-Flores, F., Sherden, N. H., Glenn, W. S., O'connor, S. E., Courdavault, V., Burlat, V., et al. (2012). An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. *Nature* 492, 138–142. doi: 10.1038/nature11692

Heitz, T., Widemann, E., Lugan, R., Miesch, L., Ullmann, P., Désaubry, L., et al. (2012). Cytochromes P450 CYP94C1 and CYP94B3 catalyze two successive oxidation steps of plant hormone jasmonoyl-isoleucine for catabolic turnover. *J. Biol. Chem.* 287, 6296–6306. doi: 10.1074/jbc.M111.316364

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297

Hu, H., He, X., Tu, L., Zhu, L., Zhu, S., Ge, Z., et al. (2016). GhJAZ2 negatively regulates cotton fiber initiation by interacting with the R2R3-MYB transcription factor GhMYB25-like. *Plant J.* 88, 921–935. doi: 10.1111/tpj.13273

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Kellner, F., Kim, J., Clavijo, B. J., Hamilton, J. P., Childs, K. L., Vaillancourt, B., et al. (2015). Genome-guided investigation of plant natural product biosynthesis. *Plant J.* 82, 680–692. doi: 10.1111/tpj.12827

Koo, A. J., Thireault, C., Zemelis, S., Poudel, A. N., Zhang, T., Kitaoka, N., et al. (2014). Endoplasmic reticulum-associated inactivation of the hormone jasmonoyl-L-isoleucine by multiple members of the cytochrome P450 94 family in Arabidopsis. *J. Biol. Chem.* 289, 29728–29738. doi: 10.1074/jbc.M114.603084

Kumar, K., Kumar, S. R., Dwivedi, V., Rai, A., Shukla, A. K., Shanker, K., et al. (2015). Precursor feeding studies and molecular characterization of geraniol synthase establish the limiting role of geraniol in monoterpene indole alkaloid biosynthesis in *Catharanthus roseus* leaves. *Plant Sci.* 239, 56–66. doi: 10.1016/j.plantsci.2015.07.007

Kurotani, K. I., Hayashi, K., Hatanaka, S., Toda, Y., Ogawa, D., Ichikawa, H., et al. (2015). Elevated levels of CYP94 family gene expression alleviate the Jasmonate response and enhance salt tolerance in rice. *Plant Cell Physiol.* 779–789. doi: 10.1093/pcp/pcv006

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., et al. (2015). AraNet v2: an improved database of co-functional gene networks for the study of *Arabidopsis thaliana* and 27 other nonmodel plant species. *Nucleic Acids Res.* 43, D996–D1002. doi: 10.1093/nar/gku1053

Li, C. Y., Leopold, A. L., Sander, G. W., Shanks, J. V., Zhao, L., and Gibson, S. I. (2013). The ORCA2 transcription factor plays a key role in regulation of the terpenoid indole alkaloid pathway. *BMC Plant Biol.* 13:155. doi: 10.1186/1471-2229-13-155

Li, J., Wang, X., and Cui, Y. (2014). Uncovering the overlapping community structure of complex networks by maximal cliques. *Phys. A Stat. Mech. Appl.* 415, 398–406. doi: 10.1016/j.physa.2014.08.025

Liu, J., Cai, J., Wang, R., and Yang, S. (2017). Transcriptional regulation and transport of terpenoid indole alkaloid in *Catharanthus roseus*: exploration of new research directions. *Int. J. Mol. Sci.* 18:E53. doi: 10.3390/ijms18010053

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, 490–495. doi: 10.1093/nar/gkt1178

Lumba, S., Toh, S., Handfield, L. F., Swan, M., Liu, R., Youn, J. Y., et al. (2014). A mesoscale abscisic acid hormone interactome reveals a dynamic signaling landscape in arabidopsis. *Dev. Cell* 29, 360–372. doi: 10.1016/j.devcel.2014.04.004

Ma, C., Xin, M., Feldmann, K. A., and Wang, X. (2014). Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in Arabidopsis. *Plant Cell* 26, 520–537. doi: 10.1105/tpc.113.121913

Ma, X., Zhao, H., Xu, W., You, Q., Yan, H., Gao, Z., et al. (2018). Co-expression gene network analysis and functional module identification in bamboo growth and development. *Front. Genet.* 9:574. doi: 10.3389/fgene.2018.00574

MacGregor, D. R., and Penfield, S. (2015). Exploring the pleiotropy of hos1. *J. Exp. Bot.* 66, 1661–1671. doi: 10.1093/jxb/erv022

Memelink, J., and Gantet, P. (2007). Transcription factors involved in terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Phytochem. Rev.* 6, 353–362. doi: 10.1007/s11101-006-9051-z

Menke, F., Parchmann, S., Mueller, M., Kijne, J., and Memelink, J. (1999). Involvement of the octadecanoid pathway and protein phosphorylation in fungal elicitor-induced expression of terpenoid indole alkaloid biosynthetic genes in *Catharanthus roseus*. *Plant Physiol.* 119, 1289–1296. doi: 10.1104/pp.119.4.1289

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., et al. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910. doi: 10.1105/tpc.111.083667

Nelson, D. R. (2009). The cytochrome p450 homepage. *Hum. Genomics* 4, 59–65. doi: 10.1186/1479-7364-4-1-59

Newton, R., and Wernisch, L. (2014). A meta-analysis of multiple matched copy number and transcriptomics data sets for inferring gene regulatory relationships. *PLoS One* 9:e105522. doi: 10.1371/journal.pone.0105522

Niu, Y., Figueroa, P., and Browse, J. (2011). Characterization of JAZ-interacting bHLH transcription factors that regulate jasmonate responses in Arabidopsis. *J. Exp. Bot.* 62, 2143–2154. doi: 10.1093/jxb/erq408

Noordewier, M. O., and Warren, P. V. (2001). Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol.* 19, 412–415. doi: 10.1016/S0167-7799(01)01735-8

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* 59:e3. doi: 10.1093/pcp/pcx191

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42, D358–D363. doi: 10.1093/nar/gkt1115

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818. doi: 10.1038/nature03607

Pan, Q., Wang, Q., Yuan, F., Xing, S., Zhao, J., Choi, Y. H., et al. (2012). Overexpression of ORCA3 and G10H in *catharanthus roseus* plants regulated alkaloid biosynthesis and metabolism revealed by NMR-metabolomics. *PLoS One* 7:e43038. doi: 10.1371/journal.pone.0043038

Pan, Y. J., Lin, Y. C., Yu, B. F., Zu, Y. G., Yu, F., and Tang, Z. H. (2018). Transcriptomics comparison reveals the diversity of ethylene and methyl-jasmonate in roles of TIA metabolism in *Catharanthus roseus*. *BMC Genomics* 19:508. doi: 10.1186/s12864-018-4879-3

Pandey, S. S., Singh, S., Babu, C. S. V., Shanker, K., Srivastava, N. K., Shukla, A. K., et al. (2016). Fungal endophytes of *Catharanthus roseus* enhance vindoline content by modulating structural and regulatory genes related to terpenoid indole alkaloid biosynthesis. *Sci. Rep.* 6:26583. doi: 10.1038/srep26583

Parage, C., Foureau, E., Kellner, F., Burlat, V., Mahroug, S., Lanoue, A., et al. (2016). Class II Cytochrome P450 Reductase Governs the Biosynthesis of Alkaloids. *Plant Physiol.* 172, 1563–1577. doi: 10.1104/pp.16.00801

Pathania, S., Bagler, G., and Ahuja, P. S. (2016). Differential network analysis reveals evolutionary complexity in secondary metabolism of *Rauvolfia serpentina* over *Catharanthus roseus*. *Front. Plant Sci.* 7:1229. doi: 10.3389/fpls.2016.01229

Patra, B., Pattanaik, S., Schluttenhofer, C., and Yuan, L. (2018). A network of jasmonate-responsive bHLH factors modulate monoterpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *New Phytol.* 217, 1566–1581. doi: 10.1111/nph.14910

Pérez-Rodríguez, P., Riaño-Pachón, D. M., Corrˆea, L. G. G., Rensing, S. A., Kersten, B., and Mueller-Roeber, B. (2009). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res* 38, D822–D827. doi: 10.1093/nar/gkp805

Rhee, S. Y., and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends Plant Sci.* 19, 212–221. doi: 10.1016/j.tplants.2013.10.006

Rischer, H., Oresic, M., Seppanen-Laakso, T., Katajamaa, M., Lammertyn, F., Ardiles-Diaz, W., et al. (2006). Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5614–5619. doi: 10.1073/pnas.0601027103

Rizvi, N. F., Weaver, J. D., Cram, E. J., and Lee-Parsons, C. W. T. (2016). Silencing the transcriptional repressor, ZCT1, illustrates the tight regulation of terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* hairy roots. *PLoS One* 11:e0159712. doi: 10.1371/journal.pone.0159712

Rombauts, S., Déhais, P., Van Montagu, M., and Rouzé, P. (1999). PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res.* 27, 295–296. doi: 10.1093/nar/27.1.295

Shen, E. M., Singh, S. K., Ghosh, J. S., Patra, B., Paul, P., Yuan, L., et al. (2017). The miRNAome of *Catharanthus roseus*: identification, expression analysis, and potential roles of microRNAs in regulation of terpenoid indole alkaloid biosynthesis. *Sci. Rep.* 7:43027. doi: 10.1038/srep43027

Shim, J. S., Jung, C., Lee, S., Min, K., Lee, Y. W., Choi, Y., et al. (2013). AtMYB44 regulates WRKY70 expression and modulates antagonistic interaction between salicylic acid and jasmonic acid signaling. *Plant J.* 73, 483–495. doi: 10.1111/tpj.12051

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Sonnhammer, E. L. L., and Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* 43, D234–D239. doi: 10.1093/nar/gku1203

Speir, M. L., Zweig, A. S., Rosenbloom, K. R., Raney, B. J., Paten, B., Nejad, P., et al. (2016). The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* 44, D717–D725. doi: 10.1093/nar/gkv1275

Steffens, N. O. (2004). AthaMap: an online resource for in silico transcription factor binding sites in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.* 32, 368D–372D. doi: 10.1093/nar/gkh017

Suttipanta, N., Pattanaik, S., Kulshrestha, M., Patra, B., Singh, S. K., and Yuan, L. (2011). The Transcription Factor CrWRKY1 positively regulates the terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Physiol.* 157, 2081–2093. doi: 10.1104/pp.111.181834

Tani, T., Sobajima, H., Okada, K., Chujo, T., Arimura, S. I., Tsutsumi, N., et al. (2008). Identification of the OsOPR7 gene encoding 12-oxophytodienoate reductase involved in the biosynthesis of jasmonic acid in rice. *Planta* 227:517. doi: 10.1007/s00425-007-0635-7

Tchieu, J. H., Fana, F., Fink, J. L., Harper, J., Nair, T. M., Niedner, R. H., et al. (2003). The PlantsP and PlantsT functional genomics databases. *Nucleic Acids Res.* 31, 342–344. doi: 10.1093/nar/gkg025

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). AgriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122-W129. doi: 10.1093/nar/gkx382

Tian, T., You, Q., Yan, H., Xu, W., and Su, Z. (2018). MCENet: a database for maize conditional co-expression network and network characterization collaborated with multi-dimensional omics levels. *J. Genet. Genomics* 45, 351–360. doi: 10.1016/j.jgg.2018.05.007

Tian, T., You, Q., Zhang, L., Yi, X., Yan, H., Xu, W., et al. (2016). SorghumFDB: sorghum functional genomics database with multidimensional network analysis. *Database* 2016:baw099. doi: 10.1093/database/baw099

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 31, 46–53. doi: 10.1038/nbt.2450

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

Van Der Fits, L., and Memelink, J. (2000). ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science* 289, 295–297. doi: 10.1126/science.289.5477.295

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G. J. E., Rombauts, S., Hasnain, G., et al. (2013). CathaCyc, a metabolic pathway database built from *catharanthus roseus* RNA-seq data. *Plant Cell Physiol.* 54, 673–685. doi: 10.1093/pcp/pct039

Van Moerkercke, A., Steensma, P., Schweizer, F., Pollier, J., Gariboldi, I., Payne, R., et al. (2015). The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8130–8135. doi: 10.1073/pnas.1504951112

Verma, M., Ghangal, R., Sharma, R., Sinha, A. K., and Jain, M. (2014). Transcriptome analysis of *Catharanthus roseus* for gene discovery and expression profiling. *PLoS One* 9:e103583. doi: 10.1371/journal.pone.0103583

Wallström, S. V., Florez-Sarasa, I., Araújo, W. L., Aidemark, M., Fernández-Fernández, M., Fernie, A. R., et al. (2014). Suppression of the external mitochondrial NADPH dehydrogenase, NDB1, in *Arabidopsis thaliana* affects central metabolism and vegetative growth. *Mol. Plant* 7, 356–368. doi: 10.1093/mp/sst115

Wang, K., Guo, Q., Froehlich, J. E., Hersh, H. L., Zienkiewicz, A., Howe, G. A., et al. (2018). Two abscisic acid responsive plastid lipase genes involved in jasmonic acid biosynthesis in *Arabidopsis thaliana*. *Plant Cell* 30, 1006–1022. doi: 10.1105/tpc.18.00250

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., and Rokas, A. (2017). A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* 29, 944-959doi: 10.1105/tpc.17.00009

Wu, T. D., and Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi: 10.1093/bioinformatics/bti310

Yamamoto, K., Takahashi, K., Mizuno, H., Anegawa, A., Ishizaki, K., Fukaki, H., et al. (2016). Cell-specific localization of alkaloids in *Catharanthus roseus* stem tissue measured with Imaging MS and Single-cell MS. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3891–3896. doi: 10.1073/pnas.1521959113

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–W103. doi: 10.1093/nar/gkt281

You, Q., Xu, W., Zhang, K., Zhang, L., Yi, X., Yao, D., et al. (2017). CcNET: Database of co-expression networks with functional modules for diploid and polyploid Gossypium. *Nucleic Acids Res.* 45, D1090–D1099. doi: 10.1093/nar/gkw910

You, Q., Zhang, L., Yi, X., Zhang, K., Yao, D., Zhang, X., et al. (2016). Co-expression network analyses identify functional modules associated with development and stress response in *Gossypium arboreum*. *Sci. Rep.* 6:38436. doi: 10.1038/srep38436

You, Q., Zhang, L., Yi, X., Zhang, Z., Xu, W., and Su, Z. (2015). SIFGD: *Setaria italica* functional genomics database. *Mol. Plant* 8, 967–970. doi: 10.1016/j.molp.2015.02.001

Yu, J., Zhang, Z., Wei, J., Ling, Y., Xu, W., and Su, Z. (2014). SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways. *BMC Genomics* 15:271. doi: 10.1186/1471-2164-15-271

Zhang, H., Hedhili, S., Montiel, G., Zhang, Y., Chatel, G., Pré, M., et al. (2011). The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis

in *Catharanthus roseus*. *Plant J.* 67, 61–71. doi: 10.1111/j.1365-313X.2011.04575.x

Zhang, L., Guo, J., You, Q., Yi, X., Ling, Y., Xu, W., et al. (2015). GraP: Platform for functional genomics analysis of *Gossypium raimondii*. *Database* 2015:bav047. doi: 10.1093/database/bav047

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhou, J., Xu, Y., Lin, S., Guo, Y., Deng, W., Zhang, Y., et al. (2018). iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations. *Nucleic Acids Res.* 46, D447–D453. doi: 10.1093/nar/gkx1041

Zhu, D., Bai, X., Chen, C., Chen, Q., Cai, H., Li, Y., et al. (2011). GsTIFY10, a novel positive regulator of plant tolerance to bicarbonate stress and a repressor of jasmonate signaling. *Plant Mol. Biol.* 77, 285–297. doi: 10.1007/s11103-011-9810-0

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: 03 April 2019
doi: 10.3389/fgene.2019.00239

Rare Variants Imputation in Admixed Populations: Comparison Across Reference Panels and Bioinformatics Tools

*Sanjeev Sariya[1,2], Joseph H. Lee[1,2,3], Richard Mayeux[1,2,3], Badri N. Vardarajan[1,2], Dolly Reyes-Dumeyer[1,2], Jennifer J. Manly[1,2,3], Adam M. Brickman[1,2,3], Rafael Lantigua[4], Martin Medrano[5], Ivonne Z. Jimenez-Velazquez[6] and Giuseppe Tosto[1,2,3]**

[1] Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, United States, [2] The Gertrude H. Sergievsky Center, College of Physicians and Surgeons, Columbia University, New York, NY, United States, [3] Department of Neurology, College of Physicians and Surgeons, New York-Presbyterian Hospital, Columbia University Medical Center, New York, NY, United States, [4] Medicine College of Physicians and Surgeons, and The Department of Epidemiology, School of Public Health, Columbia University, New York, NY, United States, [5] School of Medicine, Pontificia Universidad Catolica Madre y Maestra, Santiago, Dominican Republic, [6] Department of Medicine, Geriatrics Program, University of Puerto Rico School of Medicine, San Juan, Puerto Rico

OPEN ACCESS

Edited by:
Vinicius Maracaja-Coutinho, Universidad de Chile, Chile

Reviewed by:
Peng Zhang, Johns Hopkins University, United States
Daniela Albrecht-Eckardt, BioControl Jena GmbH, Germany

***Correspondence:**
Giuseppe Tosto
gt2260@cumc.columbia.edu

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

Received: *06 November 2018*
Accepted: *04 March 2019*
Published: *03 April 2019*

Citation:
Sariya S, Lee JH, Mayeux R, Vardarajan BN, Reyes-Dumeyer D, Manly JJ, Brickman AM, Lantigua R, Medrano M, Jimenez-Velazquez IZ and Tosto G (2019) Rare Variants Imputation in Admixed Populations: Comparison Across Reference Panels and Bioinformatics Tools. Front. Genet. 10:239. doi: 10.3389/fgene.2019.00239

Background: Imputation has become a standard approach in genome-wide association studies (GWAS) to infer *in silico* untyped markers. Although feasibility for common variants imputation is well established, we aimed to assess rare and ultra-rare variants' imputation in an admixed Caribbean Hispanic population (CH).

Methods: We evaluated imputation accuracy in CH (N = 1,000), focusing on rare (0.1% $\leq$ minor allele frequency (MAF) $\leq$ 1%) and ultra-rare (MAF $<$ 0.1%) variants. We used two reference panels, the Haplotype Reference Consortium (HRC; N = 27,165) and 1000 Genome Project (1000G phase 3; N = 2,504) and multiple phasing (SHAPEIT, Eagle2) and imputation algorithms (IMPUTE2, MACH-Admix). To assess imputation quality, we reported: (a) high-quality variant counts according to imputation tools' internal indexes (e.g., IMPUTE2 "Info" $\geq$ 80%). (b) Wilcoxon Signed-Rank Test comparing imputation quality for genotyped variants that were masked and imputed; (c) Cohen's kappa coefficient to test agreement between imputed and whole-exome sequencing (WES) variants; (d) imputation of G206A mutation in the *PSEN1* (ultra-rare in the general population an more frequent in CH) followed by confirmation genotyping. We also tested ancestry proportion (European, African and Native American) against WES-imputation mismatches in a Poisson regression fashion.

Results: SHAPEIT2 retrieved higher percentage of imputed high-quality variants than Eagle2 (rare: 51.02% vs. 48.60%; ultra-rare 0.66% vs. 0.65%, Wilcoxon p-value $<$ 0.001). SHAPEIT-IMPUTE2 employing HRC outperformed 1000G (64.50% vs. 59.17%; 1.69% vs. 0.75% for high-quality rare and ultra-rare variants, respectively, Wilcoxon p-value $<$ 0.001). SHAPEIT-IMPUTE2 outperformed MaCH-Admix. Compared to 1000G, HRC-imputation retrieved a higher number of high-quality rare and ultra-rare variants, despite showing lower agreement between imputed and WES variants

(e.g., rare: 98.86% for HRC vs. 99.02% for 1000G). High Kappa (K = 0.99) was observed for both reference panels. Twelve G206A mutation carriers were imputed and all validated by confirmation genotyping. African ancestry was associated with higher imputation errors for uncommon and rare variants (p-value < 1e-05).

Conclusion: Reference panels with larger numbers of haplotypes can improve imputation quality for rare and ultra-rare variants in admixed populations such as CH. Ethnic composition is an important predictor of imputation accuracy, with higher African ancestry associated with poorer imputation accuracy.

Keywords: rare variants, imputation, admixed population, GWAS, 1000G

INTRODUCTION

Genome-wide association studies (GWASs) are a major tool to identify common variants associated with complex diseases. GWAS can include 550 K to over 2 M Single Nucleotide Polymorphisms (SNPs) (Ha et al., 2014) to cover the human genome evenly. Although GWAS has shown to be a robust method to identify disease loci of interest, they rarely point to a causal coding variant. In fact, microarray SNP chips for GWAS are optimally designed to uncover common variants, often associated with small effect sizes mostly located in intronic and intergenic regions. The focus of genetic investigations has since shifted toward rarer alleles with larger effect sizes (Gibson, 2012). With the changing paradigm, imputation of rare variants has become an important topic to enhance the genome coverage in GWAS. Imputation is a process of inferring untyped SNP markers in the discovery population by using densely typed SNPs in external reference panel(s). These '*in silico*' markers increase the coverage of association tests while conducting genome-wide association analysis. In addition, large number of SNPs facilitate meta-analysis when merging data from different study cohorts.

The quality of imputation essentially depends on two parameters: available reference datasets and algorithms that employ those reference datasets. Previous studies have shown that imputation quality depends on how well reference panels reflect the study population. To respond to the needs, the 1000 Genome project (1000G), now in its third phase release, has proven to be one of the most frequently used reference panels (Genomes Project et al., 2015). Using these composite reference panels, a number of studies (Pei et al., 2010; Howie et al., 2012; Verma et al., 2014; Liu et al., 2015) have compared imputation accuracy using different imputation tools and algorithms, although the results are equivocal. Few studies (Browning and Browning, 2009; Zheng et al., 2012, 2015) assessed the impact of reference panel size and input data's features - such as density of SNPs - to impute rare variants, suggesting larger size of reference panels work better. Surakka et al. (2016) assessed accuracy of imputed SNPs by evaluating rate of false polymorphisms in a Finnish population using global reference panels – Haplotype Reference Consortium (HRC) release 1, 1000G phase 1 and a local reference panel. They concluded that higher false positive rate was observed in imputation from global reference panels compared to imputation performed using a local panel. Other studies (Huang et al., 2015; Das et al., 2016) found imputation accuracy increases with higher number of haplotypes, specifically for variants with MAF $\leq$ 0.5%. For Hispanic populations, Nelson et al. (2016) compared imputation performances with 1000G phase 1 (N = 1,092) vs. 1000G phase 3 (N = 2,504), concluding that phase 3 improved accuracy for variants with MAF < 1% by. Further, Nagy et al. (2017) showed that HRC reference panel provides new insight for novel variants particularly for rare variants in a family-based Scottish study cohort. Aforementioned studies highlighted the need of a larger sized reference panel to improve imputation quality. Herzig et al. (2018) assessed tools for haplotype phasing and their impact on imputation in a population isolate of Campora in southern Italy, and showed that SHAPEIT2, SHAPEIT3 and EAGLE2 were highly accurate in phasing; MINIMAC3, IMPUTE4 and IMPUTE2 were found to be reliable for imputation. Roshyara et al. (2014) compared MaCH-Admix, IMPUTE2, MACH, MACH-Minimac in different ethnicities by evaluating accuracy of correctly imputed SNPs; MaCH-Minimac outperformed SHAPEIT-IMPUTE2 in subsamples of different ethnic groups. These studies demonstrated how employed imputation algorithm determines quality of inferred SNPs.

However, no study to our knowledge has evaluated reference panels in tandem with different imputation algorithms to assess imputation quality of inferred SNPs based on MAF in a three-way admixed population. Based on these findings, we assessed imputation quality, focusing on rare and ultra-rare variants, in a large dataset of Caribbean Hispanics (CH) leveraging available GWAS and sequencing data available for our cohort.

MATERIALS AND METHODS

We will refer SNPs with MAF between 1 and 5% as "uncommon," 0.1–1% as "rare," and $\leq$ 0.1% as "ultra-rare." We considered SNPs with IMPUTE-Info metric $\geq$ 0.40 as "good-quality" and $\geq$ 0.80 as "high-quality."

GWAS Samples and Genotyping

We selected randomly 1,000 Caribbean Hispanics as part of an original genotyped cohort of 3,138 individuals: genotyped data can be downloaded at dbGaP Study Accession: phs000496.v1.p1. 719 individuals were derived from Estudio Familiar Investigar Genetica de Alzheimer (EFIGA), a study of familial LOAD; and 281 individuals from the multiethnic longitudinal cohort,

Washington Heights, Inwood, Columbia Aging Project (WHICAP). The information on study design, recruitment and GWAS methods for the EFIGA and WHICAP study was previously described in Tosto et al. (2015).

GWAS Quality Control (QC)

Genotyped data underwent quality control using PLINK (v1.90b4.9 64-bit) (Purcell et al., 2007). Briefly, we excluded SNPs with missing rate $\geq$ 5% followed by exclusion of SNPs with MAF $\leq$ 1%. We then removed SNPs with *P*-value < 1e-6 for Hardy-Weinberg Equilibrium. Samples with missing call rate $\geq$ 5% were excluded from analysis.

Global Ancestry Estimation and Selection of "True Hispanics"

Prior to imputation, we estimated global ancestry using the ADMIXTURE (v.1.3.0) software (Alexander et al., 2009; Zhou et al., 2011). We conducted supervised admixture analyses using three reference populations: African Yoruba (YRI) and non-Hispanic white of European Ancestry (CEU) from the HAPMAP project as representative of African and European ancestral populations; and eight Surui, 21 Maya, 14 Karitiana, 14 Pima and seven Colombian individuals from the Human Genome Diversity Project (HGDP) were used to represent native American ancestry (Li et al., 2008). We used ∼80,000 autosomal SNPs that were: (I) genotyped in all three datasets (Caribbean Hispanics, 1000G and HGDP); (II) common (i.e., MAF > 5 %); and III) in linkage equilibrium. Supervised admixture analyses with the three reference populations (YRI, CEU, and Native Americans) revealed that European lineage accounted for most of the ancestral origins (59%), followed by African (33%) and native American ancestry (8%). We then selected only individuals with at least 1% of all three ancestral populations.

Reference Panels

HRC reference panel contained over 39M SNPs from 27,165 individuals who participated in 17 different studies (**Table 1**). The data were downloaded from the Wellcome Trust Sanger Institute (WTSI).

1000G phase 3 reference panel contained over 81M SNPs from 2,504 individuals[1]. It includes 26 ethnic groups, with most variants rare, approximately 64 million had MAF < 0.5%; approximately 12 million had a MAF between 0.5 and 5%; and approximately eight million have MAF > 5%. In order to perform imputation with MaCH-Admix, 1000G Phase 3 pre-formatted data were downloaded from ftp://yunlianon:anon@rc-ns-ftp.its.unc.edu/ALL.phase3_v5.shapeit2_mvncall_integrated.noSingleton.tgz that contained over 47M SNPs.

The subsequent analyses were restricted to autosomal chromosomes, only.

[1]https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.tgz

TABLE 1 | SNP counts in HRC and 1000G reference panel.

Reference Panel	Individuals	Autosomal variants	Bi-allelic SNPs	Multi-allelic SNPs
1000G Phase 3	2,504	81,706,022	77,818,332	3,887,690
HRC	27,165*	39,131,600	39,131,600	NA

**For Chromosome 1, the number of individuals were 22,691.*

Phasing and Imputation Procedures

We compared SHAPEIT2 (Delaneau et al., 2013) and Eagle2 (Loh et al., 2016) by phasing and then imputing (see next section) a single chromosome (Chromosome 21), using both reference panels. We refer to SHAPEIT2 as SHAPEIT when used in tandem with IMPUTE2 for the remainder of paper.

Imputation was carried out using two bioinformatics tools: IMPUTE2 (Howie et al., 2009) and MaCH-Admix (Liu et al., 2013). For both, imputation quality ranged from 0 to 1, with 0 indicating complete uncertainty in imputed genotypes, and 1 indicating no uncertainty in imputed genotypes.

IMPUTE2 (Version 2.3.2)

IMPUTE2 uses an MCMC algorithm to integrate over the space of possible phase reconstructions for genotypes data. We conducted imputation in non-overlapping 1MB chunk regions; chunk coordinates were specified using the "*–int*" option. Other options were used with default parameters (**Supplementary Section S1**). Briefly, we used a default 250KB buffer region to avoid quality deterioration on the ends of chunk region. "-Ne" value as 2000 suggested for robust imputation which scales linkage disequilibrium and recombination error rate.

MaCH-Admix

We used MaCH-Admix because it uses a method based on IBS matching in a piecewise manner. The method breaks genomic region under investigation into small pieces and finds reference haplotypes that best represent every small piece, for each target individual separately. MaCH-Admix imputes in three steps: phasing, estimation of model parameter that includes error rare and recombination rate and lastly, haplotype-based imputation. MaCH-Admix (version Beta 2.0.185) was run on default parameters of 30 rounds, 100 states (–autoFlip flag). Details can be found in **Supplementary Section S1**. We initially compared performance between MaCH-Admix and IMPUTE2 using the 1000G reference panel for Chromosome 21 only. We then proceeded to impute all remaining chromosomes with the tool that performed better.

Imputation Performance Metrics

IMPUTE2 uses "Info" parameter to report imputation quality that measures relative statistical information about SNP allele frequency from imputed data. It reflects the information in imputed genotypes relative to the information if only the allele frequency were known. "Info" metric is used to filter poorly imputed SNPs from IMPUTE2 and is reported for all imputed SNPs. In addition, IMPUTE2 uses an internal metric known as R^2, reported for genotyped SNPs only: it measures squared correlation between genotyped SNPs and the same SNPs that

have been first masked internally and then imputed. MaCH-Admix uses *Rsq* to report imputation quality. The R^2 metric is also known as variance ratio, calculated as proportion of empirically observed variance (based on the imputation) to the expected binomial variance p(1-p), where p is the minor allele frequency. A threshold of 0.30 is recommended to filter out poorly imputed SNPs.

Despite quality measures from IMPUTE2 and MaCH-Admix being highly correlated (Marchini and Howie, 2010), we calculated a *r2hat* score to generate a single common metric to assess imputation quality across the software (Hancock et al., 2012) (v109)[2].

We compared performance of MaCH-Admix and SHAPEIT-IMPUTE2 by: (a) Reporting raw SNP counts based on quality (MaCH-Admix "Rsq" and IMPUTE2 "Info"); (b) Comparing *r2hat* for overlapping imputed SNPs from both tools; (c) Conducting a Wilcoxon Signed-Rank Test (R v3.4.2) on *r2hat* value of overlapping SNPs.

We compared performance of Eagle2 and SHAPEIT2 phasing tools in tandem with IMPUTE2 as imputation tools across reference panels by: (a) Comparing their respective IMPUTE2 R^2: (b) Conducting a Wilcoxon Signed-Rank Test on R^2 value; (c) Reporting raw counts of imputed SNPs based on IMPUTE2 "Info" metric and stratified by MAF bins (e.g., common, rare, ultra-rare).

In all comparisons, the MAFs are estimated from imputed data according to the reference panel employed. We retained monomorphic SNPs in our analyses for several reasons. A monomorphic SNP in one study might not be monomorphic in other cohorts. This has profound affects, for example, when performing meta-analysis across different studies. In addition, monomorphic SNPs provide information about MAF across studies. Without the information it is difficult to tell, for instance, if a SNP is monomorphic or failed quality control in that study.

Agreement Between Imputed and Sequence Data

To further test the quality of imputation -without relying on software's internal metrics (i.e., "Info" and R^2) - we calculated genotyped concordance between imputed and WES data using the VCF-compare tool (v0.1.14-12-gcdb80b8) (Danecek et al., 2011). First, we converted posterior probabilities obtained from imputation into genotype data using the PLINK software (v1.90b4.9) by applying a threshold of 0.9 (**Supplementary Section S1**), such that SNPs that failed on this criterion were left uncalled. For example, an imputed SNP with P(G = 0,1,2) = (0.01,0.9,0.09) would be called as a '1' (heterozygous), whereas an imputed SNP with P(G = 0,1,2) = (0.2, 0.6, 0.2) would be left uncalled. We restricted the comparison to overlapping SNPs between HRC, 1000G reference panels and whole-exome sequencing (WES) data for Chromosome 14 only, on SNPs with 0% missingness (plink –missing flag) in WES data. We also assessed variants' agreement according to different MAF bins for "high-quality" ("Info" $\geq$ 0.8) SNPs. The output resulted in number of variant "mismatches," i.e., the count of allele not matching between imputed and sequenced variants per individual. Work-flow for VCF-compare can be found in **Supplementary Figure S1**. To measure interrater reliability we computed Cohen's kappa coefficient (McHugh, 2012) for both the reference panels against WES data. Kappa coefficient $\leq$ 0 indicates no agreement, 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Work-flow for Cohen's kappa coefficient calculation can be found in **Supplementary Figure S2**.

Effects of Ancestry on Imputation Quality

To assess how ancestry affected imputation quality, we conducted a Poisson regression using R. We used percentage of global ancestry (European (CEU), Native (NAT) and African (YRI) as predictors, and total number of mismatches as the outcome; analyses were restricted to "high-quality" SNPs, only.

Imputation of G206A Mutation in PSEN1

To evaluate imputation performance of a specific rare variant, we examined a founder mutation, p.Gly206Ala (G206A - rs63750082) in the *PSEN1* gene (PSEN1-G206A) (Athan et al., 2001; Lee et al., 2015). The PSEN1-G206A mutation is a rare variant observed primarily in Puerto Ricans with familial early onset Alzheimer's disease (EOAD), but it is rare in Puerto Ricans and other populations with late-onset Alzheimer's disease (LOAD) (Arnold et al., 2013). The mutation was present in the 1000G phase 3 reference panel with an allele frequency of 0.001, but was absent in the HRC reference panel. To verify whether individuals who were found to carry the PSEN1-G206A mutation based on 1000G-imputation, they were genotyped using the KASP genotyping technology by LGC genomics[3], which uses allele-specific PCR for SNP calling. Agreement between imputed and genotype data for the PSEN1-G206A mutation was then assessed. We also tested the effect on imputation quality based on different IMPUTE2-parameters settings, more specifically by modifying the chunk size (i.e., 1 MB vs. 5 MB).

RESULTS

Comparison of Phasing Tools: Eagle2 vs. SHAPEIT2

To select the optimal tool for phasing, we compared SHAPEIT2 with Eagle2 using Chromosome 21 with 13,066 genotyped SNPs by performing subsequent imputation with IMPUTE2 on phased outputs, and using both reference panels. We found SHAPEIT2 better than Eagle2 when evaluated based on mean R^2 and "Info" metric using either the reference panels. For instance, using the 1000G, we observed higher mean R^2 for data phased with SHAPEIT2 as compared to Eagle2 (0.92 vs. 0.91; Wilcoxon p-value < 0.001). Similarly, when HRC panel was employed, mean R^2 of 0.89 was observed for SHAPEIT2 against 0.85 for Eagle2 (Wilcoxon Signed-Rank test p-value < 0.001).

SNP count comparison details can be found in **Supplementary Tables S1**, **S2**. Regardless of the reference

[2]http://csg.sph.umich.edu/yli/r2_hat.v107.tgz

[3]https://www.lgcgroup.com

panel employed, we observed higher percentage of "high-quality" rare and ultra-rare SNPs for SHAPEIT-IMPUTE2 than Eagle2-IMPUTE2. For instance, 1000G-imputation retrieved 51.02% of "high-quality" rare SNPs using SHAPEIT-IMPUTE2 vs. 48.38% with Eagle2-IMPUTE2. Detailed comparisons for different MAF bins and quality threshold can be found in **Supplementary Section S2**. Nevertheless, we found Eagle2 faster than SHAPEIT2 when computation times were compared; for instance, with HRC Eagle2 was ∼6 times faster than SHAPEIT2 (**Supplementary Table S3**). We therefore imputed the remaining chromosomes on phased output from SHAPEIT2. Comparison of phasing tools by assessing switch error rate was beyond the scope of this paper due to limited resources, for e.g., availability of phased reference panel for an admixed population.

MaCH-Admix vs. IMPUTE2

We found that SHAPEIT-IMPUTE2 performed better than MaCH-Admix. For Chromosome 21, we imputed 1,104,648 and 646,594 SNPs for SHAPEIT-IMPUTE2 and MaCH-Admix, respectively, 549,091 SNPs were overlapping. For SHAPEIT-IMPUTE2 we observed 446,591 bi-allelic SNPs with "Info" $\geq$ 0.40, in contrast with 598,943 SNPs with Rsq $\geq$ 0.30 from MaCH-Admix (**Supplementary Table S4**). SNP counts for different MAF bins based on platform-specific quality index can be found in **Supplementary Table S5**. When the two outputs were compared in terms of *r2hat*, SHAPEIT-IMPUTE2 showed a higheraverage r2hat of 0.62 against 0.36 from MaCH-Admix (Wilcoxon Signed-Rank test p-value < 0.001). Also, MaCH-Admix was 109 times slower than IMPUTE2 (**Supplementary Table S6**), thus, comparison between different panels using MaCH-Admix were excluded due to limited resources. For the remaining of this manuscript, we focused on imputation employing SHAPEIT-IMPUTE2, only.

Comparison Between HRC and 1000G Using SHAPEIT-IMPUTE2

Using SHAPEIT-IMPUTE2, we imputed 81,240,392 and 38,532,090 SNPs across all autosomal chromosomes with 1000G and HRC reference panels, respectively (**Table 2**).

Overall, we observed slightly higher mean R^2 with 1000G than with HRC panel (0.94 vs. 0.92; Wilcoxon p-value < 0.001). Nevertheless, when the analyses were restricted to only "good-" and "high-quality" SNPs, HRC consistently performed better: 60.82% of HRC-imputed SNPs were "good-quality" and 48.87% were "high-quality" (Wilcoxon Signed-Rank test p-value < 0.001). On the contrary, 40.32% of 1000G imputed SNPs were "good-quality" and 30.11% were "high-quality."

Further, we evaluated performance for uncommon, rare and ultra-rare SNPs. For "good-" and "high-quality" SNPs, HRC outperformed 1000G. For example, HRC panel produced 62.85% of "high-quality" rare SNPs, whereas 1000G had 53.83% (**Table 3**). When average imputation "Info" quality was evaluated, HRC-imputation again performed better than with 1000G (Wilcoxon p-value < 0.001) (**Figure 1**).

Next, we restricted our analyses to *overlapping* SNPs across the two reference panels only, based on their chromosome and position mapping, reference and non-reference alleles. For "good-"and "high-quality" SNPs, imputation in both panels performed similarly (**Table 2**). When restricted to uncommon, rare and ultra-rare SNPs, we observed higher percentage of "good-" and "high-quality" SNPs for HRC panel as compared to 1000G reference panel (**Table 3**). For example, 7.44% of HRC-imputed ultra-rare SNPs were "good-quality" vs. 4.95% with the 1000G. 1.69% of HRC-imputed ultra-rare SNPs were "high-quality" vs. 0.75% with the 1000G. Further, Wilcoxon test on "Info" value of "high-quality" ultra-rare SNPs (2,972) again showed better performances when HRC was employed vs. 1000G (P-value < 0.001). Complete list of counts and percentages across reference panels, MAF bins and quality score can be found in **Table 3**.

The Case of G206A and the Effect of Chromosomal Chunk Size on Imputation Quality

SNP rs63750082 is absent from HRC panel therefore no imputation was achieved. Using 1000G reference panel, 12 individuals were imputed as G206A carriers. SNP rs63750082 was imputed with an IMPUTE2 "Info" score of 0.48 using 1MB as chromosomal region parameter. When we increased the chunk size to 5MB, IMPUTE-Info score drastically improved to 0.94 (**Figure 2**). Those patients labeled as mutation-carriers according to imputation were then genotyped: all 12 were confirmed to be G206A carriers, therefore achieving a perfect imputation prediction (100% agreement) for that specific SNP.

Genotype Concordance and Kappa Coefficient

Out of the 1,000 individuals included in our study, 262 had whole exome sequencing (WES) data available (Raghavan et al., 2018). We had 14,157 overlapping SNPs in WES, HRC and 1000G reference panels with 0% missingness in WES data on Chromosome 14; SNPs imputed with each reference panel were compared against WES data separately. When concordance was evaluated, HRC panel performed slightly poorer, despite showing higher number of "high-quality" variants as compared to 1000G (**Table 4**). Using 1000G, we observed 3,542 rare and 35 ultra-rare "high-quality" SNPs; across 262 samples, we counted 1,245 {[(1,245/(3,542 × 262)] × 100 = 0.13%} and 10 (0.10%) mismatches for rare and ultra-rare, respectively. Using HRC, we retrieved 3,759 rare and 93 ultra-rare "high-quality" variants; we observed 2,439 (0.24%) and 32 (0.13%) mismatches for rare and ultra-rare variants, respectively. Details about pipeline can be found in **Supplementary Section S3**.

Next, we computed Cohen's kappa coefficient (K) for 14,157 imputed SNPs common in WES and the two reference panels. For both HRC and 1000G-imputation, we observed Kappa (K) of ∼0.99 for both rare and ultra-rare "high-quality" variants (**Table 4**). Details about pipeline can be found in **Supplementary Section S4**.

TABLE 2 | Type of imputed SNPs across reference panels.

Reference Panel	Multi-allelic SNPs			Bi-allelic SNPs			Total SNPs		
	Total SNPs	Info ≥ 0.40 (%)	Info ≥ 0.80 (%)	Total SNPs	Info ≥ 0.40 (%)	Info ≥ 0.80 (%)	Total SNPs	Info ≥ 0.40 (%)	Info ≥ 0.80 (%)
All SNPs									
1000G	3,319,815	2,586,342 (77.90)	2,061,295 (62.09)	77,920,577	31,423,926 (40.32)	23,468,086 (30.11)	81,240,392	31,423,926 (41.86)	25,529,381 (31.42)
HRC	NA	NA	NA	38,532,090	23,436,980 (60.82)	18,833,790 (48.87)	38,532,090	23,436,980 (60.82)	18,833,790 (48.79)
SNPs overlapping HRC and 1000G									
1000G	NA	NA	NA	30,090,251	22,631,112 (75.21)	18,408,585 (61.17)	30,090,251	22,631,112 (75.21)	18,408,585 (61.17)
HRC	NA	NA	NA	30,090,251	22,438,268 (74.56)	18,395,036 (61.13)	30,090,251	22,438,268 (74.56)	18,395,036 (61.13)

TABLE 3 | SNP Counts for all Bi-allelic uncommon, rare and ultra-rare SNPs.

MAF	1000G			HRC		
	Info ≥ 0	Info ≥ 0.40 (%)	Info ≥ 0.80 (%)	Info ≥ 0	Info ≥ 0.40 (%)	Info ≥ 0.80 (%)
All SNPs						
(1–5%)	6,025,281	5,989,223 (98.90)	5,441,982 (90.31)	5,434,996	5,421,257 (99.84)	5,061,904 (93.13)
(0.1–1%)	20,249,058	16,881,286 (83.36)	10,901,789 (53.83)	11,780,671	10,931,924 (92.79)	7,404,808 (62.85)
(0–0.1%)	44,562,205	1,490,434 (3.34)	242,717 (0.544)	15,055,433	828,256 (5.50)	174,673 (1.16)
SNPs overlapping HRC and 1000G						
(1–5%)	5,624,956	5,604,308 (99.63)	5,148,285 (91.52)	5,396,207	5,385,364 (99.79)	5,037,187 (93.34)
(0.1–1%)	11,875,603	10,442,603 (87.93)	7,027,312 (59.17)	10,945,899	10,268,136 (93.80)	7,060,908 (64.50)
(0–0.1%)	6,314,479	312,967 (4.95)	47,614 (0.75)	7,519,807	560,043 (7.44)	127,423 (1.69)

Effects of Ancestry on Imputation Quality

We evaluated the effect of individual ancestral component separately on SNP mismatches for Chromosome 14 on 262 individuals. For both reference panels we found that higher African ancestry (YRI) was associated with higher number of mismatches (**Supplementary Table S7**). For instance, with 1000G reference panel, for rare variants ("Info" $\geq$ 0.80), we observed an estimate of 1.46 (P-value $<$ 0.001) for YRI component (indicating that for each unit increase in YRI ancestry, it results in 1.46 additional mismatches). Details on confidence intervals and robust standard errors can be found in **Supplementary Table S7** and **Supplementary Section S5**). We did not observe significant effect of ancestry on "high-quality" ultra-rare variants in both panels.

DISCUSSION

This study examined imputation performances in a cohort Caribbean Hispanics, focusing on uncommon, rare and ultra-rare variant, by comparing different phasing and imputation tools, as well as evaluating the effects of different reference panels. Overall, uncommon and rare variants can be well imputed in this population, characterized by a unique genetic background. Caribbean Hispanics are admixed with 59% of their genetic component from European, 32% African, and 8% Native American ancestry (Tosto et al., 2015). Due to their genetic makeup and unique linkage disequilibrium patterns, admixed populations offer unique opportunity in studying complex diseases. First, disease prevalence varies across ethnic groups (Igartua et al., 2015) and certain admixed populations show higher incidence rates and prevalence (e.g., Alzheimer's disease, diabetes etc.) or lower ones (e.g., multiple sclerosis). Second, variants that are ethnic-specific may explain a higher prevalence of the disease of interest in admixed groups.

In the present study, we examined multiple parameters of imputation using the Caribbean Hispanics population. First, we found that imputation using SHAPEIT-IMPUTE2 phasing generated better results than Eagle2-IMPUTE2, and SHAPEIT-IMPUTE2 is superior to MaCH-Admix in terms of imputation performances and process time.

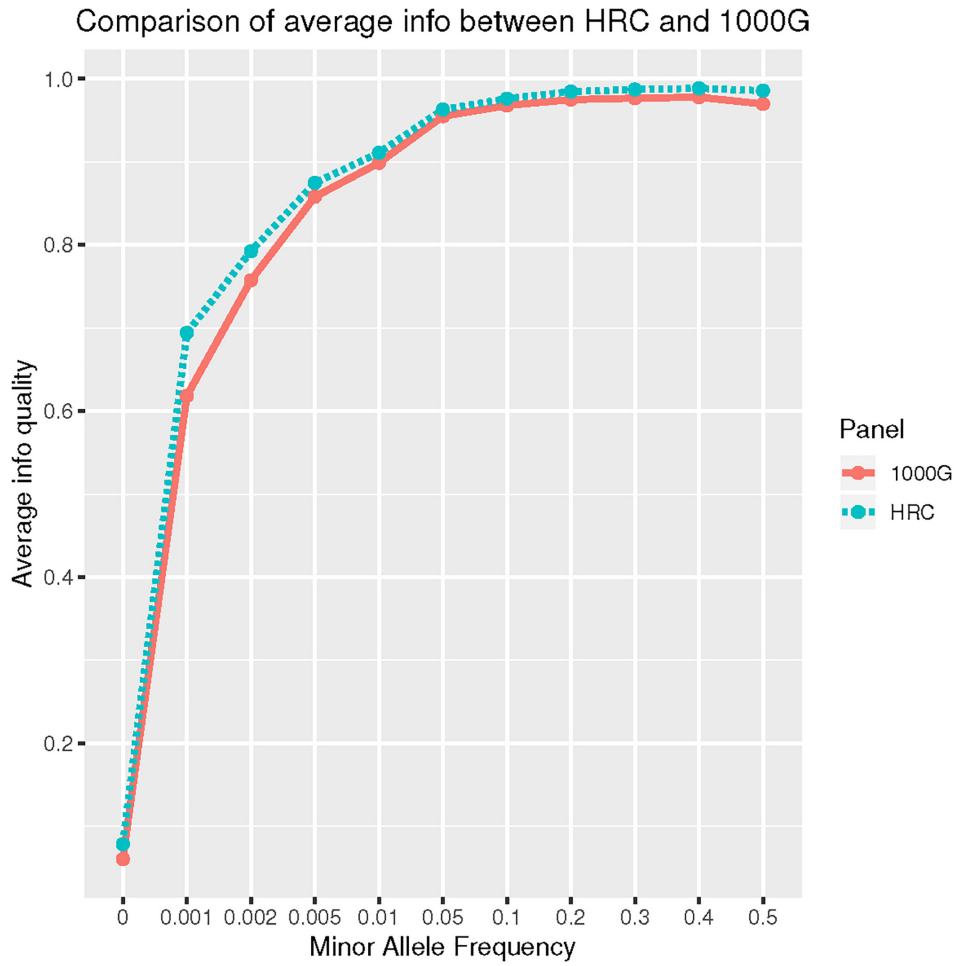


FIGURE 1 | Comparison of average Info quality between HRC and 1000G reference panel for all autosomal chromosomes.

TABLE 4 | Comparison for mismatch counts and Kappa (*K*) for HRC and 1000G using WES data on Chromosome 14.

MAF	1000G Info ≥ 0.80				HRC Info ≥ 0.80			
	SNP	Total SNPs in all persons*	Mismatch	Kappa *(K)*	SNP	Total SNPs in all persons*	Mismatch	Kappa *(K)*
(1–5%)	2,354	610,550	7,397 (1.22%)	0.99	2,264	587,961	8,963 (1.52%)	0.99
(0.1–1%)	3,542	926,109	1,245 (0.13%)	0.99	3,759	982,734	2,439 (0.24%)	0.99
(0–0.1%)	35	9,163	10 (0.10%)	0.99	93	24,348	32 (0.13%)	0.99

**Less value than 262*SNP because imputed with poor posterior probability failed to be converted from .gen to PLINK format.*

Using SHAPEIT-IMPUTE2, 1000G SNPs outnumbered HRC panel because of the higher number of SNPs included in the reference panel itself. However, when we restricted our analyses to overlapping "good-" and "high-quality" SNPs (i.e., those variants that most likely would be included in association analyses), HRC-imputation outperformed 1000G with higher. The superior performance of HRC over 1000G was confirmed also when we focused on uncommon, rare and ultra-rare SNPs only. Our findings confirm data in literature, i.e., reference panels with higher number haplotypes perform better in different scenarios.

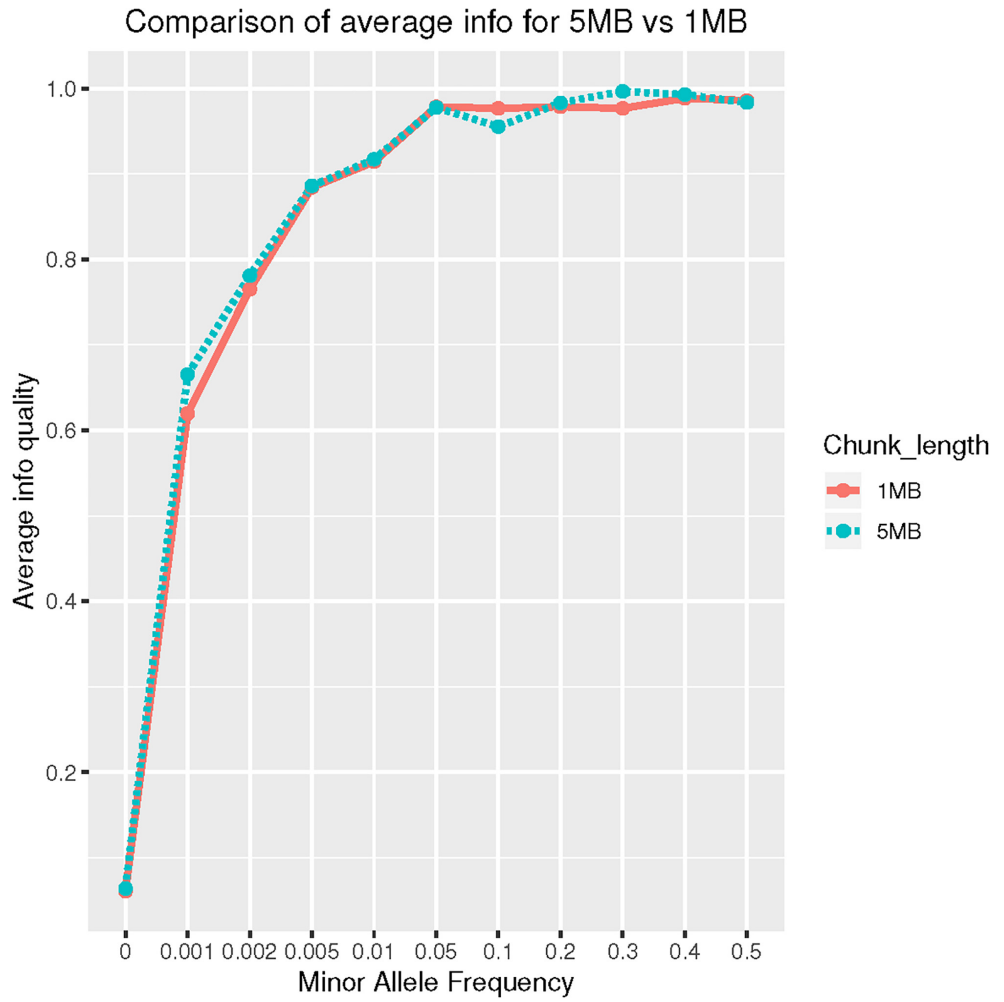


FIGURE 2 | Comparison of average Info on CHR14: 70–75 MB (5 MB) vs. 73–74 MB (1 MB) region.

Additional investigations are needed in order to apply our findings to other admixed and non-admixed populations.

Overall, higher quality of imputation for rare and ultra-rare variants was also confirmed when we tested results against sequencing data. Finally, higher YRI global ancestry was found to significantly impair SNP imputation, suggesting that imputation quality decreases with increased African ancestry.

Lastly, SHAPEIT-IMPUTE2 with 1000G reference panel was successful in identifying G206A mutation carriers. We also noticed that imputation quality drastically improved when imputation was conducted using large (5MB) chunk size as compared to small (1MB) chunks. This seems to contradict previous observation: Zhang et al. (2011) studied the effect of window size on imputation in an African-American. They concluded that window size of 1MB could be considered acceptable. Possible explanations for these different results might be the more complex admixture of CH compare to AA (three-way vs. two-way admixed) and a more complex LD pattern for the G206A region. Ultimately, we recommend to consider a wider window size to achieve high-quality imputation in specific variants that fail under default settings.

This work has limitations. First, we could carry out the comparison between the two reference panels restricting the analyses to overlapping variants only, limiting our observation to a subset of the variants included in the 1000G panel. This is a result of the HRC composition, which is composed by several studies and ended up including only a consensus number of variants. Second, we tested the agreement between imputed and sequenced variants in a smaller subset of individuals that had both GWAS and WES data available.

DATA AVAILABILITY

The datasets for this manuscript are not publicly available because data will be available soon through dbgap website. Requests to access the datasets should be directed to gt2260@cumc.columbia.edu.

ETHICS STATEMENT

All participants provided written informed consent. Ethical approval for this study was obtained from the Columbia University committee.

AUTHOR CONTRIBUTIONS

SS and GT conceived and designed the study. SS, GT, JL, BV, RM, MM, RL, IJ-V, JM, AB, and DR-D acquired and analyzed the data and drafted the manuscript or figures.

FUNDING

This study was supported by funding from the National Institute on Aging [R21AG054832 (GT); 5R37AG015473 and RF1AG015473 (RM); R56 AG051876 and R01 AG058918 (JL)] and the BrightFocus Foundation [A2015633S (JL)].

ACKNOWLEDGMENTS

We thank the EFIGA study participants and the EFIGA research and support staff for their contributions to this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00239/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Arnold, S. E., Vega, I. E., Karlawish, J. H., Wolk, D. A., Nunez, J., Negron, M., et al. (2013). Frequency and clinicopathological characteristics of presenilin 1 Gly206Ala mutation in Puerto Rican Hispanics with dementia. *J. Alzheimers Dis.* 33, 1089–1095. doi: 10.3233/JAD-2012-121570

Athan, E. S., Williamson, J., Ciappa, A., Santana, V., Romas, S. N., Lee, J. H., et al. (2001). A founder mutation in presenilin 1 causing early-onset Alzheimer disease in unrelated Caribbean Hispanic families. *JAMA* 286, 2257–2263. doi: 10.1001/jama.286.18.2257

Browning, B. L., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005

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

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Delaneau, O., Zagury, J. F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6. doi: 10.1038/nmeth.2307

Genomes Project, C., 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

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Ha, N. T., Freytag, S., and Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. *Eur. J. Hum. Genet.* 22, 1124–1130. doi: 10.1038/ejhg.2013.304

Hancock, D. B., Levy, J. L., Gaddis, N. C., Bierut, L. J., Saccone, N. L., Page, G. P., et al. (2012). Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PLoS One* 7:e50610. doi: 10.1371/journal.pone.0050610

Herzig, A. F., Nutile, T., Babron, M. C., Ciullo, M., Bellenguez, C., and Leutenegger, A. L. (2018). Strategies for phasing and imputation in a population isolate. *Genet. Epidemiol.* 42, 201–213. doi: 10.1002/gepi.22109

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959. doi: 10.1038/ng.2354

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* 6:8111. doi: 10.1038/ncomms9111

Igartua, C., Myers, R. A., Mathias, R. A., Pino-Yanes, M., Eng, C., Graves, P. E., et al. (2015). Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat. Commun.* 6:5965. doi: 10.1038/ncomms6965

Lee, J. H., Cheng, R., Vardarajan, B., Lantigua, R., Reyes-Dumeyer, D., Ortmann, W., et al. (2015). Genetic modifiers of age at onset in carriers of the G206A mutation in PSEN1 with familial Alzheimer disease among caribbean hispanics. *JAMA Neurol.* 72, 1043–1051. doi: 10.1001/jamaneurol.2015.1424

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100–1104. doi: 10.1126/science.1153717

Liu, E. Y., Li, M., Wang, W., and Li, Y. (2013). MaCH-admix: genotype imputation for admixed populations. *Genet. Epidemiol.* 37, 25–37. doi: 10.1002/gepi.21690

Liu, Q., Cirulli, E. T., Han, Y., Yao, S., Liu, S., and Zhu, Q. (2015). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Brief Bioinform.* 16, 549–562. doi: 10.1093/bib/bbu035

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443–1448. doi: 10.1038/ng.3679

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511. doi: 10.1038/nrg2796

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. Med.* 22, 276–282. doi: 10.11613/BM.2012.031

Nagy, R., Boutin, T. S., Marten, J., Huffman, J. E., Kerr, S. M., Campbell, A., et al. (2017). Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Med.* 9:23. doi: 10.1186/s13073-017-0414-4

Nelson, S. C., Stilp, A. M., Papanicolaou, G. J., Taylor, K. D., Rotter, J. I., Thornton, T. A., et al. (2016). Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Hum. Mol. Genet.* 25, 3245–3254. doi: 10.1093/hmg/ddw174

Pei, Y. F., Zhang, L., Li, J., and Deng, H. W. (2010). Analyses and comparison of imputation-based association methods. *PLoS One* 5:e10827. doi: 10.1371/journal.pone.0010827

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Raghavan, N. S., Brickman, A. M., Andrews, H., Manly, J. J., Schupf, N., Lantigua, R., et al. (2018). Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann. Clin. Transl. Neurol.* 5, 832–842. doi: 10.1002/acn3.582

Roshyara, N. R., Kirsten, H., Horn, K., Ahnert, P., and Scholz, M. (2014). Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet.* 15:88. doi: 10.1186/s12863-014-0088-5

Surakka, I., Sarin, A.-P., Ruotsalainen, S. E., Durbin, R., Salomaa, V., Daly, M. J., et al. (2016). The rate of false polymorphisms introduced when imputing genotypes from global imputation panels. *bioRxiv* [Preprint]. doi: 10.1101/080770

Tosto, G., Fu, H., Vardarajan, B. N., Lee, J. H., Cheng, R., Reyes-Dumeyer, D., et al. (2015). F-box/LRR-repeat protein 7 is genetically associated with Alzheimer's disease. *Ann. Clin. Transl. Neurol.* 2, 810–820. doi: 10.1002/acn3.223

Verma, S. S., de Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., et al. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* 5:370. doi: 10.3389/fgene.2014.00370

Zhang, B., Zhi, D., Zhang, K., Gao, G., Limdi, N. N., and Liu, N. (2011). Practical consideration of genotype imputation: sample size, window size, reference choice, and untyped rate. *Stat. Interface* 4, 339–352. doi: 10.4310/SII.2011.v4.n3.a8

Zheng, H. F., Ladouceur, M., Greenwood, C. M., and Richards, J. B. (2012). Effect of genome-wide genotyping and reference panels on rare variants imputation. *J. Genet. Genom.* 39, 545–550. doi: 10.1016/j.jgg.2012.07.002

Zheng, H. F., Rong, J. J., Liu, M., Han, F., Zhang, X. W., Richards, J. B., et al. (2015). Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One* 10:e0116487. doi: 10.1371/journal.pone.0116487

Zhou, H., Alexander, D., and Lange, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Stat. Comput.* 21, 261–273. doi: 10.1007/s11222-009-9166-3

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 Physiology

ORIGINAL RESEARCH
published: 12 April 2019
doi: 10.3389/fphys.2019.00380

Computational Modeling of Glucose Uptake in the Enterocyte

Nima Afshar[1], *Soroush Safaei*[1], *David P. Nickerson*[1], *Peter J. Hunter*[1] *and Vinod Suresh*[1,2]*

[1] *Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand,* [2] *Department of Engineering Science, University of Auckland, Auckland, New Zealand*

Absorption of glucose across the epithelial cells of the small intestine is a key process in human nutrition and initiates signaling cascades that regulate metabolic homeostasis. Validated and predictive mathematical models of glucose transport in intestinal epithelial cells are essential for interpreting experimental data, generating hypotheses, and understanding the contributions of and interactions between transport pathways. Here we report on the development of such a model that, in contrast to existing models, incorporates mechanistic descriptions of all relevant transport proteins and is implemented in the CellML framework. The model is validated against experimental and simulation data from the literature. It is then used to elucidate the relative contributions of the sodium-glucose cotransporter (SGLT1) and the glucose transporter type 2 (GLUT2) proteins in published measurements of glucose absorption from human intestinal epithelial cell lines. The model predicts that the contribution of SGLT1 dominates at low extracellular glucose concentrations (<20 mM) and short exposure times (<60 s) while the GLUT2 contribution is more significant at high glucose concentrations and long durations. Implementation in CellML permitted a modular structure in which the model was composed by reusing existing models of the individual transporters. The final structure also permits transparent changes of the model components and parameter values in order to facilitate model reuse, extension, and customization (for example, to simplify, or add complexity to specific transporter/pathway models, or reuse the model as a component of a larger framework) and carry out parameter sensitivity studies.

Keywords: computational modeling, glucose uptake, SGLT1, GLUT2, CellML, OpenCOR

OPEN ACCESS

Edited by:
Vinicius Maracaja-Coutinho, Universidad de Chile, Chile

Reviewed by:
Daniel Goldman, University of Western Ontario, Canada
Steve McKeever, Uppsala University, Sweden

***Correspondence:**
Vinod Suresh
v.suresh@auckland.ac.nz

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *14 December 2018*
Accepted: *19 March 2019*
Published: *12 April 2019*

Citation:
Afshar N, Safaei S, Nickerson DP, Hunter PJ and Suresh V (2019) Computational Modeling of Glucose Uptake in the Enterocyte. Front. Physiol. 10:380. doi: 10.3389/fphys.2019.00380

1. INTRODUCTION

Almost all of the nutrients, electrolytes, and water from food are absorbed into blood capillaries through the mucosa of the small intestine. Most absorption processes in the small intestine are driven by an electrochemical gradient of ions across the boundary of epithelial cells (enterolyses) lining the lumen. Transporter proteins embedded in the apical membrane carry ions and nutrients into the enterocyte. Other transporters in the basolateral membrane then extrude the ions into the interstitial space from where they enter capillary blood by diffusion. Carbohydrates are the main source of energy in the body. They break down to monosaccharides like glucose, which is the most important carbohydrate fuel in the cell. Therefore the uptake and transport of glucose through the small intestine epithelial cells is a vital aspect of human nutrition. Subsequent transport and metabolism of the absorbed species triggers responses such as hormone release, appetite regulation and growth via complex physiological feedback pathways. A mechanistic understanding

of these pathways and how they are disrupted in disease is lacking, partly due to the difficulties of making experimental measurements in the luminal and capillary compartments. A validated computational model of the absorption pathways can overcome these difficulties by providing quantitative predictions of concentrations and transport rates in the lumen and cell compartments (Hunter and Borg, 2003; Ingalls, 2012).

Many studies in the past few decades have focussed on mathematical modeling of the glucose-insulin control system in order to study how metabolism and the regulatory system are disrupted in diseases like diabetes (reviewed in Palumbo et al., 2013). At the cellular level, models of glucose uptake and transport in the kidneys (Weinstein, 2015), glucose homeostasis in the liver (König et al., 2012), and glucose sensing (Riz and Pedersen, 2015) have been developed. In contrast, mathematical modeling of glucose uptake by the enterolyses lining the small intestinal mucosa has attracted little attention. The first model of glucose transport in the enterocyte was developed by Thorsen et al. (2014). The model focussed on the regulation of Na,K-ATPase in enterolyses during glucose absorption. It considered SGLT1 as the sole pathway for glucose entry into the cell at the apical membrane and studied how the intracellular Na+ concentration can be maintained in the face of SGLT1-associated Na+ influx. One limitation of the model is the absence of a GLUT2 pathway for glucose entry at the apical membrane. The role of apical GLUT2 is still a matter of controversy with some studies indicating its presence and importance for glucose uptake (Kellett and Brot-Laroche, 2005; Zheng et al., 2012) while others have suggested SGLT1 as the dominant or sole pathway (Gorboulev et al., 2012; Röder et al., 2014). Differences in experimental conditions and data interpretation are partly the reason for lack of consensus (Kellett, 2012; Koepsell and Gorboulev, 2012). In this work, we developed a mathematical model that includes apical GLUT2 and parameterized it against published experimental data. We then used the model to examine the relative contributions of SGLT1 and GLUT2 in published cell culture data on glucose uptake (Zheng et al., 2012). Finally we assessed the impact of increased glucose transporter expression on uptake rates in diabetes.

The Thorsen model incorporated a mixture of mechanistic transporter models (e.g., SGLT1, basolateral GLUT2), empirical flux expressions (e.g., NaK-ATPase, an effective Na-Cl co-transporter), and diffusive membrane fluxes for Na+, K+, and Cl. We modified this framework to explicitly incorporate mechanistic models of all relevant transporters. In particular, we replaced the Na-Cl co-transporter in the original model with individual models for the anion exchanger 1 (AE1) and Na+ /H+ exchanger (NHE3) proteins at the apical membrane and incorporated ENaC and CFTR channels for apical Na+ and Cl− transport. This makes it possible to use the model to study scenarios where the expression and/or function of these transport proteins is altered, for example in gene knockout/mutation studies or the use of channel inhibitors and agonists.

The model is implemented in the open source, extensible markup language (XML)-based CellML modeling environment used to represent mathematical models of biology based on ordinary differential and algebraic equations (Cuellar et al., 2003). We adopted a modular, compositional approach to model construction by reusing CellML models of individual transport proteins encoded in an online, curated repository [Physiome Model Repository (PMR, models.physiomeproject.org)] to facilitate the sharing of models (Yu et al., 2011). The complete model, including parameter values, simulation software and simulation conditions, can be downloaded from PMR with the following link: https://models.physiomeproject.org/workspace/572.

2. METHODS

2.1. Model Construction

We constructed a mathematical model of the epithelial cell of a small intestine (enterocyte) that incorporates the relevant transport proteins identified in the literature (Barrett and Keely, 2015) and diffusion pathways (**Figure 1**). The membrane localization and function of these transporters and the source of the original mathematical models are listed in **Table 1**. The apical (luminal) and basolateral (interstitial) surface of the cell are in contact with distinct extracellular compartments. Transport of substances occurs across the membranes as well as directly between the extracellular compartments across the paracellular junctions. The variables to be solved in the model are chemical species (Na^+, K^+, H^+, Cl^-, HCO_3^-, glucose) concentrations in each compartment and the two membrane potentials. Flux balance and electric charge conservation laws yield the governing equations of the model. Water transport is not included and hence we limit ourselves to modeling iso-osmotic transport. Model equations are provided in the **Supplementary Material**.

The model was implemented in the open-source, modular CellML framework. CellML is an XML based language commonly used to encode and simulate mathematical models based on algebraic and ordinary differential equations. Encoded models are available in an online, curated repository [Physiome Model Repository-PMR(models.physiomeproject.org)] (Yu et al., 2011). Reuse of models and components within models is possible through the use of the `import` element that enables encapsulation of other CellML files within a CellML model and facilitates a modular, compositional approach to the construction of complex models. The application of this approach in the enterocyte model is shown in **Figure 2**. Existing models of the individual transporters were imported into the top level model file (`modular_model.cellml`). Units and parameters for all components as well as initial conditions for specific simulations were specified in separate `.cellml` files and also imported into the top level file. The models were encapsulated as a group into the `enterocyte` component in which the overall balance equations for the chemical species and electric currents were coded. The `mappings` element links variables that are common between the different components, e.g., glucose concentrations in `GLUT2.cellml`, `SGLT1.cellml` and `enterocyte`. The `environment` component comprises independent variables that are common to all components, which in this case, is solely time.

The model was coded and simulated in OpenCOR (Version 0.5) (Garny and Hunter, 2015). The CVODES solver was used

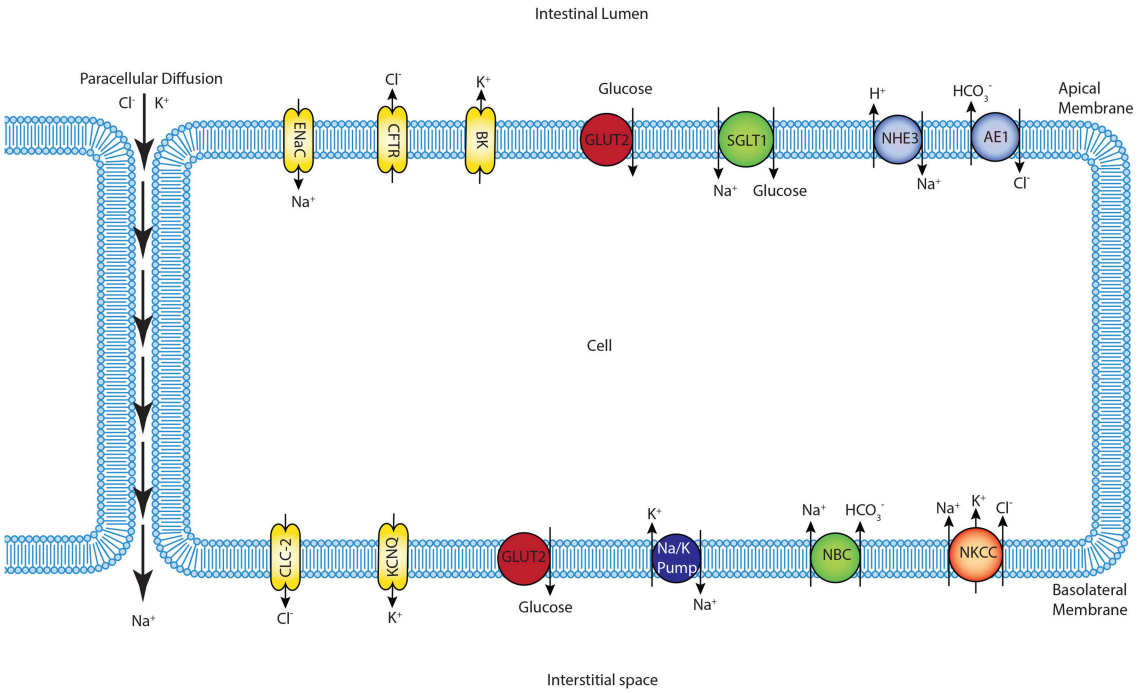


FIGURE 1 | Schematic of enterocyte showing the relevant transporters in the apical and basolateral membrane along with the apical (lumen) and basolateral (interstitium) extracellular domains.

TABLE 1 | List of transporters used in the model along with their locations and roles.

Transporter	Location	Role	Chemical Species	Source of the mathematical model
SGLT1	Apical	Cotransporter	1 Glucose, 2 Na^+	Parent et al., 1992
NaK ATPase	Basolateral	Exchange Pump	3Na^+, 2K^+	Thorsen et al., 2014
GLUT2	Apical and Basolateral	Uniporter Protein	Glucose	Pradhan et al., 2013
NHE3	Apical	Antiporter	1 Na^+, 1 H^+	Weinstein, 1995
AE1	Apical	Antiporter	1 Cl^-, 1 HCO_3^-	Weinstein, 2000
BK	Apical	Channel	$K+$	Fong et al., 2016
CFTR	Apical	Channel	Cl^-	Fong et al., 2016
CLC-2	Basolateral	Channel	Cl^-	Fong et al., 2016
ENaC	Apical	Channel	Na^+	Fong et al., 2016
IK	Basolateral	Channel	K^+	Fong et al., 2016
NBC	Basolateral	Cotransporter	1 Na^+, 3 HCO_3^-	Østby et al., 2009
NKCC1	Basolateral	Cotransporter	1 Na^+, 1 K^+, 2 Cl^-	Palk et al., 2010

with the BDF integration method and Newton iterations. All of the models including their parameters can be downloaded from PMR with the following link: https://models.physiomeproject.org/workspace/572.

2.2. Comparison With Experiments

The model was validated against published experimental measurements of glucose uptake in the human enterocyte-like cell lines Caco-2 and IEC6 Zheng et al. (2012). In the experiments, the cells were cultured on impermeable surfaces for 10–15 days in high glucose (25 mM) medium. To measure glucose uptake, varying concentrations (0.5–50 mM) of glucose were introduced into the apical chamber in a buffer solution with a baseline composition 130 mM NaCl, 4 mM KH_2PO_4, 1 mM $CaCl_2$. The osmolarity of the buffer was maintained during the measurements by modulating the NaCl content such that if the glucose concentration was x mM, NaCl concentration was $130 - x/2$ mM. After exposure to the glucose stimulus for different durations (30–600 s), cells were lysed and intracellular glucose and protein concentrations were measured. Since the measurements were reported in nanomole glucose per milligram (mg) protein, the data were converted to concentration units

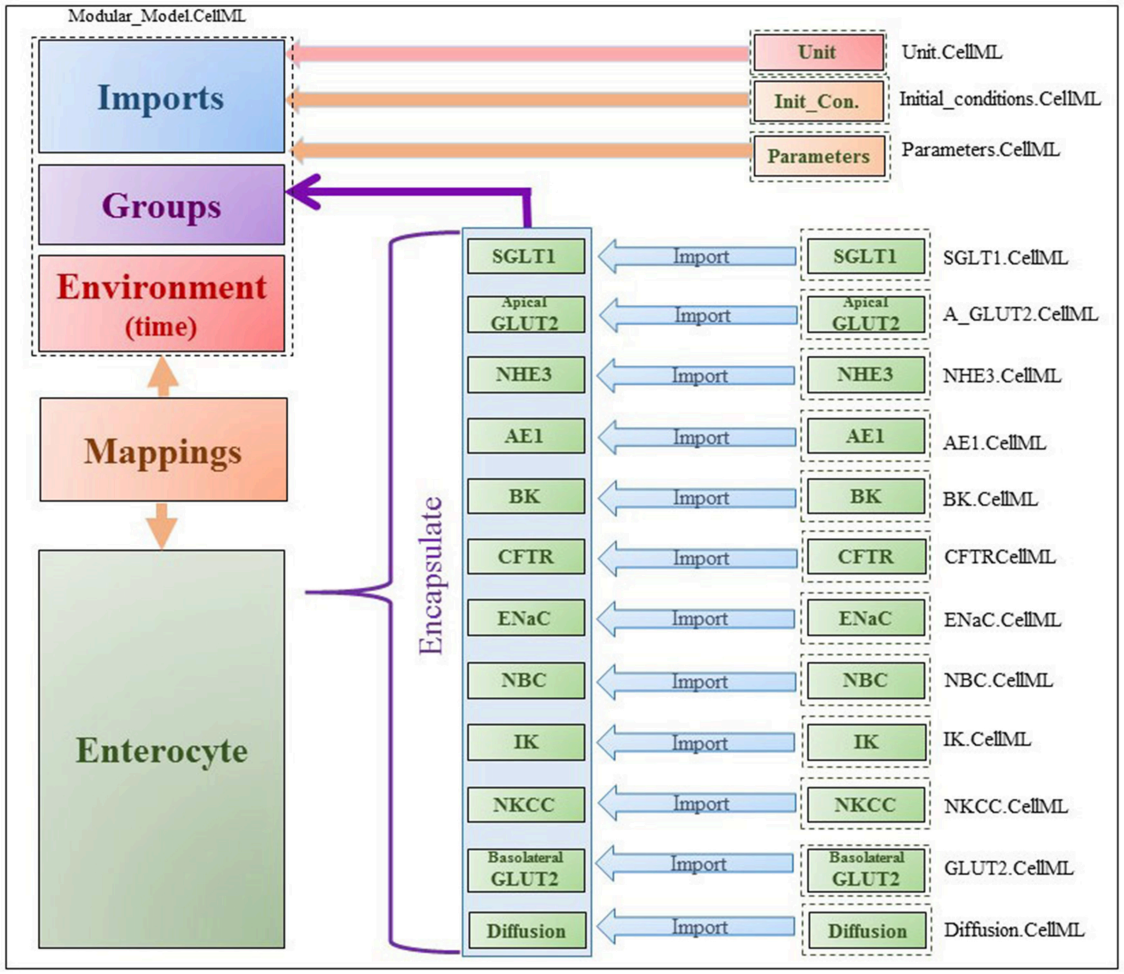


FIGURE 2 | Figure shows the modularity of CellML model. Encapsulation hierarchy (purple), the CellML model imports (blue) and the other key parts (units, parameters,components and mappings) of the top level CellML model.

(millimole per liter, mM) by doing the unit conversion from nanomole/m^3 to mM and also multiplying by the cellular protein concentration (mg protein per ml cell volume). The conversion factor a (protein density) was used as a fitting parameter in a non-linear Generalized Reduced Gradient optimization to match model outputs to the data. The optimization was done using the Microsoft Excel Solver (Microsoft Office 2013) by minimizing the least square error between model predicted and measured intracellular glucose concentration.

In the simulations, the apical compartment was treated as an infinite bath of constant composition based on the experimental conditions. Since the cells were cultured on an impermeable substrate, the volume of the basolateral compartment (V_b) was not measured. In the simulations, V_b was fixed at different multiples (m = 0.1, 1, 10) of the cell volume (V_c) and also as an infinite bath to generate a range of predictions. This allowed us to account for the uncertainty in the actual volume of the basolateral compartment. For finite values of V_b, the composition of the basolateral compartment cannot be regarded as constant and was instead determined by the flux of glucose/ions across the basolateral membrane. Since the experiments were conducted under iso-osmotic conditions, there is no water transfer between the compartments and hence V_b and V_c were held fixed for the duration of each simulation.

3. RESULTS

3.1. Steady State and Dynamic Responses

The model was first checked for physiological consistency by determining intracellular concentrations and membrane potentials in the absence and presence of a glucose stimulus. Steady state values of the model variables were computed with no glucose in the extracellular compartments. In these simulations, the composition of the apical and basolateral compartments were identical and held constant (140 mM Na^+, 5.4 mM K^+, 103 mM Cl^-). Results were consistent with reported values (**Table 2**).

Next, the dynamic response to an apical glucose stimulus was determined. The model was initialized in the steady state

TABLE 2 | Reported values for intracellular ions concentration from simulated model and literature.

Ion	Model result	Reported value	Reference
Na^+ (mM)	61	45–65	Nellans and Schultz, 1976; Okada et al., 1976
K^+ (mM)	127	120–40	Okada et al., 1976; Vogalis, 2000
Cl^- (mM)	69	50–70	Frizzell et al., 1973; Nellans et al., 1973; Okada et al., 1976
Apical (lumen-cell) membrane potential (mV)	−30	−36 ± 0.5	Rose and Schultz, 1971
Basolateral (interstitium-cell) membrane potential (mV)	−36	−40.5 ± 0.8	Rose and Schultz, 1971
pH	7.16	7.2	Shimada and Hoshi, 1987

described in **Table 2** and a time dependent, extracellular glucose stimulus previously used in the literature (Thorsen et al., 2014) was applied at $t = 60\ s$ (**Figure 3A**). Other extracellular variables were maintained at the same values used for the previous set of simulations. The stimulus causes a depolarization of both membranes (**Figure 3B**). Membrane potentials recover rapidly to baseline after around 100 s and mirror the time course of the stimulus. Transient changes in the transepithelial potential difference ($\approx$ 1.4 mV increase) are of the same direction and comparable magnitude to values reported in the literature (1.9 ± 0.1 mV ;Rose and Schultz, 1971) while changes in the apical potential ($\approx$ 12 mV increase) are higher than values reported in the same study (6 ± 0.5 mV) (**Figures 3B,C**). Intracellular ion concentrations and pH all exhibit a slower transient response than the membrane potentials that lasts for ≈200–300 s.

3.2. Comparison With the Thorsen et al. (2014) Model

Since our model is similar to that developed by Thorsen et al. (2014), we compared the responses of both models when the same parameters (**Table 3**), initial conditions and glucose stimulus were used. Model outputs were normalized against the steady state values of the Thorsen model and are shown in **Figure 4**. A few observations may be made: in the absence of a glucose stimulus, steady state values of the membrane potentials in our model are around 30% lower while the transepithelial potential is around 80% higher. Steady state values for concentration of chloride, potassium, and glucose are $5-10\%$ lower than the values in the previous model whereas for sodium it is about 10% higher. In response to a glucose stimulus, our model has a larger change in membrane potentials and intracellular glucose, but smaller changes in sodium and potassium. Chloride responses are of almost the same magnitude in both models. The duration of the transients are similar in both models, except for glucose where our model has a similar rise time, but a slower decay (around 2 times slower).

3.3. Comparison Against Cell Culture Data

Finally, model predictions were compared against measurements carried out in cell culture studies (Zheng et al., 2012). The experiments used Caco-2 and IEC6 cell lines. While Caco-2 expresses both SGLT1 and GLUT2, IEC6 cells do not express GLUT2. We therefore turned off the expression of GLUT2 in the apical membranes to simulate these cells.

Model predictions of the intracellular glucose concentrations are in good agreement with the measurements over the entire range of time points and apical glucose concentrations for both cell lines (**Figure 5**). As shown in **Figures 5A,B** at 30 and 60 s of exposure, glucose concentrations in both cell lines have a tendency to level off at higher concentration of glucose in the apical compartment. IEC6 still has the same behavior for longer exposure times (300 and 600 s) whereas concentrations in Caco2 do not saturate with increasing glucose concentration in the apical compartment (**Figures 5C,D**). The protein density parameter a are quite close to each other for Caco-2 cell line and varies for IEC-6 cell line to fit the four exposure durations (**Table 4**).

Together, these results indicate that the model is able to reproduce a range of independent experimental observations. Next we present applications of the validated model to address questions about glucose uptake pathways in health and disease.

3.4. Role of Apical GLUT2 in Glucose Uptake

In the original study of Zheng et al., the experimental data in **Figures 5A–D** were interpreted as indicating the presence of GLUT2-mediated uptake at the apical membrane (Zheng et al., 2012). We investigated if an alternative explanation was possible whereby SGLT1 expression levels in the model could be tuned to reproduce the same trends in intracellular glucose concentration. In **Figures 6A–D**, the data for Caco-2 cells at the 600 s time point are compared to the model with varying levels of apical GLUT2 and SGLT1. The baseline model with normal expression of SGLT1 and apical GLUT2 provides a good fit to the data over the full range of apical glucose concentrations (**Figure 6A**). When apical GLUT2 is turned off with no changes in SGLT1 expression (**Figure 6B**), model predictions of intracellular glucose are low compared to the data for apical glucose concentrations higher than 10 mM. In addition, model predictions saturate after around 20 mM of apical glucose while the data shows an increasing trend. A higher expression of SGLT1 was also examined and can provide a better match to the data in the absence of apical GLUT2. With no apical GLUT2 and 2-fold levels of baseline SGLT1 (**Figure 6C**) the model overpredicts the data at low apical glucose concentrations (<10 mM) and underpredicts the data at apical glucose concentrations >40 mM. When SGLT1 levels are increased to 3 times the baseline value, the model overpredicts the data over the whole range, except at a apical glucose of 50 mM (**Figure 6D**).

In order to explain these results, the contribution of SGLT1 and GLUT2 to the apical glucose flux is shown in **Figure 7** following 600 s of exposure to apical glucose. It is seen that

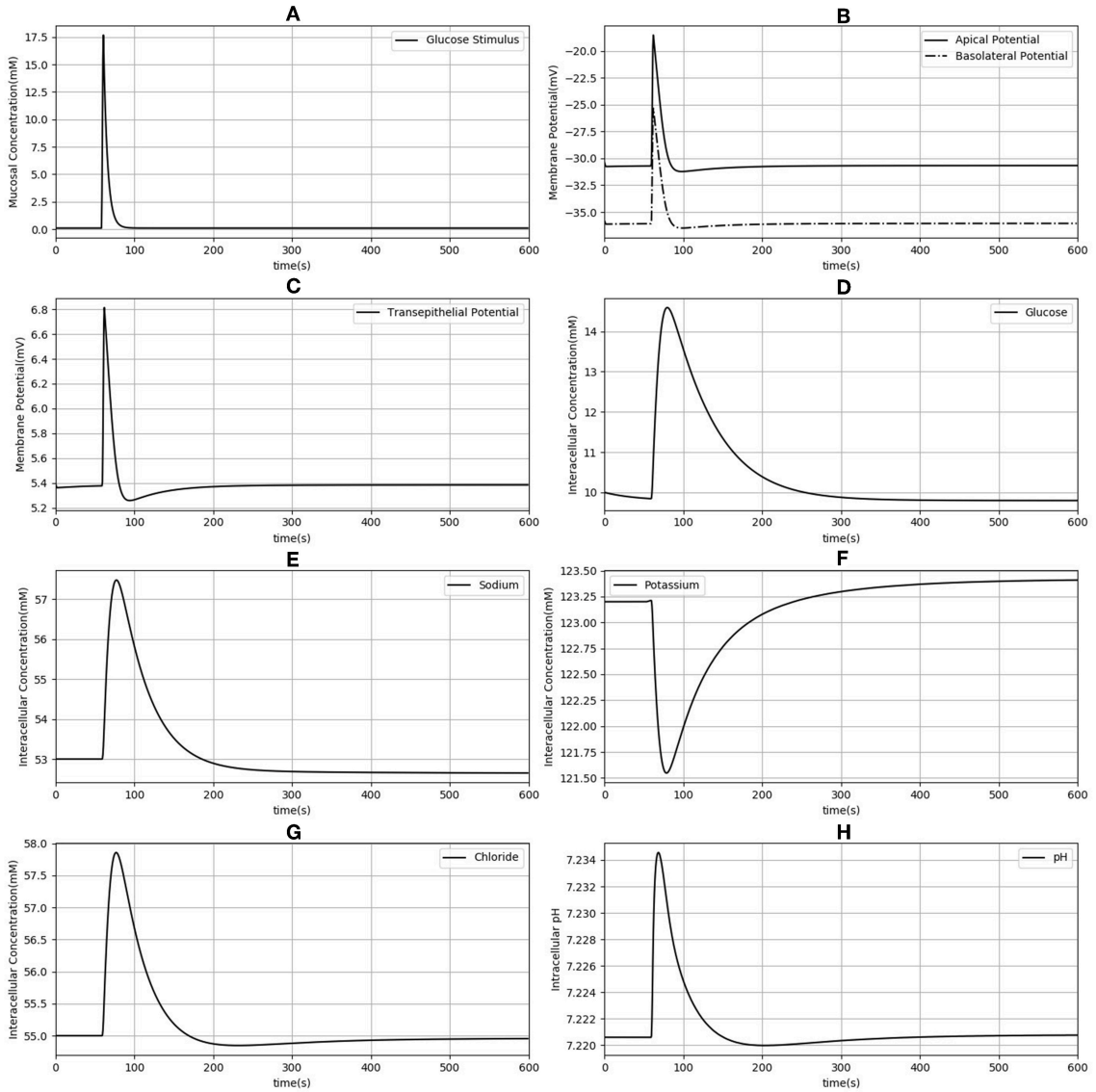


FIGURE 3 | Dynamic response of the model to an extracellular glucose stimulus. The stimulus consists of a step increase followed by an exponential decay **(A)**. Apical and basolateral membrane potentials **(B)**, transepithelial potential **(C)**, and intracellular concentrations of glucose **(D)**, sodium **(E)**, potassium **(F)**, chloride **(G)**, and pH **(H)** are shown.

TABLE 3 | Parameter values used in the simulations.

Parameter	Value in our model (Figures 3, 4)	Value in our model (Other figures)	Unit
nSGLT1	18×10^7	4×10^7	–
nA_{GLUT2}	0	42×10^7	–
nB_{GLUT2}	14×10^6	14×10^7	–
V_{cell}	6×10^{-16}	2×10^{-15}	m^3
Capacitance	1×10^{-5}	1×10^{-5}	μF

for apical glucose concentrations up to around 25 mM, the flux through SGLT1 is higher than GLUT2 flux but after that it starts to saturate, while the GLUT2 flux continues to increase and get higher than SGLT1 flux. This behavior looks similar to the previous experimental study Kellett and Helliwell (2000) which at the apical glucose concentration of 50 mM the glucose flux through GLUT2 is about 2 times higher than flux via SGLT1, Thus, varying the level of SGLT1 in the absence of apical GLUT2 is unable to capture the shape and magnitude of the experimental measurements since transport through SGLT1 saturates at an apical glucose concentration of about 25 mM. This suggests that apical GLUT2 is essential to account for the data from Zheng et al. (2012).

3.5. Glucose Uptake in Diabetes

In diabetes, expression levels of SGLT1 and GLUT2 in the small intestine are reported to be increased 3 to 4-fold compared

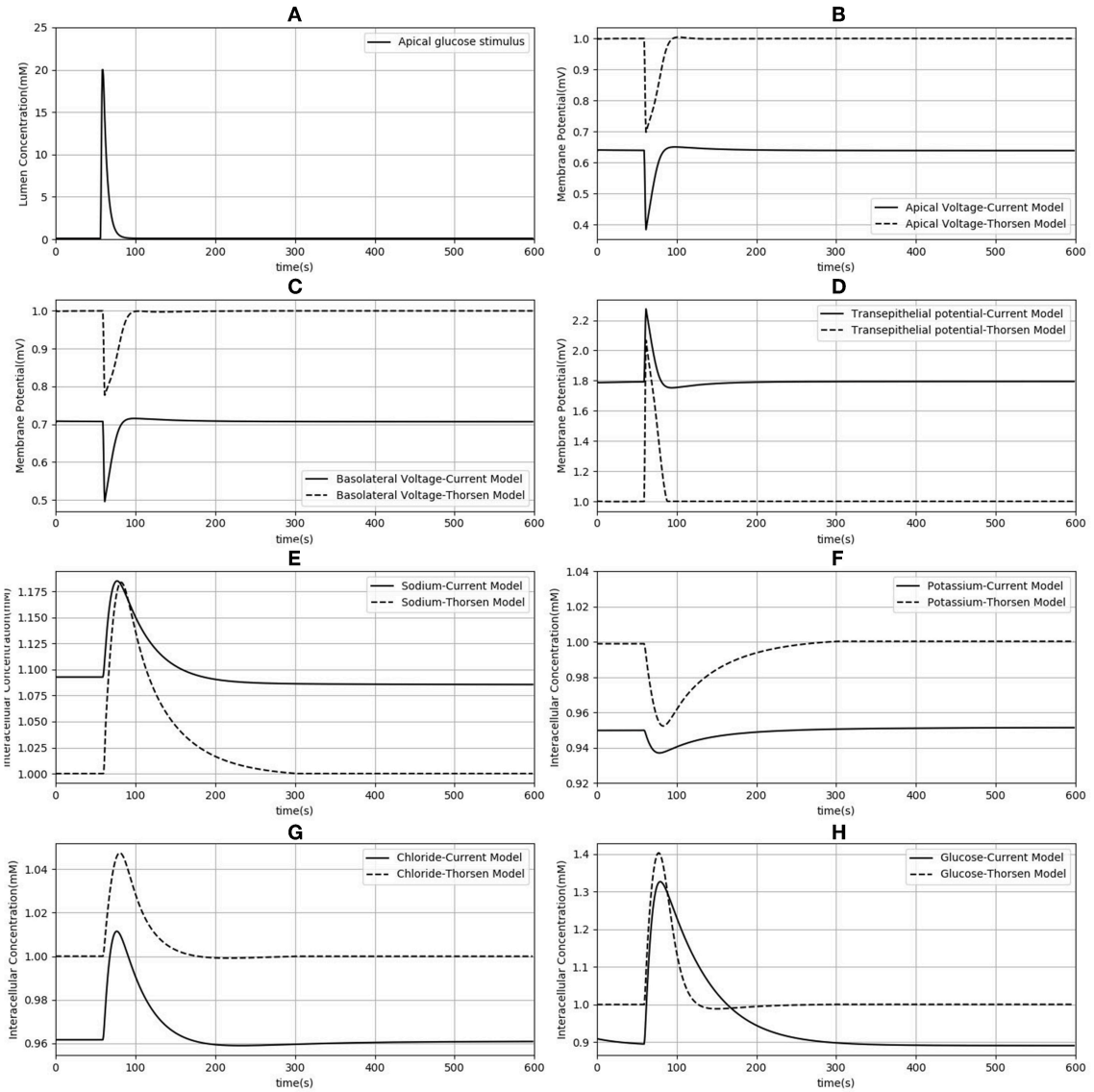


FIGURE 4 | Comparison of model responses against the model of Thorsen et al. (2014). Each variable has been normalized against the corresponding steady value from the Thorsen model. **(A)** Apical glucose stimulus. **(B,C)** Apical and basolateral membrane potential respectively. **(D)** Transepithelial potential. **(E–H)** Sodium, potassium, chloride and glucose intracellular concentration.

to non-diabetic controls in both human and animal studies (Fedorak et al., 1991; Burant et al., 1994; Dyer et al., 1997, 2002). The surface area of the villi has also been reported to increase in diabetes (Schedl and Wilson, 1971). Together these factors are expected to lead to higher rates of glucose absorption to the blood. However, the magnitude of the effect is not known. We used our developed model to study the effect of a 3-fold elevated SGLT1 and GLUT2 expression levels on glucose flux into the basolateral compartment. **Figure 8** shows the ratio of steady state glucose flux into the basolateral compartment, normalized to the flux at baseline conditions over a range of apical glucose concentrations. The increase in glucose absorption is less than the increase in transporter expression levels. For apical glucose concentrations up to 50 mM, 3-fold increase in SGLT1 levels causes a small increase in the basolateral flux over the whole range of apical glucose concentration. This increase is <1.1 times the baseline value. On the other hand increasing the level of GLUT2 by 3-fold increases the basolateral flux to almost 3 times the baseline value. This increase is observed over the whole range of glucose concentration. The result shows that higher levels of GLUT2 in diabetics may lead to a proportional increase in glucose absorption. In contrast, increases in SGLT1 cause a much smaller increase in absorption. However, SGLT1 may indirectly increase absorption rates since studies have shown that apical GLUT2 expression is dependent on SGLT1 activity (Kellett and Helliwell, 2000).

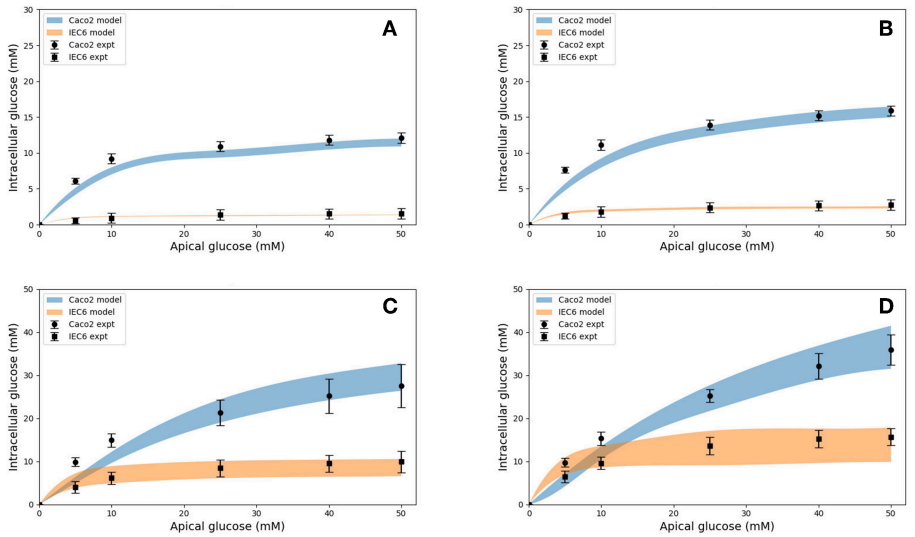


FIGURE 5 | Intracellular glucose concentrations for a range of extracellular glucose concentrations in Caco2 and IEC6 cells and exposure times (**A**: 30 s, **B**: 60 s, **C**: 300 s, **D**: 600 s). Experimental data points and error bars were digitally extracted from Zheng et al. (2012). Strips for the model predictions represent the range of values generated by setting $V_b = mV_c, m = 0.1, 1, 10, 100, \infty$.

TABLE 4 | Best fit values of the protein density (*a*) used to generate the simulated curves in **Figure 5** for different exposure times and both cell lines.

Exposure duration (seconds)	*a* (g protein/ml) Caco2	*a* (g protein/ml) IEC6
30	0.021	0.009
60	0.026	0.013
300	0.032	0.035
600	0.03	0.047

4. DISCUSSION

We have developed a computational model of glucose transport in the enterocyte that includes the full set of relevant transporters. The model is able to reproduce measurements reported in the literature and can be used to answer physiologically relevant questions about glucose uptake rates and mechanisms. In addition, the capabilities of the CellML framework were exploited to compose existing validated models of individual transporters to create the final model, which provides greater confidence in the implementation and facilitates model reuse and sharing.

4.1. Comparison With Existing Models

Our model differs from the Thorsen et al. (2014) model in some important respects.

One of the differences between the two models is in the treatment of sodium and chloride transport at the apical membrane. Thorsen et al. postulate electro neutral one-for-one fluxes of these ions to account for the sodium-hydrogen (NHE3) and chloride-bicarbonate (AE1) exchangers and use Goldman-Hodgkin-Katz (GHK) diffusion to model ENaC and CFTR. In contrast, our model takes a more general approach by incorporating the individual transport pathways at the apical membrane (**Figure 1**). We examined the implications of these modeling choices in **Figure 9**. **Figure 9A** shows the ratio of the AE1 flux to NHE3 flux for the simulation conditions of **Figure 4**. In the Thorsen model this ratio is equal to 1, whereas the ratio lies in the range 7–8 in our model. Our decision to explicitly model AE1 and NHE3 offers some advantages and testable consequences. First, our model produces the intracellular pH as an output since H+ concentration is a variable in the model and this provides an additional consistency check. Second, our model can be used to investigate conditions in which the expression/function of AE1 and NHE3 are altered, e.g., impaired absorption in NHE3 knockout mice (Schultheis et al., 1998), reduced chloride absorption and pH imbalance in AE1 mutations (Noonan et al., 2005).

Thorsen et al. used sodium and chloride diffusion through both apical and basolateral membrane of the cell. We replaced them with ENaC and CFTR transporters for sodium and chloride flux in the apical membrane, respectively. **Figure 9B** shows the ratio of sodium and chloride flux through transporters in our model to the sodium and chloride flux through diffusion in the Thorsen model. It is seen that Chloride flux via CFTR is around 4 times higher than Cl^- diffusion and also sodium via ENaC has around 2 times higher flux compared to Na^+ diffusion in Thorsen model. Thus, the contributions of individual transport pathways are significantly different between the models while still providing similar steady state predictions (**Figure 4**). By incorporating individual transporters our model offers the flexibility to study effects of drugs or diseases that influence the function of these transporters.

4.2. Parameter Choice and Data Fitting

Published values from the literature were used for the majority of transport protein kinetic parameters in our model, with a great

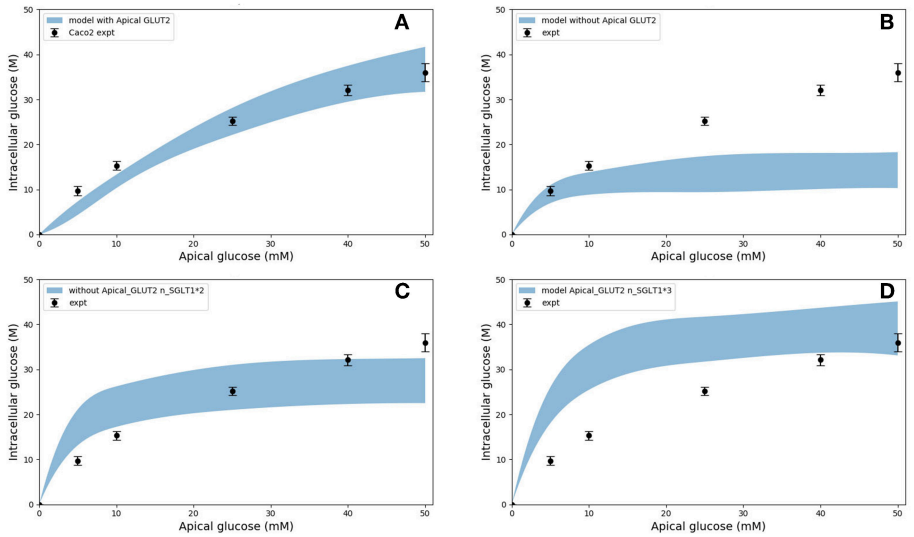


FIGURE 6 | Intracellular glucose concentration vs. extracellular glucose concentration in Caco2 in the presence/absence of Apical GLUT2 with different number of SGLT1 transporter **(A)** Output of model with apical GLUT2 **(B)** Model does not have apical GLUT2 **(C)** model does not have apical GLUT2 and the number of SGLT1 is doubled **(D)** model does not have apical GLUT2 and the number of SGLT1 is 3-fold higher. Experimental data points and error bars were digitally extracted from Zheng et al. (2012). Strips for the model predictions represent the range of values generated by setting $V_b = mV_c, m = 0.1, 1, 10, 100, \infty$.

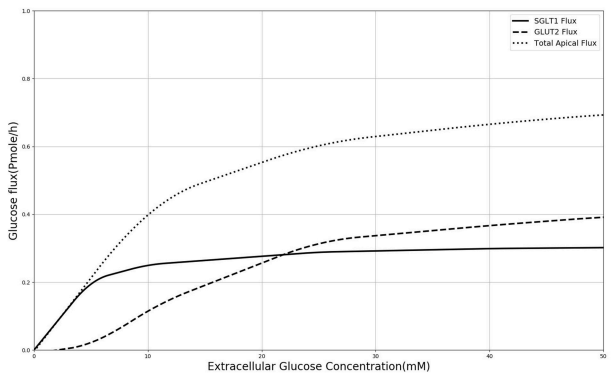


FIGURE 7 | Glucose Flux through SGLT1 and GLUT2 in 600 s of simulation along with total apical flux of glucose.

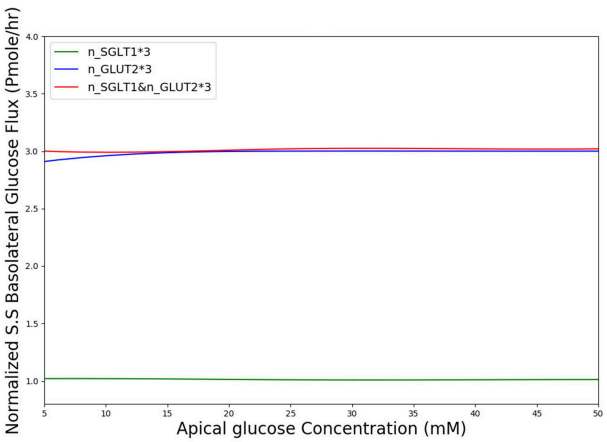


FIGURE 8 | Normalized steady state basolateral glucose flux vs. different stimulus of glucose in the lumen when number of SGLT1 is 3-fold higher(green), number of GLUT2 is 3-fold higher(blue) and number of both SGLT1 & GLUT2 are 3-fold higher(red).

deal of information for ohmic models provided by Fong et al. (2016). However, we used a different number of transporters in order to obtain a better fit to the experimental data. These values are shown in the **Supplementary Material**. The total cellular protein density (a) was used as a fitting parameter to match the model predictions of intracellular glucose with the data of Zheng et al. (2012). For both cell types, the fitted value increased with the duration of glucose exposure (**Table 4**). Since $c_{pred} = a\ c_{expt}$, this indicates that measured uptake increased at a slower rate with exposure time than predicted by the model. Possible reasons could include desensitization or inactivation of transporters and variations in cell protein density between different experiments. Also, the fitted protein densities are lower than indicative values for the mammalian cells (0.1–0.2 g/ml, Milo, 2013). The actual values of a do not hold much significance since they depend on the cell volume, which were not estimated in the experiments and were assigned arbitrary, realistic values (volume = $1400\mu m^3$) (Buschmann and Manke, 1981; MacLeod et al., 1991; Crowe and Marsh, 1993). Given these caveats, the model produces reasonable fits without the requirement of fine tuning.

It was also necessary to make an assumption about the volume of the basolateral compartment in the comparisons with Zheng et al. (2012) as explained before. Rather than treat this as a fitting parameter, we generated a range of model predictions by

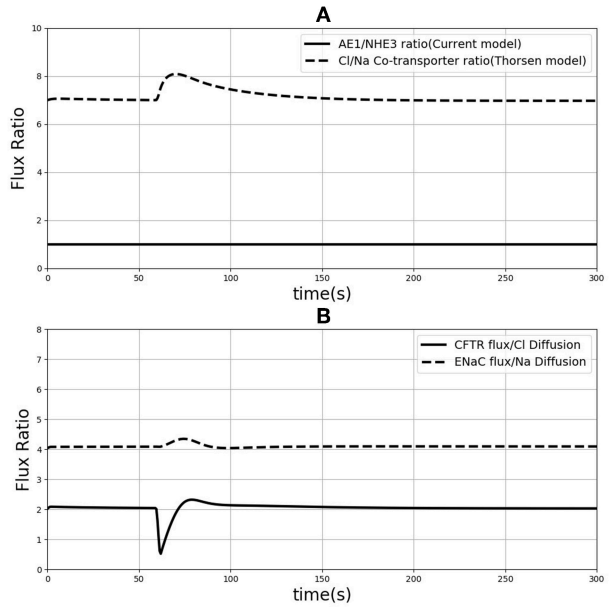


FIGURE 9 | (A) Ratio of sodium flux and Chloride flux through NHE3 and AE1 compare to NaCl co-transporter flux in the Thorsen model. **(B)** Chloride flux through CFTR and chloride diffusion ratio compare to the ration of sodium flux through ENaC and sodium diffusion.

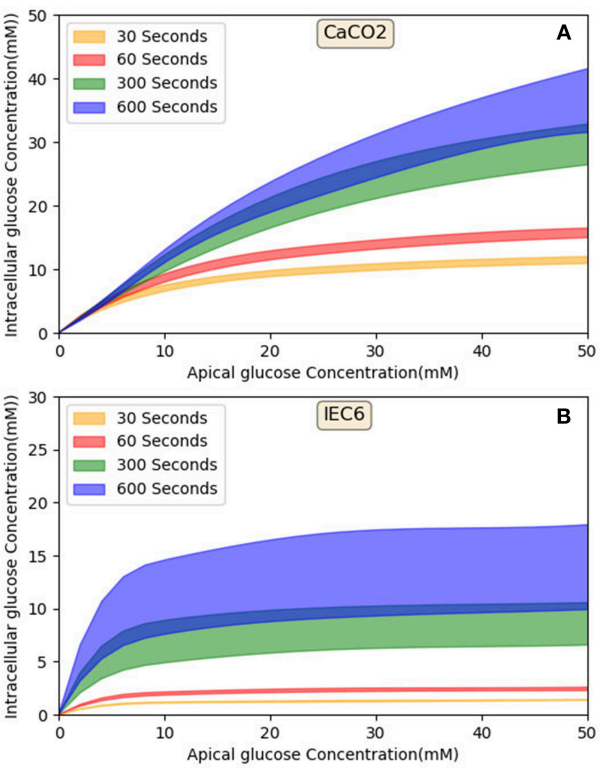


FIGURE 10 | Model output—Intracellular glucose concentration vs. Extracellular glucose concentration in **(A)** Caco2 cell line and **(B)** IEC6 cell line.

varying the parameter from small (0.1 times the cell volume) to large (an infinite compartment) values. The model predictions varied by <5% at short exposure durations and about 50% at long durations (**Figure 5**) and bracketed the experimental observations in all cases, except few shortest exposures for Caco2. This once again points to the robustness of the model predictions.

4.3. Role of Apical GLUT2 in Glucose Uptake and Effect of Time

Figure 10 shows that in both cell lines at short exposure times the glucose uptake has a tendency to be saturated (in 30 and 60 s), in longer term (>300 s) Caco2 shows non-saturation glucose uptake (**Figure 10A**) however IEC6 has a greater tendency to level off even at higher apical glucose concentration (**Figure 10B**). It has been reported that increasing the glucose concentration in the lumen can cause the apical translocation of GLUT2 (Scow et al., 2011) however, in our model results do not require acute translocation of GLUT2 to the apical membrane. Also Western blots experiments showed higher level of GLUT2 expression in higher extracellular glucose concentration (Kellett and Brot-Laroche, 2005); however, in our model density of apical GLUT2 was the same in different concentrations and exposure times. This shows that apical GLUT2 is highly crucial in order to explain the behavior of intracellular glucose absorption.

According to **Figure 8**, in diabetic patients GLUT2 plays much more important role in the increased glucose absorption compared to SGLT1 regarding the number of transporters. This is in fact a very interesting finding which could be a potential subject of future research into the role of glucose transporter expression levels in diabetic patients.

In summary, we have developed an integrative model of glucose uptake in the enterocyte that incorporates mechanistic descriptions of all relevant transporters and validated it against published measurements and models with minimal parameter tuning. The work utilizes the CellML modeling framework and the Physiome Model Repository to provide a portable, publically available implementation that facilitates sharing, reuse and extension of the model. We expect that the model will provide insight into transport pathways and guide the design and interpretation of experiments to generate and test hypotheses. We have used the model to determine the relative contribution of SGLT1 and GLUT2 to glucose absorption under a range of conditions. We have also evaluated the consequences of altered SGLT1 and GLUT2 expression in diabetes on glucose absorption rates. Potential applications in the future can include predictive modeling of the effect of drugs such as SGLT1 and GLUT2 inhibitors on glucose uptake and ion transport. This model of cellular uptake can be coupled with models of blood flow and metabolism to develop a more complete predictive framework of glucose homeostasis in the body (Nickerson et al., 2015).

AUTHOR CONTRIBUTIONS

NA, SS, DPN, PJH, and VS contributed conception and design of the study. NA performed the statistical analysis and modeling, wrote the first draft of the manuscript along with sections of the

manuscript. VS and SS checked the model and validation. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

NA was funded by a doctoral scholarship from the Riddet Centre of Research Excellence, one of ten currently funded by the New Zealand Tertiary Education Commission, Funding from RSNZ Marsden (contract UOA1411) is acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2019.00380/full#supplementary-material

REFERENCES

Barrett, K. E. and Keely, S. J. (2015). "Electrolyte secretion and absorption in the small intestine and colon," in *Yamada's Textbook of Gastroenterology*, eds D. K. Podolsky, M. Camilleri, J. G. Fitz, A. N. Kalloo, F. Shanahan, and T. C. Wang (New Jersey, NJ: John Wiley & Sons, Ltd.), 420–449. doi: 10.1002/9781444303254.ch14

Burant, C. F., Flink, S., DePaoli, A. M., Chen, J., Lee, W. S., Hediger, M. A., et al. (1994). Small intestine hexose transport in experimental diabetes. Increased transporter mrna and protein expression in enterocytes. *J. Clin. Invest.* 93, 578–585.

Buschmann, R. J. and Manke, D. J. (1981). Morphometric analysis of the membranes and organelles of small intestinal enterocytes. i. fasted hamster. *J. Ultra. Res.* 76, 1–14. doi: 10.1016/S0022-5320(81)80046-9

Crowe, P. T. and Marsh, M. N. (1993). Morphometric analysis of small intestinal mucosa iv. determining cell volumes. *Virchows Archiv. A* 422, 459–466. doi: 10.1007/BF01606454

Cuellar, A. A., Lloyd, C. M., Nielsen, P. F., Bullivant, D. P., Nickerson, D. P., and Hunter, P. J. (2003). An overview of cellml 1.1, a biological model description language. *Simulation* 79, 740–747. doi: 10.1177/0037549703040939

Dyer, J., Garner, A., Wood, I., Sharma, A., Chandranath, I., and Shirazi-Beechey, S. (1997). Changes in the levels of intestinal na+/glucose co-transporter (sglt1) in experimental diabetes. *Biochem. Soc. Trans.* 25:479S.

Dyer, J., Wood, I., Palejwala, A., Ellis, A., and Shirazi-Beechey, S. (2002). Expression of monosaccharide transporters in intestine of diabetic humans. *Am. J. Physiol. Gastroint. Liver Physiol.* 282, G241–G248. doi: 10.1152/ajpgi.00310.2001

Fedorak, R. N., Cheeseman, C. I., Thomson, A., and Porter, V. M. (1991). Altered glucose carrier expression: mechanism of intestinal adaptation during streptozocin-induced diabetes in rats. *Am. J. Physiol. Gastroint. Liver Physiol.* 261, G585–G591. doi: 10.1152/ajpgi.1991.261.4.G585

Fong, S., Chiorini, J. A., Sneyd, J., and Suresh, V. (2016). Computational modeling of epithelial fluid and ion transport in the parotid duct after transfection of human aquaporin-1. *Am. J. Physiol. Gastroint. Liver Physiol.* 312, G153–G163. doi: 10.1152/ajpgi.00374.2016

Frizzell, R. A., Nellans, H. N., Rose, R. C., Markscheid-Kaspi, L., and Schultz, S. G. (1973). Intracellular cl concentrations and influxes across the brush border of rabbit ileum. *Am. J. Physiol. Legacy Content* 224, 328–337.

Garny, A. and Hunter, P. J. (2015). Opencor: a modular and interoperable approach to computational biology. *Front. Physiol.* 6:26. doi: 10.3389/fphys.2015.00026

Gorboulev, V., Schürmann, A., Vallon, V., Kipp, H., Jaschke, A., Klessen, D., et al. (2012). Na+-d-glucose cotransporter sglt1 is pivotal for intestinal glucose absorption and glucose-dependent incretin secretion. *Diabetes* 61, 187–196. doi: 10.2337/db11-1029

Hunter, P. J. and Borg, T. K. (2003). Integration from proteins to organs: the physiome project. *Nat. Rev. Mol. Cell Biol.* 4:237. doi: 10.1038/nrm1054

Ingalls, B. (2012). Mathematical modelling in systems biology: an introduction. *Appl. Math. Univ. Waterloo* 396, 1–15.

Kellett, G. L. (2012). Comment on: Gorboulev et al. na+-d-glucose cotransporter sglt1 is pivotal for intestinal glucose absorption and glucose-dependent incretin secretion. *Diabetes* 61, e4–e5. doi: 10.2337/db11-1793

Kellett, G. L. and Brot-Laroche, E. (2005). Apical glut2: a major pathway of intestinal sugar absorption. *Diabetes* 54, 3056–3062. doi: 10.2337/diabetes.54.10.3056

Kellett, G. L. and Helliwell, P. A. (2000). The diffusive component of intestinal glucose absorption is mediated by the glucose-induced recruitment of glut2 to the brush-border membrane. *Biochem. J.* 350, 155–162. doi: 10.1042/bj3500155

Koepsell, H. and Gorboulev, V. (2012). Response to comment on: Gorboulev et al. na+-d-glucose cotransporter sglt1 is pivotal for intestinal glucose absorption and glucose-dependent incretin secretion. *Diabetes* 61, e5–e5. doi: 10.2337/db12-0061

König, M., Bulik, S., and Holzhütter, H.-G. (2012). Quantifying the contribution of the liver to glucose homeostasis: a detailed kinetic model of human hepatic glucose metabolism. *PLoS Comput. Biol.* 8:e1002577. doi: 10.1371/journal.pcbi.1002577

MacLeod, R. J., Hamilton, J., Bateman, A., Belcourt, D., Hu, J., Bennett, H., and Solomon, S. (1991). Corticostatic peptides cause nifedipine-sensitive volume reduction in jejunal villus enterocytes. *Proc. Natl. Acad. Sci. U.S.A.* 88, 552–556. doi: 10.1073/pnas.88.2.552

Milo, R. (2013). What is the total number of protein molecules per cell volume? a call to rethink some published values. *Bioessays* 35, 1050–1055. doi: 10.1002/bies.201300066

Nellans, H. N., Frizzell, R. A., and Schultz, S. G. (1973). Coupled sodium-chloride influx across the brush border of rabbit ileum. *Am. J. Physiol. Legacy Content* 225, 467–475.

Nellans, H. N., and Schultz, S. G. (1976). Relations among transepithelial sodium transport, potassium exchange, and cell volume in rabbit ileum. *J. Gen. Physiol.* 68, 441–463.

Nickerson, D. P., Ladd, D., Hussan, J. R., Safaei, S., Suresh, V., Hunter, P. J., et al. (2015). Using cellml with opencmiss to simulate multi-scale physiology. *Front. Bioeng. Biotechnol.* 2:79. doi: 10.3389/fbioe.2014.00079

Noonan, W. T., Woo, A. L., Nieman, M. L., Prasad, V., Schultheis, P. J., Shull, G. E., et al. (2005). Blood pressure maintenance in nhe3-deficient mice withtransgenic expression of nhe3 in small intestine. *Am. J. Physiol. Regul. Integrat. Compar. Physiol.* 288, R685–R691. doi: 10.1152/ajpregu.00209.2004

Okada, Y., Irimajiri, A., and Inouye, A. (1976). Intracellular ion concentrations of epithelial cells in rat small intestine effects of external potassium ions and uphill transports of glucose and glycine. *Jap. J. Physiol.* 26, 427–440.

Østby, I., Øyehaug, L., Einevoll, G. T., Nagelhus, E. A., Plahte, E., Zeuthen, T., et al. (2009). Astrocytic mechanisms explaining neural-activity-induced shrinkage of extraneuronal space. *PLoS Comput Biol.* 5:e1000272. doi: 10.1371/journal.pcbi.1000272

Palk, L., Sneyd, J., Shuttleworth, T. J., Yule, D. I., and Crampin, E. J. (2010). A dynamic model of saliva secretion. *J. Theor. Biol.* 266, 625–640. doi: 10.1016/j.jtbi.2010.06.027

Palumbo, P., Ditlevsen, S., Bertuzzi, A., and De Gaetano, A. (2013). Mathematical modeling of the glucose–insulin system: a review. *Math. Biosci.* 244, 69–81. doi: 10.1016/j.mbs.2013.05.006

Parent, L., Supplisson, S., Loo, D. D., and Wright, E. M. (1992). Electrogenic properties of the cloned na+/glucose cotransporter: Ii. a transport model under nonrapid equilibrium conditions. *J. Membr. Biol.* 125, 63–79.

Pradhan, R. K., Vinnakota, K. C., Beard, D. A., and Dash, R. K. (2013). *Chapter 5: Carrier-mediated Transport Through Biomembranes* (Elsevier).

Riz, M., and Pedersen, M. G. (2015). Mathematical modeling of interacting glucose-sensing mechanisms and electrical activity underlying glucagon-like peptide 1 secretion. *PLoS Comput. Biol.* 11:e1004600. doi: 10.1371/journal.pcbi.1004600

Röder, P. V., Geillinger, K. E., Zietek, T. S., Thorens, B., Koepsell, H., and Daniel, H. (2014). The role of sglt1 and glut2 in intestinal glucose transport and sensing. *PLoS ONE* 9:e89977. doi: 10.1371/journal.pone.0089977

Rose, R. C., and Schultz, S. G. (1971). Studies on the electrical potential profile across rabbit ileum. *J. Gen. Physiol.* 57, 639–663. doi: 10.1085/jgp.57.6.639

Schedl, H. P., and Wilson, H. D. (1971). Effects of diabetes on intestinal growth in the rat. *J. Exp. Zool.* 176, 487–495. doi: 10.1002/jez.1401760410

Schultheis, P. J., Clarke, L. L., Meneton, P., Miller, M. L., Soleimani, M., Gawenis, L. R., et al. (1998). Renal and intestinal absorptive defects in mice lacking the nhe3 na+/h+ exchanger. *Nat. Genet.* 19:282. doi: 10.1038/969

Scow, J. S., Tavakkolizadeh, A., Zheng, Y., and Sarr, M. G. (2011). Acute "adaptation" by the small intestinal enterocyte: a posttranscriptional mechanism involving apical translocation of nutrient transporters. *Surgery* 149, 601–605. doi: 10.1016/j.surg.2011.02.001

Shimada, T., and Hoshi, T. (1987). Role of na+/h+ antiport in intracellular ph regulation by rabbit enterocytes. *Biochim. Biophys. Acta* 901, 265–272. doi: 10.1016/0005-2736(87)90123-4

Thorsen, K., Drengstig, T., and Ruoff, P. (2014). Transepithelial glucose transport and na+/k+ homeostasis in enterocytes: an integrative model. *Am. J. Physiol. Cell Physiol.* 307, C320–C337. doi: 10.1152/ajpcell.00068.2013

Vogalis, F. (2000). Potassium channels in gastrointestinal smooth muscle. *J. Auton. Pharmacol.* 20, 207–219.

Weinstein, A. M. (1995). A kinetically defined na+/h+ antiporter within a mathematical model of the rat proximal tubule. *J. Gen. Physiol.* 105, 617–641.

Weinstein, A. M. (2000). A mathematical model of the outer medullary collecting duct of the rat. *Am. J. Physiol. Renal Physiol.* 279, F24–F45. doi: 10.1152/ajprenal.2000.279.1.F24

Weinstein, A. M. (2015). A mathematical model of the rat nephron: glucose transport. *Am. J. Physiol. Renal Physiol.* 308, F1098–F1118. doi: 10.1152/ajprenal.00505.2014

Yu, T., Lloyd, C. M., Nickerson, D. P., Cooling, M. T., Miller, A. K., Garny, A., et al. (2011). The physiome model repository 2. *Bioinformatics* 27, 743–744. doi: 10.1093/bioinformatics/btq723

Zheng, Y., Scow, J. S., Duenes, J. A., and Sarr, M. G. (2012). Mechanisms of glucose uptake in intestinal cell lines: role of glut2. *Surgery* 151, 13–25. doi: 10.1016/j.surg.2011.07.010

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

frontiers in Genetics

TECHNOLOGY REPORT
published: 14 May 2019
doi: 10.3389/fgene.2019.00356

Simplicity DiffExpress: A Bespoke Cloud-Based Interface for RNA-seq Differential Expression Modeling and Analysis

Cintia C. Palu[1,2], Marcelo Ribeiro-Alves[3], Yanxin Wu[2,4], Brendan Lawlor[2,4], Pavel V. Baranov[1,5], Brian Kelly[2] and Paul Walsh[2,4]**

[1] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland, [2] NSilico Life Science Ltd., Cork, Ireland, [3] Laboratory of Clinical Research on STD/AIDS, National Institute of Infectology Evandro Chagas (INI) – Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [4] Cork Institute of Technology, Cork, Ireland, [5] Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia

OPEN ACCESS

Edited by:
Helder Nakaya,
University of São Paulo, Brazil

Reviewed by:
Qiong-Yi Zhao,
The University of Queensland, Australia
Daniel Paul Heruth,
Children's Mercy Hospital, United States

***Correspondence:**
Cintia C. Palu
cintia.palu@nsilico.com
Paul Walsh
paul.walsh@nsilico.com

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

Received: 04 December 2018
Accepted: 02 April 2019
Published: 14 May 2019

Citation:
Palu CC, Ribeiro-Alves M, Wu Y, Lawlor B, Baranov PV, Kelly B and Walsh P (2019) Simplicity DiffExpress: A Bespoke Cloud-Based Interface for RNA-seq Differential Expression Modeling and Analysis. Front. Genet. 10:356. doi: 10.3389/fgene.2019.00356

One of the key challenges for transcriptomics-based research is not only the processing of large data but also modeling the complexity of features that are sources of variation across samples, which is required for an accurate statistical analysis. Therefore, our goal is to foster access for wet lab researchers to bioinformatics tools, in order to enhance their ability to explore biological aspects and validate hypotheses with robust analysis. In this context, user-friendly interfaces can enable researchers to apply computational biology methods without requiring bioinformatics expertise. Such bespoke platforms can improve the quality of the findings by allowing the researcher to freely explore the data and test a new hypothesis with independence. *Simplicity DiffExpress* is a data-driven software platform dedicated to enabling non-bioinformaticians to take ownership of the differential expression analysis (DEA) step in a transcriptomics experiment while presenting the results in a comprehensible layout, which supports an efficient results exploration, information storage, and reproducibility. *Simplicity DiffExpress'* key component is the bespoke statistical model validation that guides the user through any necessary alteration in the dataset or model, tackling the challenges behind complex data analysis. The software utilizes *edgeR*, and it is implemented as part of the Simplicity™ platform, providing a dynamic interface, with well-organized results that are easy to navigate and are shareable. Computational biologists and bioinformaticians can also benefit from its use since the data validation is more informative than the usual DEA resources. Wet-lab collaborators can benefit from receiving their results in an organized interface. *Simplicity DiffExpress* is freely available for academic use, and it is cloud-based (https://simplicity.nsilico.com/dea).

Keywords: differential expression analysis, differential gene expression, statistical modeling, edgeR, transcriptomics, RNA-seq, data-driven

INTRODUCTION

OMICs techniques open the doors to researching organisms from a comprehensive perspective, enabling the exploration of the intricate network of relationships, as opposed to analyzing point biological variations. The scaling up of the analysis enabled the understanding of many of the molecular aspects behind an organism's features, while at the same time revealed that the mechanisms involved in the control of transcription, translation and the organism physiology, in general, are more complex than once thought. Therefore, it is no surprise that OMICs research requires the effort of multi-disciplinary teams and it is quite common to see a publication with more than ten co-authors. From this new perspective, many challenges arise, one being the knowledge transfer and communication between professionals with different backgrounds. Other challenges are well explored, such as analysis and storage of complex data, with every new technique for high-throughput molecular biology research requiring new methods for data analysis and interpretation (Finotello and Di Camillo, 2015; Han et al., 2015; Yuryev, 2015; Byron et al., 2016; Conesa et al., 2016).

Among the OMICs techniques, transcriptomics rapidly became a popular methodology for profiling gene expression through RNA-seq (Nagalakshmi et al., 2008). Transcriptomics can be applied to the analysis of messenger RNAs, non-coding RNAs (such as long non-coding RNAs, microRNAs, and transfer RNAs), the investigation of mRNA isoforms and can be combined with other methods to enhance analysis (Byron et al., 2016; Conesa et al., 2016). Originally, RNA-seq techniques were developed for sequences from pooled cells, which is known as "bulk RNA-seq." Later on, single-cell RNA-seq methods were developed, requiring not only new laboratory procedures but also the development of novel approaches to process and analyze the data (Tang et al., 2009).

The relative abundance of the set of RNAs found in a sample reflects the level of expression of the corresponding genes, indicating the cells' state and the aspects involved in the determination of a certain condition (Finotello and Di Camillo, 2015). The objective of DEA is to identify the mRNAs (or other transcribed sequences) that have changed significantly in abundance across treatment groups in an experiment. A typical DEA based workflow firstly requires the mapping of the sequenced reads of each sample to a reference genome or a transcriptome (when available). The following step is the estimation of how many reads matched to different loci or transcripts, the organization of the retrieved information in the "read-count table," and finally the completion of the necessary corrections such as distribution and coverage normalization. All of the above steps must be done using quality checkpoints, and the analysis strategies may vary according to the organism being studied and the research objective (Oshlack et al., 2010; Conesa et al., 2016).

Summarizing the sequenced data into a read-count table presents important challenges and, on top of that, only precise and powerful tests can efficiently detect the differential expression (Oshlack et al., 2010; Finotello and Di Camillo, 2015; Han et al., 2015). Regardless of the challenges involved in the read-mapping step to generate the read-count table, it was shown that most tools that run this step perform equally (Costa-Silva et al., 2017). On the other hand, the methods applied to DEA have the greatest influence on the final results, and no current strategy offers optimum results (Costa-Silva et al., 2017). Therefore, the real challenge is to identify which transcripts are affected by the phenomena targeted by the research (treatment, cell types, etc.), among all the observed expression changes. Moreover, this is highly dependent on the accurate modeling of technical and biological variability (Finotello and Di Camillo, 2015).

It is undeniable that there is a heavy demand on the bioinformatics skills needed to process the high-throughput sequencing raw data files and the subsequent statistical skills to apply the methods that can uncover the relevant features in the data. A common mistake is to assume that the transcriptomics analysis ends with the list of genes differentially expressed, whereas it is more likely to lead to the next stage of the research. The research team still needs to explore the biological meaning behind the data analysis results, carry out gene set enrichment analysis or similar strategies, retrieve literature to support understanding the biological context and, ideally, test hypothesis by carrying out new wet-lab experiments (Han et al., 2015).

Simplicity DiffExpress tackles the statistical analysis steps that are required after the raw RNA-seq data is summarized in read-count tables. The main objective is to improve discovery by facilitating the statistical modeling of the DEA, with no programming skills required. It also offers an interactive interface with guided steps and the presentation of the results in a practical and shareable interface. These features are critical in the study of complex biological questions where multiple factors define the observed phenotype. *Simplicity DiffExpress* opens the doors to non-bioinformatician researchers to explore the data, and we believe it improves the discovery process by enabling the person who knows best about the biological aspects to be hands-on with the statistical analysis without an intermediary bioinformatician. Nonetheless, bioinformaticians benefit from the validation feedback and the practicality of results reproducibility that can be re-visited in any time-point and shared.

METHODS

Interface Implementation

The workflow was implemented as part of *Simplicity*™, a cloud-based software designed for supporting bioinformatics services to non-bioinformaticians (Walsh et al., 2013). Simplicity workflows' architecture is built using a combination of JavaScript, .NET, Java, and Python based components, which implement the UI, middleware, message queue, and storage (Azure Blob, **Figure 1**). *Simplicity's* UI is implemented in HTML5, JavaScript, and CSS. In the case of *DiffExpress* input, .NET also submits the data to R scripts in the middleware to run

Abbreviations: CPM, counts per million; DEA, differential expression analysis; GLM, generalized linear model; UI, user interface.

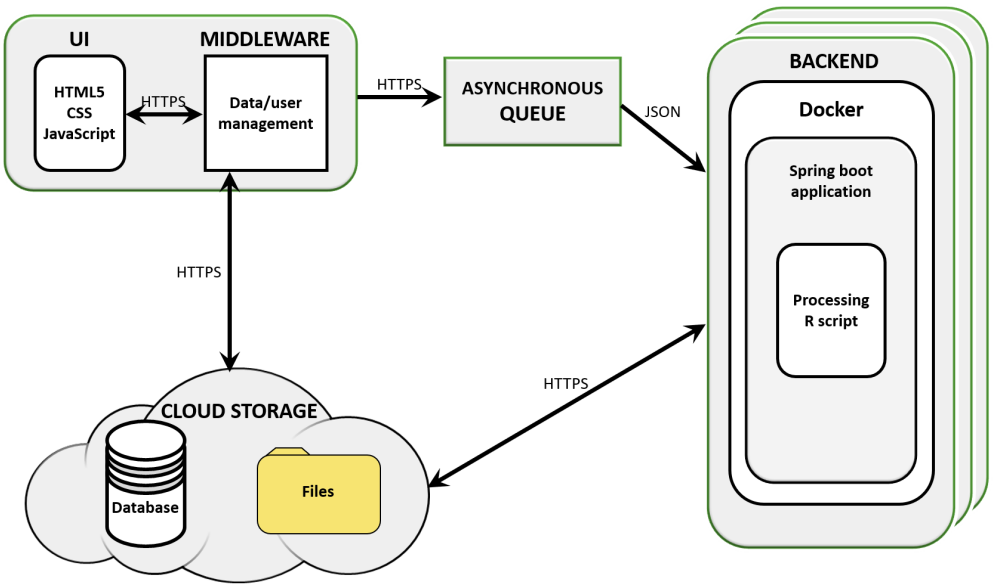


FIGURE 1 | *Simplicity*'s cloud-based structure. *Simplicity* is available on-line and presents a user interface (UI) that carries out real-time processing through its middleware storage and scripts. User requests are managed by an asynchronous queue and data is passed to cloud storage. Once a processing node is available, the backend retrieves the analysis parameters and data from the cloud storage and analyses it, delivering results to the user and compiling reports and interactive visualizations in the UI.

on-the-fly validations. The communication interface is made up of the queue system and storage elements. The queue system was developed in Java and Spring Boot and runs in a Docker container, controlling which jobs and data are sent to the backend to be processed. In the backend, an agent written in Java interacts with the queue, the storage and the service, which runs in a Docker container on a Linux server host. The resulting output files, when complete, are uploaded back to storage. During the processing, a JSON file containing information on the pipeline progress is constantly updated into the storage. Once this JSON file signals that the job was completed, an email is sent to the user. After he/she securely authenticates his/her login credentials in *Simplicity*, the user is granted access to the results interface, implemented using the same strategy as the input interface. The communication interface pulls the output results from the storage and presents them.

Data Validation

The first validation step is to check if the sample names on both read-count and metadata table match and to remove unwanted characters from the labels (implemented on .NET). Missing data ("NA") is retrieved using R scripts and is with dealt by removing samples or transcripts. The model fitness test is written in R and first evaluates if there are enough degrees of freedom, then, it applies QR decomposition to the statistical model design matrix to verify if it is full rank (all rows and columns are linearly independent) and, finally, checks if there are at least two samples for all the factor combinations generated by an interaction. The user is always informed of any detected issue and, when possible, offered an option on how to deal with it.

Differential Expression Analysis Implementation

The DEA of *DiffExpress* is fully implemented on Ubuntu 16.04.4 LTS, R version 3.5.2 (R Core Team, 2017), and based on *edgeR* version 3.22.5 (Robinson et al., 2010; McCarthy et al., 2012). *EdgeR* and *DESeq2* (Love et al., 2014) are among the best DEA performers (Finotello and Di Camillo, 2015), enabling multi-group comparisons (Oh et al., 2014), and, in our experience, *edgeR* offers the best approach to model complex data, therefore it was chosen to be the basis of our workflow. We use the library *jsonlite* version 1.6 (Ooms, 2014) to recover the analysis parameters passed as JSON files and *pheatmap* version 1.0.12 (Kolde, 2012) to generate heatmaps. The DEA scripts were initially tested on a dataset investigating changes in the modulation of rat small non-coding RNA due to exercise intensity, which required the modeling of a continuous variable (Oliveira et al., 2018). The environment information with the updated version of the libraries and programs used are presented on the results report, allowing the user keep track of upgrades done in the future.

Input Files

Simplicity DiffExpress requires two tables as input, which can be a CSV or TXT file. The interface provides options to set the parameters to read the files and on-the-fly visualization of how the data is being processed, enabling flexible input format. The current files' size is unlimited.

The first table to be uploaded is the read-count table which presents the raw read counts mapped to each genomic tag (genes). There are no requirements regarding transcript

IDs formats, although they must be presented in the first column of the file. The remaining columns should be numeric (with the sample name as heading). It is a requirement that the data is not transformed because *edgeR* automatically takes into account the total size (total read number) of each sample/library in all calculations of fold-changes, concentration, and statistical significance. In other words, RPKM, FPKM, and TPM -transformed data are not compatible (Robinson and Smyth, 2008).

The second table contains the metadata and must have (1) a row for each sample/library in the count table; (2) a column for each variable(s) of interest. *Simplicity DiffExpress* automatically removes any sample that is not present in both tables (the user receives a warning). The metadata table may contain any relevant information to understand the data, such as phenotypic features, clinical outcomes or experimental information (such as collection day, batch, institution). Later, the user will inform which of the information will be used in the statistical design, therefore there is no issue if the table contains variables beyond the ones that are intended to be used in the analysis.

Low-Count Filtering

A dataset usually has thousands of genomic features, and not all of them have enough reads to contribute to the DEA. In addition, these low counts may interfere with some of the statistical methods used in the pipeline. Therefore, it is strongly recommended to filter them out prior to further analysis. Nonetheless, the user can either opt to not filter out low counts or to decide what is the minimum CPM that a genomic feature must have in order to be kept in the analysis.

Normalization

In *Simplicity DiffExpress*, normalization is a mandatory step. The dataset is normalized for RNA composition by trimmed means of *M*-values (Robinson and Oshlack, 2010), which is the default methodology implemented on *edgeR*. The normalization step adjusts the RNA composition effect, avoiding the issue that the remaining genes falsely appear to be down-regulated in that sample/library.

Dispersion

The genomic features dispersion estimation is necessary so that it is consistent across replicates and in *Simplicity DiffExpress* it is based on the weighted likelihood empirical Bayes method (Robinson and Smyth, 2007). *Simplicity DiffExpress* uses *edgeR*'s Cox-Reid profile-adjusted likelihood method for all genomic features. It fits a GLM from an informed design matrix, allowing for all systematic sources of variation to be accounted for in the estimations (McCarthy et al., 2012; Chen et al., 2014; Oh et al., 2014). In addition, the user may decide whether the analysis should be robustified against potential outliers.

GLM Fitting

Once the above steps are completed, a Negative Binomial GLM is ready to be fitted to the dataset, as described by McCarthy et al. (2012). It conducts a gene-wise statistical test for a given coefficient or coefficient contrast of the variable(s) of interest.

Likelihood Ratio Test for the Selected Variables

This method is applied to test the ratio of deviances between nested models with and without the estimation of coefficients or coefficient contrast of the variable(s) of interest in the Negative Binomial-GLM model, respectively. It is at this *stage* of the analysis that genes differentially expressed between groups/conditions are actually identified and the gene-wise *p*-values are corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate method (Hochberg and Benjamini, 1990).

Case-Study

The public dataset GSE68086[1] for the use case, was originally published by Best et al. (2015). This dataset consists of RNA-seq data of 283 blood platelet samples obtained from 228 patients with six types of malignant tumor and 55 healthy donors. The large number of samples and the availability of metadata allows for a great data modeling opportunity. The statistical model used was "$\sim cancer + Metastasis + batch + Gender + Age$" for estimation of cancer and Metastasis effect, respectively. In both models, *batch*, *Gender*, and *Age* were included as confounders to minimize sample bias in the estimations of interest associated with variables *cancer* and *Metastasis* (**Supplementary Figure S1**).

Not all series on GEO are suitable for DiffExpress since there are assorted types of data that can be available. For GSE68086, the read-count table was available as **Supplementary File** and the metadata was obtained from the "Series Matrix" file. It was necessary to explore the "Series Matrix" file, select relevant information, such as the sample IDs that matched the read-count table, batch dates, cancer type, age, and gender. Some further formatting was done to remove the field names from the table cells (e.g., "*cancer type: BrCa*" became "*BrCa*"). The current series publicly available on GEO no longer provides information on *Age*, *Gender,* and *Metastasis*.

RESULTS

In this section, we do an overview of the features provided by the interface and present a case study using the public available dataset retrieved from GEO under the series identification GSE68086 (see text footnote 2), containing the RNA-seq data of 283 blood platelet samples obtained from 228 patients with six types of malignant tumor and 55 healthy donors (Best et al., 2015). The dataset size and metadata availability offered a great opportunity to test different statistical models. More information on the implementation and how to use *Simplicity DiffExpress* are provided in the documentation[2], tutorial page[3], and video[4].

Input Interface

Format-Flexible Data Input

The RNA-seq data must be processed and organized in a read-count table in order to be analyzed in *Simplicity DiffExpress*.

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68086

[2]https://simplicity.nsilico.com/Home/Document

[3]https://simplicity.nsilico.com/Home/Support

[4]https://www.youtube.com/watch?v=QKZu46c4HfU

The originally observed expression counts are required, so no data transformation is necessary prior to analysis, once the workflow will handle it subsequently (Robinson et al., 2010). A second table, called "metadata table," containing the features that characterize the samples must also be provided since it will be used to define the samples groups. The file upload buttons are the first features made available to the user once they access the platform (**Figure 2A**).

Simplicity DiffExpress was designed to deal with variable table formats, and the user can inspect how the software interprets their files on the go and can change the settings as required (**Figure 2B**). One of the key features of this platform is to allow the user to define the statistical design that best represents the experiment. *Simplicity DiffExpress* offers a form where the user should pick at least one of the features in the metadata table and fit a statistical model (**Figures 2C,D**). It is also recommended that the sources of bias are informed when defining the statistical model in the input interface. By including all known sources of undesired bias (e.g., batch effects) in the statistical model, the data analysis will take into consideration all those factors which will provide more precise estimations. **Supplementary Figure S1** exemplifies how the statistical model used in this work was set-up.

On-the-Fly Validation

In order to make the DEA on *Simplicity DiffExpress* more accessible, a graphical UI was implemented to provide clear and immediate feedback for the user. Therefore, the validation steps

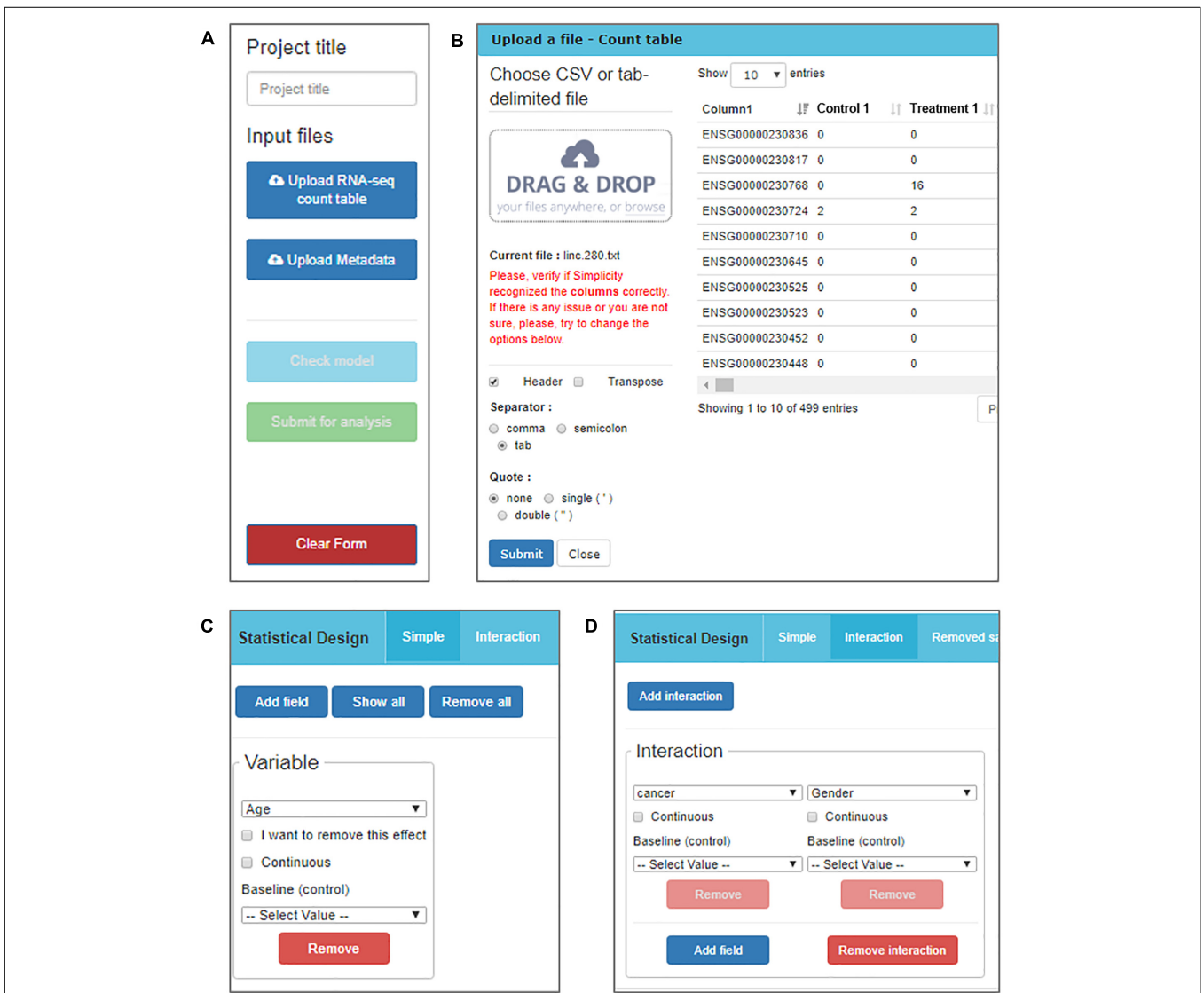


FIGURE 2 | *DiffExpress* input options. **(A)** The user is initially required to input a project title and upload two text files containing the read counts and metadata. **(B)** The upload window displays the table being uploaded in real-time allowing for verification if the interface is reading it correctly. The options on the left side can be altered to adjust the file-reading. **(C)** Once the tables are uploaded, a menu to include variables in the statistical model is enabled. The user should inform what is the baseline between the categories of a factorial variable or mark it as continuous. **(D)** It is also possible to study the interaction between two or more variables.

are key features of the software, to ensure that the data are adequate before submitting for the full analysis. The software informs the user if there are any missing value (represented by "NA"), any mismatch between samples IDs, and checks if it is possible to fit the model (described in "Methods"). This is performed by calling specialized validation R routines (R Core Team, 2017) in the *Simplicity DiffExpress* software validation module, implemented in the middleware. If any issue is identified, the interface presents possible solutions to the user (**Supplementary Figure S2A**) and specifies the variables that are causing it (**Supplementary Figure S2B**). In other words, the validation module secures the chances of a successful run of the DEA prior to submitting data to the server.

Other Features

All features in the input interface have a short user guide displayed on the bottom of the page, known as the "*Step Wizard*" (not shown). This area is designated to provide a brief overview on which options are available and which actions the user is expected to take. Moreover, some parameters can be customized, and they are made available at the "*Statistics*" tab (**Supplementary Figure S3A**).

Furthermore, when the validation procedure detects modeling issues, *Simplicity DiffExpress* will offer the option to resolve them by either removing a variable from the model or by eliminating some samples. It is important to highlight that the interface always restores the samples when changes in the statistical model resolve the issue. To allow the user to keep track of all the tested models and adjustments made, the interface provides the "*Removed samples*" tab (**Supplementary Figure S3B**), which is dedicated to specific details of samples that were eliminated from the current model and a "*History*" tab (**Supplementary Figure S3C**) that enlists all models tested.

Once all information is provided, and the statistical model has passed the fitness test ("*Check model*" button on **Figure 2A**), the analysis can be submitted. The user may finish the session at this point or choose to keep using the uploaded data, an option offered to facilitate the creation of new statistical models for the same dataset. Meanwhile, once the workflow management receives the job request, the analysis can take from a couple of minutes to a few hours, depending on the complexity of the statistical models and the dataset size. Once it is finished, the user will receive an email informing that the results are ready to be accessed.

Results Interface

The *Simplicity* system presents a list of all pipelines run by the user, highlighting the completion status and submission date (**Supplementary Figure S4**). The users may grant access to a pipeline to specific researchers of their choice, and this access can be revoked at any time by the user. Although the results are sharable, the invitees do not have access to the original input files. This feature favors collaborative work by enabling the whole team to explore the results in an organized presentation. Additionally, it supports reproducibility since all information regarding the analysis parameters is documented and stored at *Simplicity* and permanently linked to the pipeline.

Once a pipeline result is chosen, the user is brought to a new page enlisting all information regarding the DEA (**Figure 3**), including the chosen setting and analysis log. The log describes all the steps carried on the analysis and summarizes how many genes were found differentially expressed. It also provides access to the biological coefficient of variation plot and multidimensional scaling plots. The later plot presents the leading log-fold-change between each pair of samples and supports identifying structure and heterogeneity in the relative expression data (**Figure 4**). In the example presented here, it is possible to notice that there is a data structure due to batch. The buttons on the left (**Figure 3**) offer further functionalities, such as exploratory analysis of the results ("*Output Explorer*"), download all results, information on how to cite the *Simplicity DiffExpress* methods and the possibility to contact support.

By clicking on the "*Output Explorer*" button, the users have access to a window (**Figure 5A**) containing heatmaps and providing an overview of the data (**Figures 5B,D**) and a list with all comparisons between variables done in the DEA (**Figure 5C**). Once a comparison is chosen for further exploration, they are taken into a page where the results table is displayed (**Figure 5E**). In the case where more than two transcripts are differentially expressed, a specific MA plot (**Figure 5F**) and heatmaps are made available to enable an exploratory analysis of the results.

Simplicity DiffExpress will generate multiple tables with the DEA results. The number of tables depends on (1) the variables included in the statistical design; (2) the number of levels which each of the categorical/nominal variables has (e.g., in our case-study, variable *cancer* has seven levels: healthy donor and six cancer types); and, (3) if the user sets the program to carry out DEA between every level of the categorical/nominal variables or only contrasts the levels against the baseline. The researcher should interpret the differential expression significance based on the chosen false discovery rate; by default, it is set as 0.05.

Furthermore, it is recommended that the user follow the citation guidelines to ensure all credit is correctly presented; all information is available at the "*Citation and References*" button. Finally, results are restricted to the user and available upon login, and all images and tables can be saved locally through the button "*Download All Files*" (**Figure 3**).

DISCUSSION

The primary objective of *Simplicity DiffExpress* is to allow researchers with or without prior bioinformatics knowledge to create DEA models in order to study quantitative changes in gene expression levels between experimental groups. *Simplicity DiffExpress* achieves this through a user-friendly, intuitive, flexible and interactive cloud-based platform (**Figure 6**). The platform also provides clarity, real-time answers, and data validation. *Simplicity DiffExpress* is available at https://simplicity.nsilico.com/DEA, and it is free for academic use.

In the context of RNA-seq DEA, there are two major types of experimental designs: (1) pairwise group comparisons, where the samples were collected in a single time point and targets differences across two or more biological groups; and, (2)

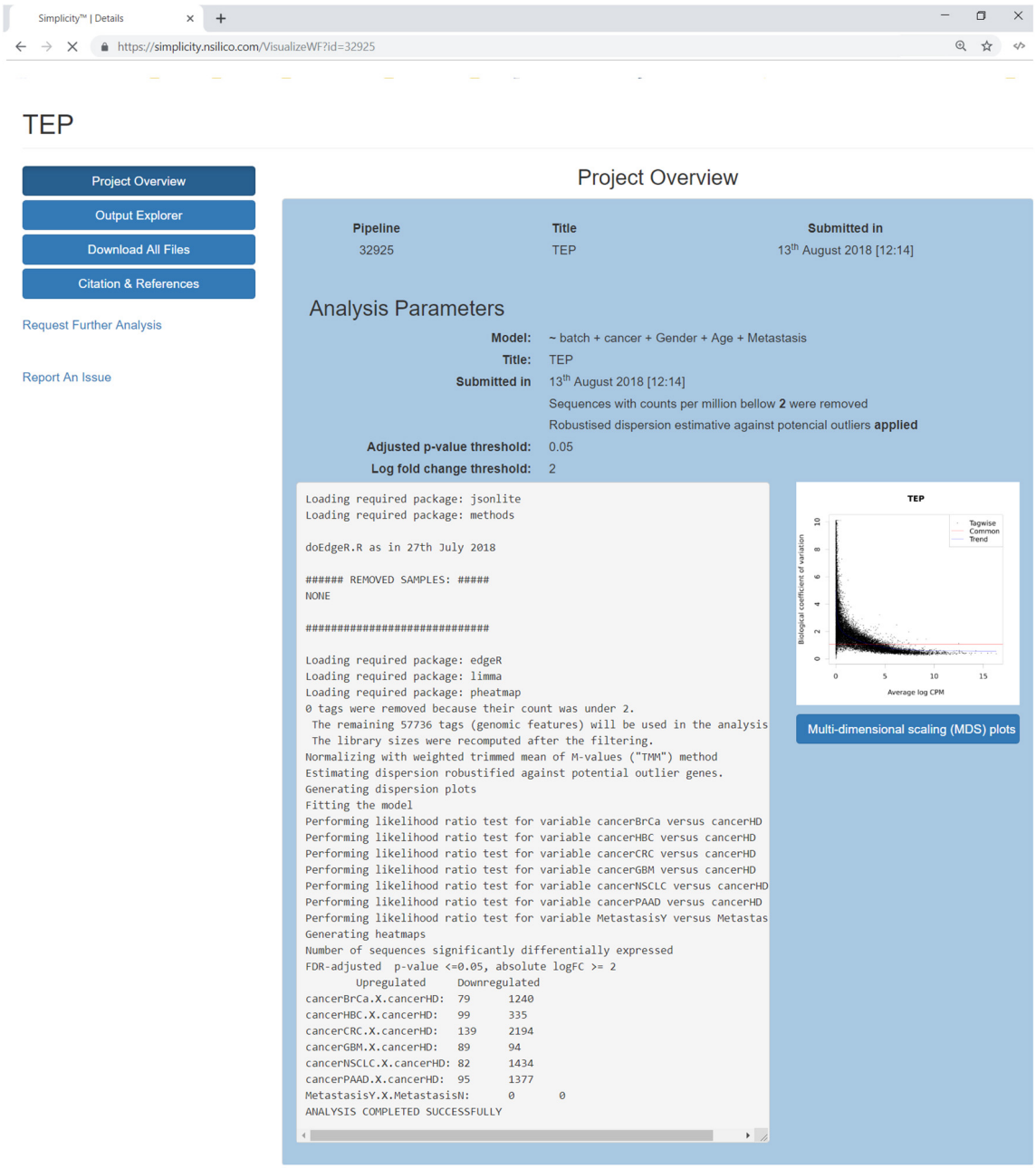


FIGURE 3 | Project's results overview interface. The blue buttons on the top left allow the user to navigate through the results features, it is also possible to report issues or request further analysis. The main panel reviews the analysis settings and its log, which lists all the steps done during the pipeline run and generates a summary of the genes found up and down-regulated. On the right, a dispersion plot showing the data distribution based on the biological coefficient of variation log CPM average. The biological coefficient of variation represents the coefficient of variation that would remain between biological replicates if sequencing depth could be increased indefinitely. Finally, the user can explore the generated multidimensional scaling plot clicking on the button on the right.

progression experiments, aiming to characterize the dynamics of a biological phenomenon (Oh et al., 2014). Time-series are the most common examples of the progression experiments, where the samples are collected in different points over a time window, but they can also relate to analyses of samples submitted to different intensities of interventions, such as drug dosages. Ideally, the experimental design should account for other sources of nuisances, such as different batches, age, sex, and replicates (Oh et al., 2014; Han et al., 2015). Controlling the sources of variation when designing and modeling correctly all these factors reflects directly in the capability of successfully identifying differentially expressed sequences. Therefore, it is critical to understand those variables, correctly identifying if they are continuous or categorical and how they relate to each

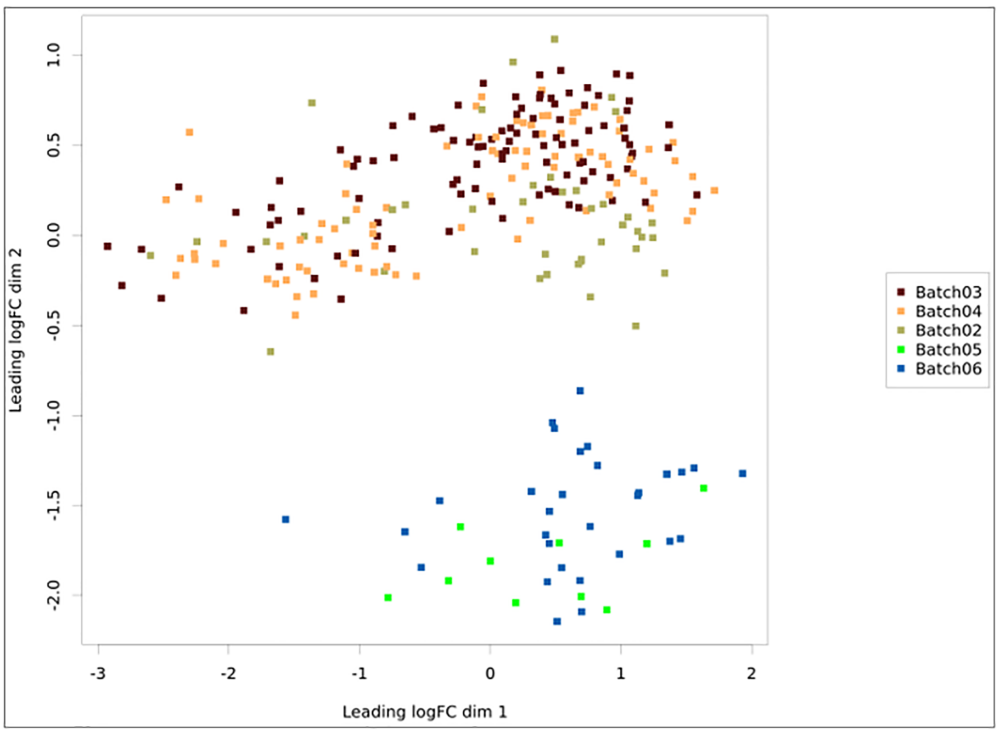


FIGURE 4 | Multidimensional scaling plot with the samples colored according to batch. The *x*- and *y*-axes represent the leading log-fold-changes between each pair of RNA samples, which is given by the root-mean-square of the largest absolute log-fold-changes between each pair of samples. In this example, there is a batch bias, whereas samples from batches 2, 3, and 4 are separated from 5 and 6.

other (Is there an interaction effect? Are they independent?). *Simplicity DiffExpress* tackles this challenge providing support to modeling both continuous and categorical variables, regardless of how many levels a category can have, enabling interaction analysis while providing feedback on issues preventing the statistical model fitting.

For example, in the work by Oliveira et al. (2018) rats were submitted to low, moderate and high-intensity treadmill protocols to investigate the impact of exercise on serum extracellular vesicles and their small RNAs. If the exercise intensity was modeled as a factor of four levels ("no exercise," "low," "moderate," and "high" intensities), the experimental design would be misrepresented because the intensity levels would be interpreted as unrelated treatments. What should be done instead, is to include the average treadmill speed applied to each group and modeled it as a continuous variable, enabling to capture potential gradual expression changes in relation to the speed. Going back to the analysis of the blood platelet samples, in **Figure 3** we can observe that no transcripts related to *Metastasis* were found differentially expressed. This is likely because the metastasis features and onset changes depend on the cancer type, therefore a better model would include an interaction between the variables *Cancer* and *Metastasis*.

Simplicity DiffExpress core analysis is based on the well-known and broadly used resources offered by the R (R Core Team, 2017) package *edgeR* (Robinson et al., 2010; McCarthy et al., 2012). *Simplicity DiffExpress* makes the valuable *edgeR* features available to a non-bioinformatician public and augments the use of *edgeR* with key validation support, used to identify issues in the dataset and statistical model prior to running the analysis. This is a crucial feature since, when running a script for an *edgeR*-based analysis, many errors are only identified after some time is elapsed, therefore a strong validation is a valuable contribution toward the analysis process. Moreover, the technical aspects of the analysis (input format, validation issues, statistical parameters) are presented in clear language in order to make it accessible to non-specialists. All these features are combined with detailed documentation, which includes insights into the statistical aspects of the analysis and a step-by-step tutorial.

In comparison to other web applications that provide DEA for the user without programming experience, like DEApp (Li and Andrade, 2017) and DEBrowser (Kucukural et al., 2019), the key advantages of *Simplicity DiffExpress* are related to input files and complex data modeling. *Simplicity DiffExpress* has no limit for file size and offers clearer feedback regarding issues when reading the files and incompatibilities between count-table and metadata. To our knowledge, *Simplicity DiffExpress* is the only platform of this type that allows the analysis of variables as continuous, which is very important as explained above, and it offers more flexible options to define interactions for multi-factorial analysis because the interactions are not mandatory and can be done with specific features. Moreover, both DEApp and DEBrowser require the user to inform manually each paired comparison to be studied, which can be not practical when dealing with

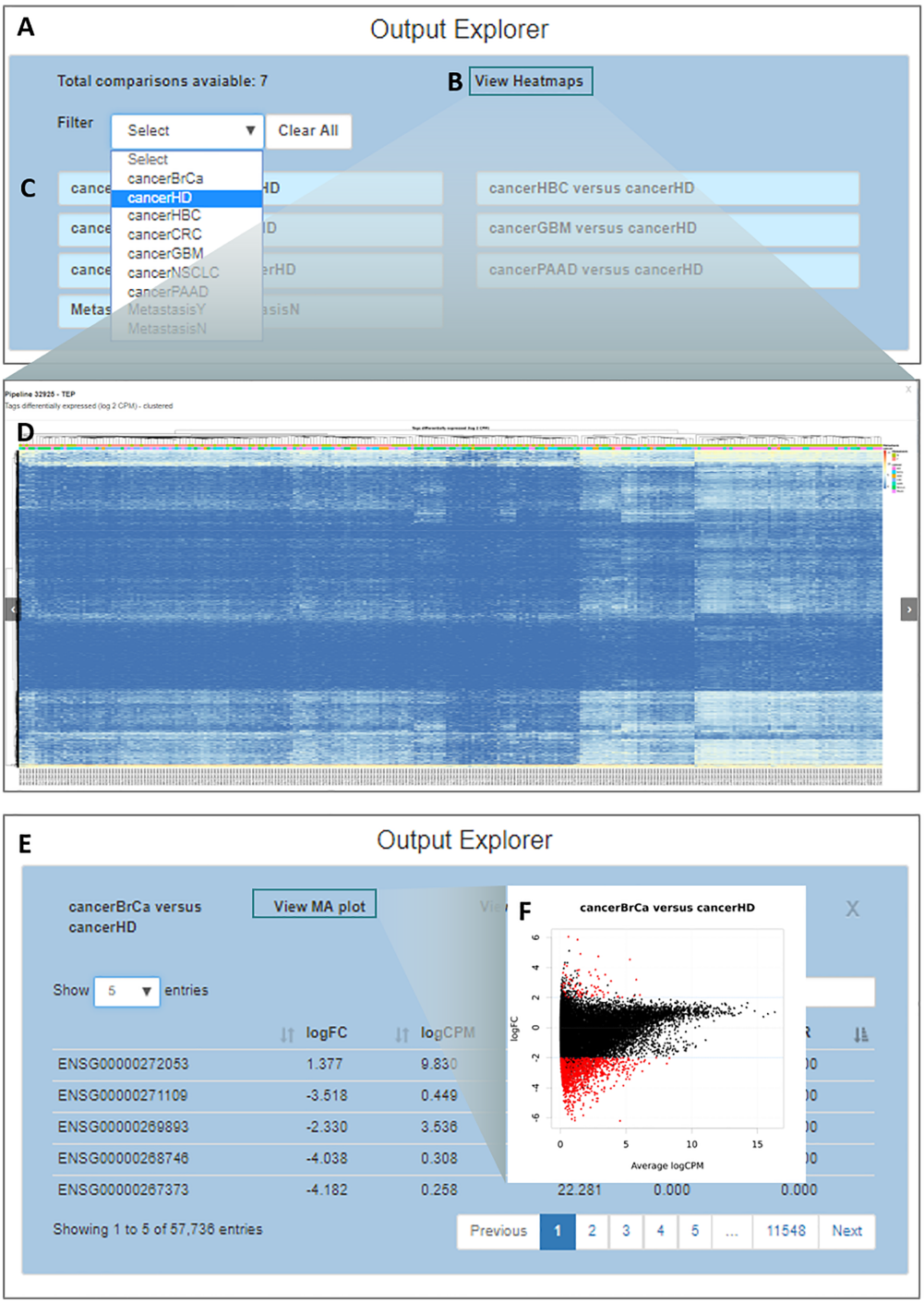


FIGURE 5 | Output explorer options. **(A)** The initial window where **(B,D)** heatmaps can be accessed and **(C)** listing all comparisons across variables. It is possible to filter the comparison list based on the variable category. **(E)** Results of a selected comparison (in this case BrCa vs. HD – breast cancer versus healthy donor). **(F)** An MA plot displaying the log (base 2) fold-change observed for the average log CPM of each group of interest (e.g., BrCa and HD), with genes differentially expressed highlighted in red.

many variables or categorical variables with many levels. In summary, *Simplicity DiffExpress* structure is more robust to deal with datasets with complex metadata, besides the fact it is able to store and share the results.

Simplicity DiffExpress can be used on a broad range of data sources, as long as the RNA-seq data is summarized in the read-count table, without any transformation. The investigation of complex biological outcomes will greatly benefit from *Simplicity*

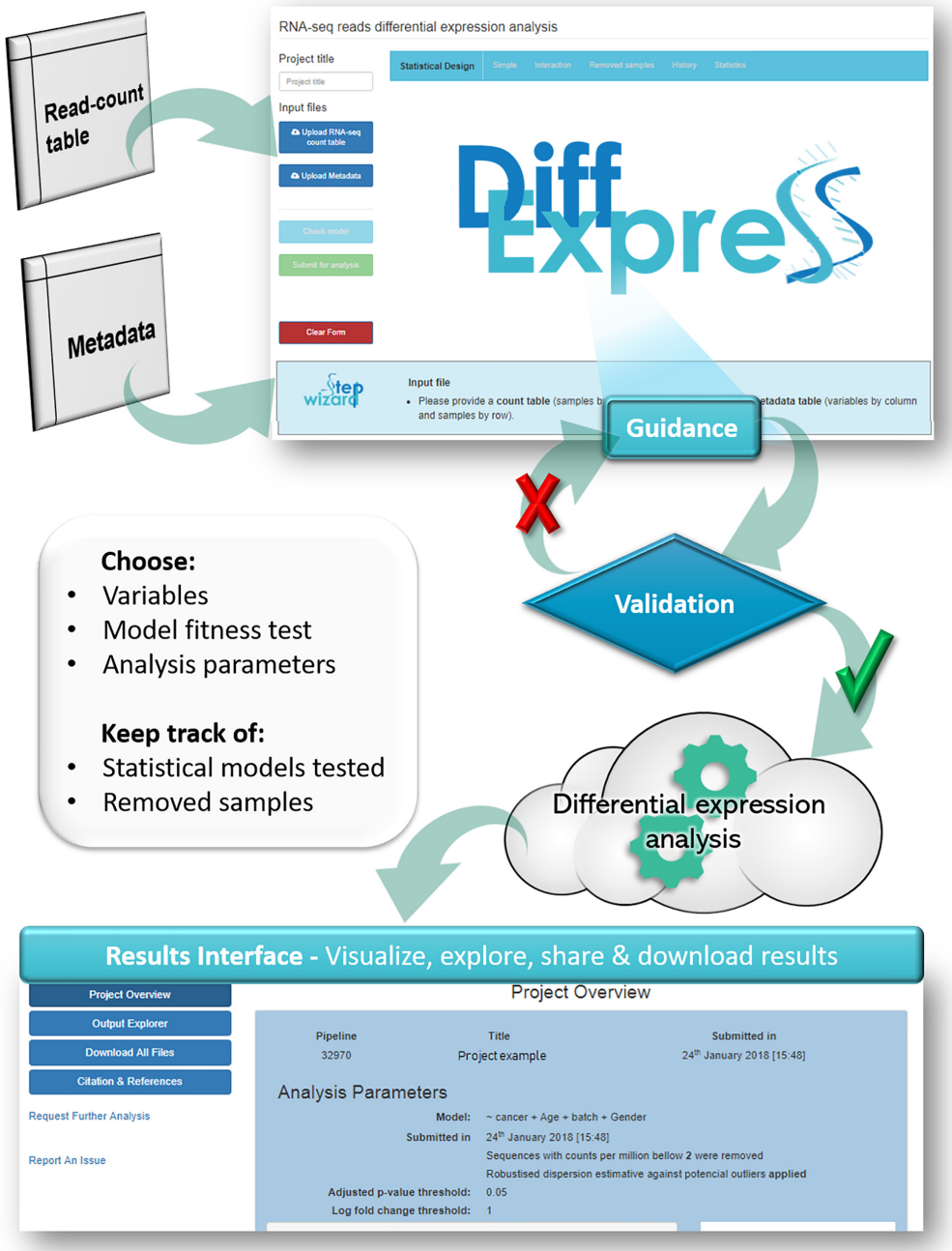


FIGURE 6 | Overview of the *DiffExpress* workflow. The web-interface requires a read-count table and a metadata as input. The users will set-up their analysis which will undergo a real-time validation prior to submission for analysis. If any issue is encountered, the interface will guide the users through the possible actions. Once the statistical model is successfully fitted, the pipeline can be submitted. The data will be sent to the cloud and once the analysis is finished, the users can retrieve and explore the results in a secondary web interface, linked to their accounts.

DiffExpress features. For example, RNA-based measurements can be applied across diverse areas of human health, including disease diagnosis, prognosis, and therapeutic decisions. At the moment, it supports clinical practice for infectious diseases, cancer, transplant medicine, and fetal monitoring (Byron et al., 2016). *Simplicity DiffExpress* features offer useful assistance for health-care because it provides functionality for guiding users on modeling multi-factorial and temporal designs. When dealing with cohort studies there are many bias sources beyond the obvious genetic variability across individuals. By enabling investigators with clinical knowledge to run their own DEA, our software increases the possibilities of discovery because users can

combine variables, correct for sources of bias and test hypotheses themselves and at their convenience since they no longer depend on an intermediary researcher between them and the analysis. It can also be used as a means to support the communication between bioinformatician and wet-lab researchers because it presents the data in a user-friendly set-up.

Differential expression analysis can generate a high number of outputs depending on the experimental design. *Simplicity DiffExpress* also addresses file management issues by saving the analysis parameters and organizing the output files systematically. This feature supports research reproducibility and reporting. Moreover, the *sharing* feature facilitates the exchange between collaborators, avoids e-mail clutter and promotes transparency.

CONCLUDING REMARKS

Simplicity DiffExpress aims to support the research of differentially expressed sequences by providing an intuitive interface with guidance through the steps and, on overcoming data modeling issues. Another critical advance provided by *Simplicity DiffExpress* is the data validation: besides checking the correspondence between samples IDs in the input files, it tests the statistical model fitness prior to the DEA enabling the immediate identification of any issues in the design and indicating solutions for it. This feature advances the functionalities provided by the R library *edgeR* (Robinson et al., 2010; McCarthy et al., 2012). Moreover, the results interface was designed to present the outputs of the DEA in an organized and easy to navigate format, addressing an issue regarding files management that can be critical since, depending on the experimental design, the output results can be extensive.

AUTHOR CONTRIBUTIONS

CP conceptualized the study, performed statistical analysis, validated the software, and wrote the original draft. MR-A conceptualized the study, performed statistical analysis, wrote, reviewed and edited the manuscript. YW performed software validation. BL performed software validation, and reviewed the manuscript. PB conceptualized the study, supervised, reviewed, and edited the manuscript. BK conceptualized and supervised the study, and reviewed the manuscript. PW conceptualized the study, performed software validation, supervised the study and reviewed the manuscript.

FUNDING

This work was supported by the grants from the Irish Research Council Enterprise Partnership; the SFI Industry Fellowship for Multi-Gene Assay Cloud Computing Platform – (16/IFA/4342) and SAGE-CARE (Project ID: 644186).

ACKNOWLEDGMENTS

We thank Olumakinde Adeyemi for co-developing *DiffExpress*' tutorial and Getulio Pereira de Oliveira Jr. for testing the interface from the early stages.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00356/full#supplementary-material

REFERENCES

Best, M. G., Sol, N., Kooi, I., Tannous, J., Westerman, B. A., Rustenburg, F., et al. (2015). RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* 28, 666–676. doi: 10.1016/j.ccell.2015.09.018

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., and Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* 17, 257–271. doi: 10.1038/nrg.2016.10

Chen, Y., Lun, A. T. L., and Smyth, G. K. (2014). "Differential expression analysis of complex RNA-seq experiments using edgeR," in *Statistical Analysis of Next Generation Sequence Data*, eds S. Datta and D. S. Nettleton (New York, NY: Springer), 1–25. doi: 10.1007/978-3-319-07212-8_3

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8

Costa-Silva, J., Domingues, D., and Lopes, F. M. (2017). RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS One* 12:e0190152. doi: 10.1371/journal.pone.0190152

Finotello, F., and Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief. Funct. Genomics* 14, 130–142. doi: 10.1093/bfgp/elu035

Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinform. Biol. Insights* 9(Suppl. 1), 29–46. doi: 10.4137/BBI.S28991

Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat. Med.* 9, 811–818. doi: 10.1002/sim.4780090710

Kolde, R. (2012). *pheatmap: Pretty Heatmaps. R Packag. version 1.0.8, 1–7.* .

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., and Garber, M. (2019). DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* 20:6. doi: 10.1186/s12864-018-5362-x

Li, Y., and Andrade, J. (2017). DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code Biol. Med.* 12:2. doi: 10.1186/s13029-017-0063-4

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

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi: 10.1126/science.1158441

Oh, S., Song, S., Dasgupta, N., and Grabowski, G. (2014). The analytical landscape of static and temporal dynamics in transcriptome data. *Front. Genet.* 5:35. doi: 10.3389/fgene.2014.00035

Oliveira, G. P., Porto, W. F., Palu, C. C., Pereira, L. M., Petriz, B., Almeida, J. A., et al. (2018). Effects of acute aerobic exercise on rats serum extracellular vesicles diameter, concentration and small RNAs content. *Front. Physiol.* 9:532. doi: 10.3389/fphys.2018.00532

Ooms, J. (2014). *The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv*. Available at: https://arxiv.org/abs/1403.2805(accessedMay12,2018).

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biol.* 11:220. doi: 10.1186/gb-2010-11-12-220

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Available at: https://www.R-project.org/ (accesses October 31, 2017)

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 Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Robinson, M. D., and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887. doi: 10.1093/bioinformatics/btm453

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

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Walsh, P., Carroll, J., and Sleator, R. D. (2013). Accelerating in silico research with workflows: a lesson in simplicity. *Comput. Biol. Med.* 43, 2028–2035. doi: 10.1016/j.compbiomed.2013.09.011

Yuryev, A. (2015). Gene expression profiling for targeted cancer treatment. *Expert Opin. Drug Discov.* 10, 91–99. doi: 10.1517/17460441.2015.971007

Conflict of Interest Statement: CP, YW, and BL work in collaboration with the company NSilico Life Science Ltd. BK is the CEO and PW is the CTO of the company NSilico Life Science Ltd.

The remaining 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

TECHNOLOGY REPORT
published: 31 May 2019
doi: 10.3389/fgene.2019.00499

ABioTrans: A Biostatistical Tool for Transcriptomics Analysis

Yutong Zou[1†], Thuy Tien Bui[2†] and Kumar Selvarajoo[2]*

[1] Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore,
[2] Biotransformation Innovation Platform (BioTrans), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

Here we report a bio-statistical/informatics tool, ABioTrans, developed in R for gene expression analysis. The tool allows the user to directly read RNA-Seq data files deposited in the Gene Expression Omnibus or GEO database. Operated using any web browser application, ABioTrans provides easy options for multiple statistical distribution fitting, Pearson and Spearman rank correlations, PCA, *k*-means and hierarchical clustering, differential expression (DE) analysis, Shannon entropy and noise (square of coefficient of variation) analyses, as well as Gene ontology classifications.

Keywords: transcriptomics, correlation, entropy, noise, DEG (differentially expressed genes), RNA-seq, clustering, gene expression data

OPEN ACCESS

Edited by:
Juilee Thakar,
University of Rochester, United States

Reviewed by:
Gaurav Sablok,
Finnish Museum of Natural History, Finland
Frederico Moraes Ferreira,
University of São Paulo, Brazil

***Correspondence:**
Kumar Selvarajoo
kumar_selvarajoo@biotrans.a-star.edu.sg

† 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: *17 August 2018*
Accepted: *07 May 2019*
Published: *31 May 2019*

Citation:
Zou Y, Bui TT and Selvarajoo K (2019) ABioTrans: A Biostatistical Tool for Transcriptomics Analysis. Front. Genet. 10:499. doi: 10.3389/fgene.2019.00499

INTRODUCTION

Large-scale gene expression analysis requires specialized statistical or bioinformatics tools to rigorously interpret the complex multi-dimensional data, especially when comparing between genotypes. There are already several such tools developed with fairly user-friendly features (Russo and Angelini, 2014; Poplawski et al., 2016; Velmeshev et al., 2016). Nevertheless, there still is a need for more specialized, focused and "click-and-go" analysis tools for different groups of bioinformatics and wet biologists. In particular, software tools that perform gene expression variability through entropy and noise analyses are lacking. Here, we focused on very commonly used statistical techniques, namely, Pearson and Spearman rank correlations, Principal Component Analysis (PCA), *k*-means and hierarchical clustering, Shannon entropy, noise (square of coefficient of variation), differential expression (DE) analysis, and gene ontology classifications (Tsuchiya et al., 2009; Piras et al., 2014; Piras and Selvarajoo, 2015; Simeoni et al., 2015).

Using R programming as the backbone, we developed a web-browser based user interface to simply perform the above-mentioned analyses by a click of a few buttons, rather than using a command line execution. Our interface is specifically made simple considering wet lab biologists as the main users. Nevertheless, our tool will also benefit bioinformatics and computational biologists at large, as it saves much time for running the R script files for analyses and saving the results in pdf.

MAIN INTERFACE AND DATA INPUT

Upon loading ABioTrans.R, the homepage window pops up and displays a panel to choose the RNA-Seq data and supporting files (**Figure 1**). The data file, in comma-separated value (.csv) format, should contain the gene names in rows and genotypes (conditions: wildtype, mutants, and replicates, etc.) in columns, following the usual format of files deposited in the GEO database (Clough and Barrett, 2016). Supporting files (if applicable) include gene length, list of negative

control genes, and metadata file. If the data files contain raw read counts, the user can perform normalization using 5 popular methods: FPKM, RPKM, TPM, Remove Unwanted Variation (RUV), or upper quartile in the pre-processing step (Mortazavi et al., 2008; Trapnell et al., 2010; Wagner et al., 2012; Risso et al., 2014). FPKM, RPKM, and TPM normalization requires inputting gene length file, which should provide matching gene name and their length in base pair in two-column csv file. RUV normalization requires a list of negative control genes (genes that are stably expressed in all experimental conditions), which should be contained in a one-column csv file. If negative control genes are not available, upper quartile normalization option will replace RUV. The metadata file is required for DE analysis, and should specify experimental conditions (e.g., Control, Treated, etc.) for each genotype listed in the data file. Otherwise, the user can move to the next option to perform/click all available analysis buttons (scatter plot, distribution fit, and Pearson Correlation, etc.) once a data file is loaded (whether normalized or in raw count).

DATA PRE-PROCESSING

Upon submitting data files and all supporting files (gene length, negative control genes, and metadata table), the user can filter the lowly expressed genes by indicating the minimum expression value and the minimum number of samples that are required to exceed the threshold for each gene. If input data contain raw read counts, user can choose one of the normalization options (FPKM, RPKM, TPM, upper quartile, and RUV) listed upon availability of supporting files. FPKM, RPKM, and TPM option perform normalization for sequencing depth and gene length, whereas RUV and upper quartile eliminate unwanted variation between samples. To check for sample variation, Relative Log Expression (RLE) plots (Gandolfo and Speed, 2018) of input and processed data are displayed for comparison.

SCATTER PLOT AND DISTRIBUTIONS

The scatter plot displays all gene expressions between any two columns selected from the datafile. This is intended to show, transcriptome-wide, how each gene expression varies between any two samples. The lower the scatter, the more similar the global responses and vice-versa (Piras et al., 2014). That is, this option allows the user to get an indication of how variable the gene expressions are between any two samples (e.g., between 2 different genotypes or replicates).

After knowing this information, the next process is to make a distribution (cumulative distribution function) plot and compare with the common statistical distributions. As gene expressions are known to follow certain statistical distributions such as power-law or lognormal (Furusawa and Kaneko, 2003; Bengtsson et al., 2005; Beal, 2017; Bui et al., 2018), we included the distribution test function. Previously, we have used power-law distribution to perform low signal-to-noise expression cutoff with FPKM expression threshold of less than 10 (Simeoni et al., 2015). Thus, this mode allows the user to check the deviation of their expression pattern with appropriate statistical distributions to select reliable genes for further analysis.

ABioTrans allows the comparison with (i) log-normal, (ii) Pareto or power-law, (iii) log-logistic (iv) gamma, (v) Weibull, and (vi) Burr distributions. To compare the quality of statistical distribution fit, the Akaike information criterion (AIC) can also be evaluated on this screen.

PEARSON AND SPEARMAN CORRELATIONS

This mode allows the user to compute linear (Pearson) and monotonic non-linear (Spearman) correlations, (i) in actual values in a table or (ii) as a density gradient plot between the samples.

PCA AND K-MEANS CLUSTERING

The PCA button plots the variance of all principal components and allows 2-D and 3-D plots of any PC-axis combination. There is also a slide bar selector for testing the number of *k*-means clusters.

ENTROPY AND NOISE

These functions measure the disorder or variability between samples using Shannon entropy and expressions scatter (Shannon, 1948; Bar-Even et al., 2006). Entropy values are obtained through binning approach and the number of bins are determined using Doane's rule (Doane, 1976; Piras et al., 2014).

To quantify gene expressions scatter, the noise function computes the squared coefficient of variation (Gandolfo and Speed, 2018), defined as the variance (σ^2) of expression divided by the square mean expression (μ^2), for all genes between all possible pairs of samples (Piras et al., 2014).

DIFFERENTIAL EXPRESSION ANALYSIS

ABioTrans provides users with 3 options to carry out DE analysis on data with replicates: edgeR, DESeq2, and NOISeq (McCarthy et al., 2012; Love et al., 2014; Tarazona et al., 2015). In case there are no replicates available for any of the experimental condition, technical replicates can be simulated by NOISeq. edgeR and DESeq2 requires filtered raw read counts, therefore, it is recommended that the user provide input data file containing raw counts if DE analysis is required using either of the two methods. On the other hand, if only normalized gene expression data is available, NOISeq is recommended.

FIGURE 1 | ABioTrans main interface and snapshots of various analysis mode.

To better visualize DE analysis result by edgeR and DESeq, volcano plot (plot of $\log_{10}$-p-value and $\log_2$-fold change for all genes) distinguishing the significant and insignificant, DE and non-DE genes, is displayed. Plot of dispersion estimation, which correlates to gene variation, is also available in accordance to the selected analysis method.

HIERARCHICAL CLUSTERING AND HEATMAP

This function allows clustering of differentially expressed genes. User can either utilize the result from DE analysis, or carry out clustering independently by indicating the minimum fold change between 2 genotypes.

For clustering independently, normalized gene expression (output from pre-processing tab) first undergo scaling defined by $Z_j(p_i) = (x_j(p_i) - (\bar{x}_j))/\sigma_{x_j}$ where $Z_j(p_i)$ is the scaled expression of the jth gene, $x_j(p_i)$ is expression of the jth gene in sample p_i, $\bar{x}_j$ is the mean expression across all samples and σ_{x_j} is the standard deviation (Simeoni et al., 2015). Subsequently, Ward hierarchical clustering is applied on the scaled normalized gene expression.

ABioTrans also lists the name of genes for each cluster.

GENE ONTOLOGY

This function is used to define the biological processes or enrichment of differentially regulated genes in a chosen sample or cluster. User can select among 3 gene ontology enrichment test: enrichR, clusterProfiler and GOstats (Falcon and Gentleman, 2007; Yu et al., 2012; Kuleshov et al., 2016).

The user needs to create a new csv file providing the name of genes (for each cluster) in 1 column (foreground genes). Background genes (or reference genes), if available, should be prepared in the same format. Next, the sample species, gene ID type (following NCBI database (Clough and Barrett, 2016)) and one of the three subontology (biological process, molecular function, or cellular component) need selection. The output results in a gene list, graph (clusterProfiler), and pie chart (clusterProfiler and GOstats) for each ontology.

TYPICAL ANALYSIS TIME ESTIMATION

The loading time of ABioTrans for a first time R user is about 30 min on a typical Windows notebook or Macbook. This is due to the installation of the various R-packages that are prerequisite to run ABioTrans. For regular R users, who have installed most packages, the initial loading can take between a few to several minutes depending on whether package updates are required. Once loaded, the subsequent re-load will take only a few seconds.

The typical time taken from pre- to post-processing using all features in ABioTrans is between 10–20 min. **Table 1** below highlights the typical time taken for each execution for 3 sample data deposited in ABioTrans Github folder (*zfGenes*, *Biofilm-Yeast*, and *Yeast-biofilm2*).

ABioTrans has also been compared with other similar freely available RNA-Seq GUI tools, and it

TABLE 1 | Time comparison of functionalities for different test data.

Type of analysis		Time (s)		
		Test 1*	Test 2#	Test 3^
Pre-processing	TPM/RPKM/FPKM and RLE plot	–	–	0.6 s
	Upper quartile normalization and RLE plot	–	0.5 s	0.6 s
	RUV normalization and RLE plot	1.7 s	–	–
Scatter plot		0.01 s	0.01 s	0.01 s
Distribution fitting (for all 6 distributions)		4.3 s	3.1 s	2.5 s
Correlation matrix		0.01	0.01 s	0.01 s
PCA calculation and plotting		0.01	0.01	0.01
DE analysis	edgeR	7.89	1.52 s	5.23 s
	DESeq2	15.4 s	3.1 s	11.3 s
	NOISeq	29.6 s	22.87 s	31.0 s
Heat map and hierarchical clustering	DE (using edgeR result) (5 clusters)	0.36 s	1.7 s	0.25 s
	Independent (5 clusters)	30.4 s	7.7 s	4.6 s
Noise		3.2 s	1.3 s	3.9 s
Shannon entropy		0.03 s	0.02 s	0.08 s
GO analysis (using edgeR result)	clusterProfiler	20.2 s	10.3 s	9.1 s
	GOstats	26.6 s	10.2 s	12.3 s
	EnrichR	–	–	–

Risso et al., 2014: GEO accession number: GSE53334. #Bendjilali et al., 2017: GEO accession number: GSE85595. ^Cromie et al., 2017: GEO accession number: GSE85843.

demonstrates better functionalities and capabilities (**Supplementary Table S1**).

SUMMARY

ABioTrans is a user-friendly, easy-to-use, point-and-click statistical tool tailored to analyse RNA-Seq data files. It can also be used to analyse any high throughput data as long as they follow the format listed in this technology report. The complete user manual to operate ABioTrans is available as **Supplementary Data Sheet S1** in **Supplementary Material** posted online.

AVAILABILITY AND IMPLEMENTATION

ABioTrans is available at: https://github.com/buithuytien/ABioTrans, Operating system(s): Platform independent (web browser), Programming language: R (RStudio), Other requirements: Bioconductor genome wide annotation databases, R-packages (shiny, LSD, fitdistrplus, actuar, entropy, moments, RUVSeq, edgeR, DESeq2, NOISeq, AnnotationDbi, ComplexHeatmap, circlize, clusterProfiler, reshape2, DT, plotly, shinycssloaders, dplyr, ggplot2). These packages will automatically be installed when the ABioTrans.R is executed in RStudio. No restriction of usage for non-academic.

AUTHOR CONTRIBUTIONS

YZ and TTB developed the software tool. KS planned, designed the tool and wrote the manuscript. **Supplementary Data Sheet S1** (user manual) was prepared by TTB.

FUNDING

The authors thank the Biotransformation Innovation Platform (BioTrans, A*STAR) for funding the work.

ACKNOWLEDGMENTS

The authors thank Lin Yifeng, Ng Shi Yuan, and Nic Lindley for discussion of the work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00499/full#supplementary-material

TABLE S1 | Comparison of functionalities of ABioTrans with other RNA-Seq tools.

REFERENCES

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., et al. (2006). Noise in protein expression scales with natural protein abundance. *Nat. Genet.* 38, 636–643.

Beal, J. (2017). Biochemical complexity drives log-normal variation in genetic expression. *IET Eng. Biol.* 1, 55–60.

Bendjilali, N., MacLeon, S., Kalra, G., Willis, S. D., Hossian, A. K., Avery, E., et al. (2017). Time-course analysis of gene expression during the saccharomyces cerevisiae hypoxic response. *G3* 7, 221–231.

Bengtsson, M., Stahlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* 15, 1388–1392.

Bui, T. T., Giuliani, A., and Selvarajoo, K. (2018). Statistical Distribution as a Way for Lower Gene Expressions Threshold Cutoff. *J. Biol. Sci.* 2, 55–57. doi: 10.13133/2532-5876_4.6

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Cromie, G. A., Tan, Z., Hays, M., and Jeffery, E. W. (2017). Dissecting gene expression changes accompanying a ploidy-based phenotypic switch. *G3* 7, 233–246. doi: 10.1534/g3.116.036160

Doane, D. P. (1976). Aesthetic frequency classification. *Am. Stat.* 30, 181–183.

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* 23, 257–258.

Furusawa, C., and Kaneko, K. (2003). Zipf's law in gene expression. *Phys. Rev. Lett.* 90:088102.

Gandolfo, L. C., and Speed, T. P. (2018). RLE plots: visualizing unwanted variation in high dimensional data. *PLoS One* 13:e0191629. doi: 10.1371/journal.pone.0191629

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

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.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Piras, V., and Selvarajoo, K. (2015). The reduction of gene expression variability from single cells to populations follows simple statistical laws. *Genomics* 105, 137–144. doi: 10.1016/j.ygeno.2014.12.007

Piras, V., Tomita, M., and Selvarajoo, K. (2014). Transcriptome-wide variability in single embryonic development cells. *Sci. Rep.* 4:7137. doi: 10.1038/srep07137

Poplawski, A., Marini, F., Hess, M., Zeller, T., Mazur, J., and Binder, H. (2016). Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective. *Brief. Bioinform. Mar.* 17, 213–223. doi: 10.1093/bib/bbv036

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931

Russo, F., and Angelini, C. (2014). RNASeqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics* 30, 2514–2516. doi: 10.1093/bioinformatics/btu308

Shannon, C. E. (1948). A mathematical theory of communication. *Bell. Syst. Tech. J.* 379–423, 623–656.

Simeoni, O., Piras, V., Tomita, M., and Selvarajoo, K. (2015). Tracking global gene expression responses in T cell differentiation. *Gene* 569, 259–266. doi: 10.1016/j.gene.2015.05.061

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* 43:e140. doi: 10.1093/nar/gkv711

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tsuchiya, M., Piras, V., Choi, S., Akira, S., Tomita, M., Giuliani, A., et al. (2009). Emergent genome-wide control in wildtype and genetically mutated lipopolysaccarides-stimulated macrophages. *PLoS One* 4:e4905. doi: 10.1371/journal.pone.0004905

Velmeshev, D., Lally, P., Magistri, M., and Faghihi, M. A. (2016). CANEapp: a user-friendly application for automated next generation transcriptomic data analysis. *BMC Genomics* 17:49. doi: 10.1186/s12864-015-2346-y

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285. doi: 10.1007/s12064-012-0162-3

Yu, G., Wang, L., Han, Y., and He, Q. (2012). Clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

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

METHODS
published: 21 June 2019
doi: 10.3389/fgene.2019.00594

BioNetStat: A Tool for Biological Networks Differential Analysis

Vinícius Carvalho Jardim[1,2], *Suzana de Siqueira Santos*[1], *Andre Fujita*[1] and *Marcos Silveira Buckeridge*[2]*

[1] Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil,
[2] Department of Botany, Institute of Biosciences, University of São Paulo, São Paulo, Brazil

The study of interactions among biological components can be carried out by using methods grounded on network theory. Most of these methods focus on the comparison of two biological networks (e.g., control vs. disease). However, biological systems often present more than two biological states (e.g., tumor grades). To compare two or more networks simultaneously, we developed `BioNetStat`, a `Bioconductor` package with a user-friendly graphical interface. `BioNetStat` compares correlation networks based on the probability distribution of a feature of the graph (e.g., centrality measures). The analysis of the structural alterations on the network reveals significant modifications in the system. For example, the analysis of centrality measures provides information about how the relevance of the nodes changes among the biological states. We evaluated the performance of `BioNetStat` in both, toy models and two case studies. The latter related to gene expression of tumor cells and plant metabolism. Results based on simulated scenarios suggest that the statistical power of `BioNetStat` is less sensitive to the increase of the number of networks than Gene Set Coexpression Analysis (GSCA). Also, besides being able to identify nodes with modified centralities, `BioNetStat` identified altered networks associated with signaling pathways that were not identified by other methods.

Keywords: differential network analysis, coexpression network, correlation network, systems biology, systems biology tool, differential coexpression, differential correlation

OPEN ACCESS

Edited by:
Helder Nakaya,
University of São Paulo, Brazil

Reviewed by:
Diego Bonatto,
Federal University of Rio Grande do Sul, Brazil
Ling-Yun Wu,
Academy of Mathematics and Systems Science (CAS), China

***Correspondence:**
Marcos Silveira Buckeridge
msbuckeridge@gmail.com

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

Received: *21 February 2019*
Accepted: *05 June 2019*
Published: *21 June 2019*

Citation:
Jardim VC, Santos SS, Fujita A and Buckeridge MS (2019) BioNetStat: A Tool for Biological Networks Differential Analysis. Front. Genet. 10:594. doi: 10.3389/fgene.2019.00594

1. INTRODUCTION

In the last two decades, the high-dimensional data production, such as metabolomics, proteomics, transcriptomics, and genomics, increased considerably (Zhu et al., 2008; McKenzie et al., 2016). It brings out the high complexity of the biological systems, posing the challenge to understand how they work. In science, it is fundamental to compare the many states assumed by a system, such as sick against healthy patients or developmental stages of a living being. A range of strategies can be applied for comparing different states depending on the study hypothesis, such as the *t*-test (to compare two means), the analysis of variance—ANOVA (to compare two or more means) (de Souza et al., 2008; Wu et al., 2016) or Gene Set Enrichment Analysis (GSEA), to test whether a gene set is differentially expressed between two conditions (Subramanian et al., 2005). However, none of these methods takes into account the relationship among several biological components at the same time. In this sense, methods based on networks represent the association between each pair of components and may help to understand the role each variable plays in the system (Barabási and Oltvai, 2004).

Biological systems can be assessed by correlation networks, in which the nodes represent the elements (variables) and edges represent the statistical relations among its elements. Some approaches have been proposed to qualitatively analyze the correlation networks by performing a visual inspection of their structure (Caldana et al., 2011; Weston et al., 2011), while others are based on formal strategies to search for differences between biological networks (Sun et al., 2013; Li et al., 2016; Zhang and Yin, 2016). However, these studies do not apply statistical tests or formal control of false positives.

Over the last years, several tools have been developed to statistically test whether correlation networks are different across conditions. Examples include `DCGL` (Liu et al., 2010), `EBcoexpress` (Dawson et al., 2012), `DiffCorr` (Fukushima, 2013), and `CoDiNA` (Gysi et al., 2018), which evaluate whether the correlations between pairs of nodes are different among biological states. `DiffCoEx` (Tesson et al., 2010) `coXpress` (Watson, 2006) searches for cohesive subgroups of variables in one of the states and evaluates whether these groups change their correlation patterns among states. `DINGO` (Ha et al., 2015), `DECODE` (Lui et al., 2015), `dCoxS` (Cho et al., 2009), `GSCA` (Choi and Kendziorski, 2009), `GSNCA` (Rahmatallah et al., 2014), and `CoGA` (Santos et al., 2015) compare predefined sets of variables (Santos et al., 2015). Here we focus on the last group, in which the tests are performed for each predefined group of variables.

Although several biological studies compare more than two networks (Caldana et al., 2011; Weston et al., 2011; Hochberg et al., 2013; Zhang and Yin, 2016), to the best of our knowledge, there are only two tools that perform statistical tests to compare two or more networks simultaneously: `DiffCoEx` and `GSCA`. However, only `GSCA` performs tests for predefined groups of variables. `GSCA` builds correlations matrices and compares the biological condition networks by using Euclidean distance (Choi and Kendziorski, 2009). Pairwise comparison between the networks obtains the `GSCA` generalization for comparing more than two networks. However, this strategy, in general, gives an inadequate control of type I error (Fujita et al., 2017). Besides, since the network structure may vary over time and also across systems from the same biological class, searching for precisely similar structures between two graphs is not an effective strategy to compare the behavior of biological pathways (Santos et al., 2015).

In the context of functional brain network studies, a generalization of `CoGA`, named by `GANOVA`, has been proposed to compare more than two populations of graphs (Fujita et al., 2017). This tool is specific for datasets containing several graphs in each biological condition. `GANOVA` is not useful when only one network is available per condition, such as in the case of physiological or genes correlations networks. Here we combined the methods proposed by Santos et al. (2015) and Fujita et al. (2017) to compare two or more biological states, namely `BioNetStat`. `BioNetStat` is available at `Bioconductor` and includes a graphical user interface. We performed simulation experiments and applied the proposed method in two biological data sets.

2. MATERIALS AND METHODS

We propose a method for comparing simultaneously two or more biological correlation networks. In the following subsections, we explain the construction of correlation networks (graphs), the structural graph analysis, and the statistical test performed by `BioNetStat`.

2.1. Construction of Correlation Networks

A correlation network is an undirected graph, where each node corresponds to a biological variable, and each edge connects a pair of nodes indicating the association between two variables. In our context, the edge corresponds to the statistical dependence between two variables. To measure and detect monotonic relations, `BioNetStat` includes the Pearson (1920), Spearman (1904), and Kendall (1938) correlation coefficients. Given a measure of statistical dependence, `BioNetStat` provides three scales of association degree: the absolute correlation coefficient, one minus the p-value of the dependence test, and one minus the p-value adjusted by the False Discovery Rate method (Benjamini and Hochberg, 1995). Each association degree is a real number varying from zero to one. The user can choose between unweighted (zero or one) and weighted network (values from zero to one). Zero means no monotonic association between variables, while one means a monotonic association between them. To construct a graph, the user can choose a threshold for edges insertion, based on some association measure (correlation or p-value of the independence test).

The proposed method is based on graph topological features. In the following sections, we describe how `BioNetStat` performs the comparisons based in the Probability Distribution of a Feature of the Graph (PDFG), in the vector of some network centrality, and in each node centrality measure.

2.2. Differential Network Analysis of Multiple Graphs Based on PDFG

A random graph G is a graph generated by a random process. In the last decades, several random graph models have been proposed for studying biological networks. For example, Barabasi and Albert (1999) proposed the scale-free model, in which a few nodes have many connections (hubs) and many nodes present a lower number of connections (Jeong et al., 2000). An example where to which the scale-free model suits well is in the representation of the protein-protein interactions networks, in which only a few essential proteins interact with many others and are central to metabolism, whereas many proteins display lower numbers of interactions because they participate in a few specific metabolic pathways.

Consider a set of nodes $V = \{v_1, v_2, \ldots, v_{n_v}\}$ of the graph, r states $S_1, S_2, \ldots, S_r$, and o_i samples (number of observations) for each state S_i, for $i = 1, 2, \ldots, r$. We want to test whether the r graphs $G_1, G_2, \ldots, G_r$ (each one representing a state) were generated by the same random graph model. In case the PDFG are different, it would be assumed that the graphs were generated by different random graph models. As will be seen next, here we analyzed correlation networks in which the elements correspond to variables such as genes, proteins, metabolites, and phenotypic

variables. Examples of states include different treatments or conditions. An alteration in the structure of the network, detected by a change in the PDFG, could mean that a healthy human cell may be turning into a tumor cell or the tumor tissue might be entering in a new degree of aggressiveness.

The *differential network analysis* consists of the following steps: (i) construction of a correlation network for each state, which are denoted by $G_1, G_2, \ldots, G_r$, (ii) computation of the statistic test, denoted by θ, which quantifies the differences among the networks, and (iii) a permutation test.

The PDFG is the probability density function of some topological feature x and has n_v elements $x_1, x_2, \ldots, x_{n_v}$. Examples of topological features are the set of eigenvalues of the adjacency matrix of the graph, or graph centrality measures. Let δ be the Dirac's delta and the brackets "$\langle\rangle$" denote the expectation according to the probability law of a random graph. Formally, the PDFG (g) is defined as:

$$\rho_g(x) = \lim_{n_v \to \infty} \langle \frac{1}{n_v} \sum_{j=1}^{n_v} \delta(x - x_j/\sqrt{n_v}) \rangle \tag{1}$$

In real systems, the PDFG is unknown. To estimate the PDFG, `BioNetStat` uses the Gaussian Kernel estimator implemented by the function *density* of the `R base` package. The user can choose between the Sturges' (Sturges, 1926) and the Silverman's (Silverman, 1986) criteria to define the Kernel bandwidth for the Gaussian Kernel estimator. In the analyses performed in this work, we used the Sturges' criterion.

2.2.1. Computation of the Statistic Test

The *differential network analysis* is a comparison between two or more graphs based on their PDFG.

The θ statistic is calculated as follows:

1. For each graph G_i ($i = 1, \ldots, r$), compute the PDFG denoted by ρ_{g_i}.
2. Calculate the average PDFG as:

$$\rho_{g_M} = \frac{\sum_{i=1}^{r} \rho_{g_i}}{r}. \tag{2}$$

3. Calculate the Kullback-Leiber (KL) divergence between (ρ_{g_i}) and ρ_{g_M}:

$$D_i = KL(\rho_{g_i} | \rho_{g_M}) \tag{3}$$

4. The statistic θ, which measures the difference among graphs, is the average distance:

$$\theta = \frac{\sum_{i=1}^{r} D_i}{r}. \tag{4}$$

The KL divergence measures the discrepancy between two probability distributions. For graphs, we can use the KL divergence to select the graph model that best describes the observed graph or to discriminate PDFGs (Takahashi et al., 2012). Formally, we define the KL divergence between graphs as follows. Let g_1 and g_2 be two random graphs with densities ρ_{g_1} and ρ_{g_2}, respectively. If the support of ρ_{g_2} contains the support of ρ_{g_1}, then the KL divergence between ρ_{g_1} and ρ_{g_2} is (Takahashi et al., 2012):

$$KL(\rho_{g_1} | \rho_{g_2}) = -\int_{-\infty}^{+\infty} \rho_{g_1}(x) \log \frac{\rho_{g_1}(x)}{\rho_{g_2}(x)} dx \tag{5}$$

where $0 \log 0 = 0$ and ρ_{g_2} is called the reference measure. If the support of ρ_{g_2} does not contain the support of ρ_{g_1}, then $KL(\rho_{g_1} | \rho_{g_2}) = +\infty$. The KL divergence is non-negative, and it is zero if and only if ρ_{g_1} and ρ_{g_2} are equal. For many cases, $KL(\rho_{g_1} | \rho_{g_2})$ and $KL(\rho_{g_2} | \rho_{g_1})$ are different when ρ_{g_1} and ρ_{g_2} are not equal, i.e., KL is an asymmetric measure.

2.3. Differential Network Analysis of Multiple Graphs Based on Graph Centralities

As in section 2.2, consider a set of nodes $V = \{v_1, v_2, \ldots, v_{n_v}\}$ and a set of edges $E = \{e_1, e_2, \ldots, e_{n_e}\}$ of the graph, r states $S_1, S_2, \ldots, S_r$, and o_i samples (number of observations) of each state S_i, for $i = 1, 2, \ldots, r$. The aim is to test if the centrality values of r graphs $G_1, G_2, \ldots, G_r$, of each state, are the same among all graphs. `BioNetStat` considers five node centrality measures, namely degree, eigenvector, closeness, betweenness, and clustering coefficient, and one edge centrality (edge betweenness). The centrality measures quantify the importance of each node/edge according to its position in the network. The degree centrality counts the number of connections of a node (Barabási and Oltvai, 2004). In correlation networks, a node with high degree centrality is correlated with several other nodes/variables. This, such a node may be involved in numerous biological processes. The eigenvector centrality of a node is proportional to the centralities of its neighbors weighted by the strength of the connections (Bonacich, 1972). That is, a node is progressively more important as it connects with higher numbers of strongly connected neighbors nodes. The closeness and betweenness centralities are related to the shortest paths in the network (Rubinov and Sporns, 2010). The closeness centrality measures the average proximity of a node to all other nodes (Freeman, 1978). The betweenness centrality measures the importance of a node in the communication of the network. It counts how many shortest paths pass through the node (Freeman, 1978). The clustering coefficient quantifies how connected the neighbors of a node are (Watts and Strogatz, 1998). Finally, the edge betweenness centrality is similar to the betweenness centrality for nodes (Girvan and Newman, 2002). It quantifies how many shortest paths pass through an edge, measuring its importance in the communication of the network. The mathematical definitions of these six measures are shown in the **Table S5**.

Alterations in the centrality measures among networks means that the importance of the gene/protein/metabolite changed, i.e., its connectivity was altered regarding the main issues associated. Our tool, therefore, affords evaluation of data by assessing: (i) importance of a node in relation to the entire population of nodes in the network; (ii) proximity among nodes; (iii) importance of a node in the communication within the network, and (iv) the connectivity strength of the network as a whole.

The differential analysis consists of the same steps described in section 2.2.1. However, since in this case we are comparing the graphs centralities, the PDFG ρ_{g_i} is replaced by the vector of centrality measure and the D_i by the Euclidean distance between the vector of nodes/edges centralities of graph G_i and the vector containing the average centralities among the graphs (steps 2 and 3 of section 2.2.1).

2.4. Differential Node Analysis of Multiple Graphs Based on Node Centralities

Consider a set of nodes $V = \{v_1, v_2, \ldots, v_{n_v}\}$ and a set of edges $E = \{e_1, e_2, \ldots, e_{n_e}\}$ of the graph, r states $S_1, S_2, \ldots, S_r$, and o_i samples (number of observations) of each state S_i, for $i = 1, 2, \ldots, r$. The aim is to test if the importance (centrality value) of node v_j, for $j = 1, 2, \ldots, n_v$, or for the edge e_l, for $l = 1, 2, \ldots, n_e$, is the same among r graphs $G_1, G_2, \ldots, G_r$, of each state. In the same way that was done in section 2.3, here we considers the five node centrality measures (degree, eigenvector, closeness, betweenness, and clustering coefficient) and the edge centrality (edge betweenness).

The *differential node analysis* consists in similar steps as in section 2.2: (i) construction of a correlation network for each state, which are denoted by $G_1, G_2, \ldots, G_r$, (ii) computation of the statistic test, denoted by θ, which quantifies the differences among the node centralities of each network, and (iii) a permutation test.

2.4.1. Computation of the Test Statistic for Node Comparison

The θ statistic is calculated as follows:

1. For each node V_j ($j = 1, \ldots, n_v$) or for each edge E_l ($l = 1, \ldots, e_v$) in graph G_i ($i = 1, \ldots, r$), compute the node centrality denoted by C_i^j, or edge centrality, replacing j for l.
2. From the r centralities of each node/edge in each graph, we obtain an average node/edge centrality as:

$$M^j = \frac{\sum_{i=1}^{r} C_i^j}{r}. \quad (6)$$

3. Calculate the distance between the centrality of nodes/edges in each graph G_i (C_i^j) and the average node/edge centrality (M^j):

$$D_i^j = |C_i^j - M^j|. \quad (7)$$

4. The statistic θ, which measures the difference among centralities for each node/edge j of graphs, is the average distance:

$$\theta = \frac{\sum_{i=1}^{r} D_i^j}{r}. \quad (8)$$

2.5. Permutation Test

The hypotheses to be tested are defined as:

$H_0 : \theta = 0$ vs. $H_1 : \theta > 0$.

To construct the null hypothesis we perform a permutation test as follows:

1. Compute $\hat{\theta}$.
2. Construct r new graphs by resampling the observations without replacement.
3. Compute $\hat{\theta}^*$ by using the graphs constructed in step 2.
4. Repeat steps 2 and 3 until obtaining the desired number of permutation replications.
5. Test if $\hat{\theta} = 0$ using the empirical distribution obtained in steps 2–4. Gather the information from the empirical distribution of $\hat{\theta}^*$ to obtain a p-value for $\hat{\theta} = 0$, by analyzing the probability of obtaining values equal or greater than $\hat{\theta}$.

2.6. Description of the BioNetStat Package

`BioNetStat` is implemented in `R` http://cran.r-project.org/, provides a graphical interface, and is used to study correlation networks. It is based on the following packages: (i) `CoGA` to calculate the PDFG measures and the Kullback-Leibler divergence; (ii) `shiny`, `shinyBS`, `yaml`, `whisker`, and `RJSONIO` for browser interface; (iii) `igraph` to compute graph topological properties; (iv) `Hmisc` and `psych` for graph inference; and (v) `ggplot2`, `pathview`, `pheatmap`, and `RColorBrewer` for plotting.

`BioNetStat` receives two files as input. One is the *Biological samples file*, with the pre-processed data, containing the values of the variables (e.g. gene expression levels or metabolites concentration). This file must be a table, in which the columns indicate the variables and rows indicate the biological samples. At least one of these columns should indicate the label of rows (e.g. state to which each biological sample is related to). A second file, *variable set file*, contains the pre-defined set of variables (e.g., sets of biological variables belonging to the same metabolic pathways, sharing the same Gene Ontology terms). As an example of gene set collections, we suggest the use of Molecular Signature Database (MSigDB in http://www.broadinstitute.org/gsea/msigdb/index.jsp) (Subramanian et al., 2005), which is available for download.

For *differential network analysis*, presented in sections 2.2 and 2.3, `BioNetStat` returns a table containing the set name, the number of compared graphs, the size of each set, the statistics of the test, the permutation-based p-values, and the adjusted p-values by False Discovery Rate method (Benjamini and Hochberg, 1995) for multiple tests (q-values). An example of the output is shown in **Supplementary Data Sheet 1**. If the user performs the node differential analysis (section 2.4), the software returns a table containing the variable name, the statistics of the test, the permutation-based p-values, the q-values, and the node/edge centrality in each network, as shown in **Table 1**.

`BioNetStat` also includes a visual inspection of alterations in the correlation networks (heatmaps of the adjacency matrices). It also includes a list of the differences in the pairwise correlations, a table of variable set properties (e.g., spectral entropy, average node centrality, and average clustering coefficient) for each biological state, a rank of the centrality and local clustering coefficients, and a comparison of the measurements obtained in each state by heatmaps and boxplots. Also, `BioNetStat` provides a metabolic KEGG pathway view, using `pathview` `R` package. This functionality allows the user to visualize the gene expression, the concentration of proteins

TABLE 1 | Differential node analysis based on the degree centrality.

				Degree centrality			
	θ **Statistic**	***p*-value**	***q*-value**	**AST**	**OAST**	**ODG**	**GBM**
MAPK3	25.151	0.001	0.017	25	28.1	18.7	9.3
MAPK10	19.904	0.001	0.017	29	30.7	22.2	17.5
MAPK9	18.653	0.001	0.017	27.9	30.9	22.4	17.8
TOLLIP	17.877	0.002	0.026	25	28.2	20	15.3
TAB1	17.393	0.001	0.017	27.2	30.8	25.2	16.1
PIK3R1	17.098	0.001	0.017	28.9	30.7	24.5	18
AKT3	17.013	0.001	0.017	31.1	31.4	24.2	21.3
PIK3CB	15.215	0.002	0.026	29.1	31.8	23.9	21.7

Statistical tests results of the comparison of four glioma subtypes on Toll like receptor signaling pathway geneset networks. Eight genes were selected as differentially coexpressed, by considering a q-value threshold at 0.05. AST, astrocytoma; OAST, oligoastrocytoma; ODG, oligodendroglioma; GBM, glioblastoma. BioNetStat identified genes on important cell regulatory pathways involved in gliomas formation.

and metabolites, and the centrality of nodes at the KEGG pathway maps.

The `BioNetStat` pipeline is summarized in **Figure 1**. For a detailed tutorial and manual, we refer the user to the `Bioconductor` page: doi: 10.18129/B9.bioc. BioNetStat.

2.7. Example Datasets

To illustrate the utility of `BioNetStat`, we considered two different datasets: (i) gene expression dataset from glioma, and (ii) a plant metabolism dataset. The first dataset was selected because the cancer gene expression data contain thousands of variables and hundreds of samples (common features in this area), allowing robust analysis. The second dataset was motivated by a large number of experiments in plant studies that use a small number of replicates.

The glioma dataset was obtained from a public database (TCGA) (Tomczak et al., 2015). The glioma is a brain tumor that occurs in glial cells, a tissue in charge of protecting and nourishing the neurons (Purves et al., 2001). We used gene expression data of 19,947 genes obtained from 612 samples divided into four cancer cell types: 174 oligodendroglioma samples, 169 astrocytoma samples, 114 oligoastrocytoma samples, and 155 glioblastoma multiforme (GBM) samples. The tumor tissues have different degrees of aggressiveness. GBM is the most aggressive, while astrocytoma, oligodendrogliomas, and oligoastrocytoma are less aggressive than GBM (Louis et al., 2016). To approximate the genes expression levels distribution to a normal distribution, we transformed the values by their logarithm to the base two. For constructing the correlation networks, we performed Spearman's independence test between each pair of genes and inserted an edge for those whose p-value is smaller than 0.05. The absolute Spearman's correlation coefficient weights all edges.

The plant metabolism dataset contains 73 metabolites from whole-plant sorghum development (de Souza et al., 2015). The data were obtained from five organs (leaves, culm, roots, prop roots, and grains) of six biological samples. We consider correlation graphs in which the edges are weighted by Pearson's correlations >0.75, as used by Jeong et al. (2001) and Ding et al. (2015).

3. RESULTS AND DISCUSSION

To evaluate the performance of `BioNetStat`, we applied it on two datasets, namely glioma, and sorghum, and compared it to `GSCA`. The results for these comparisons are described in the following sections.

3.1. Analyses Using Glioma's Data Set

We performed Monte Carlo experiments to verify the ability of `BioNetStat` (based on the PDFG and the degree centrality) and `GSCA` to control the rate of false positives (control the proportion of type I error). We combined all 612 biological samples from four cancerous tissues (astrocytoma, oligoastrocytoma, oligodendroglioma, and GBM). For each test, we randomly selected, from a uniform distribution, 120 biological samples, and 50 genes to build each network. Thus, we consider that they come from the same dataset (i.e., under the null hypothesis). To analyze the results, we estimated the proportion of false positives to each p-value threshold. We analyzed the performance of the three methods (`BioNetStat` based on the PDFG and the degree centrality, and `GSCA`) when comparing five and ten networks (**Figures S1A,B**). Under the null hypothesis, we expect that the observed proportion of false positives is similar to the expected proportion set by the p-value threshold. In **Figure S1**, we observe that all methods indeed control the rate of false positives as expected.

To measure the statistical power (the ability to detect differences among two or more networks when indeed they are different) of the methods, we build r networks similarly to described in the previous paragraph. However, for one of the networks, we permuted the measurements of some gene expressions to change its co-expression pattern. The proportion of permuted genes is denoted by γ. In other words, for one of the networks we set $\gamma > 0$ (the network is different from the others) and $\gamma = 0$ for the others. Therefore, we expect that the methods detect that there is a different network. Then, to estimate the rate of false positives, we apply the tests in two networks selected from the $r - 1$ networks that are under the null hypothesis ($\gamma = 0$). Here, we expect to obtain a rate of false positives similar to the level of significance set by the p-value threshold. We carried out this experiment 1,000 times for different proportions of altered genes ($\gamma = 0.05, 0.1, 0.2, 0.3, 0.5$) and number of networks ($r = 2, 3, 5, 10, 15, 20$).

To summarize the statistical power of the test, we constructed Receiver Operating Characteristic (ROC) curves. The x and y axes of the ROC curves are the empirical false and true positive rates, respectively. The area under this ROC curve (AUC) summarizes the empirical power of the test. Under the alternative hypothesis (when at least one of the networks are generated by a different model), we expect that the proposed test present a ROC curve above the diagonal and consequently an AUC > 0.5.

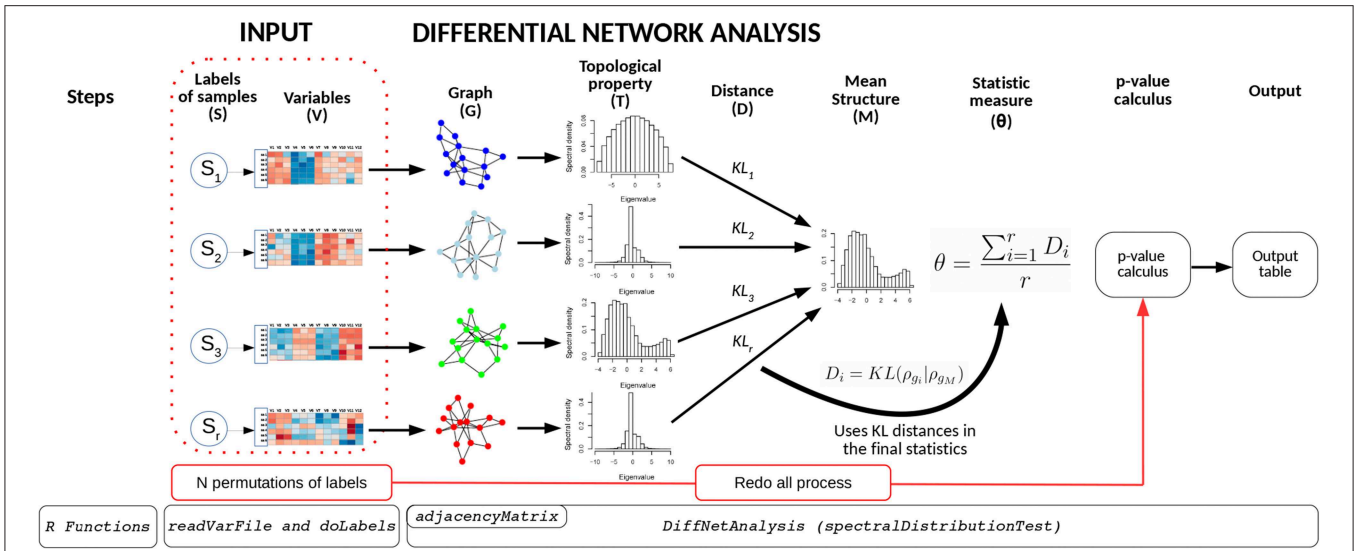


FIGURE 1 | Schematic diagram of BioNetStat. BioNetStat receives an input file containing the values of the variables to be analyzed and r biological states ($S_1, \ldots, S_r$). This figure illustrates the method performed with PDFG, however it can be replaced by centralities (such as Degree, Betweenness, and Closeness) without loss of generality.

In **Figures 2A,B**, we show the AUC when we compare five and ten biological states/networks (denoted by r), respectively, to $\gamma = 0.05, 0.1, 0.2, 0.3, 0.5$. In **Figures 2C,D**, we show the AUC for each $r = 2, 3, 5, 10, 15, 20$, and a fixed $\gamma = 0.1, 0.2$, respectively.

As expected, we observe in **Figures 2A,B** that both `BioNetStat` (based on PDFG and the degree centrality) and `GSCA` increase the statistical power proportionally to the increase of γ. Moreover, the performance of `BioNetStat` based on the PDFG presented lower power than `BioNetStat` based on the degree centrality and `GSCA` for $0.05 \leq \gamma \leq 0.2$ (**Figure 2A**). By comparing ten networks, we observe that the power of `GSCA` becomes lower than `BioNetStat` based on the degree centrality for $\gamma \geq 0.05$, and similar to `BioNetStat` based on PDFG for $\gamma \geq 0.2$ (**Figure 2B**).

We also observed that for a fixed γ, the empirical power decreases with the increase of the number of networks, as shown in **Figures 2C,D**. By comparing the performance of the methods, we observe that the empirical power of `GSCA` is greater than `BioNetStat` when the number of networks is small ($r = 2, 3$) and the changes in the networks are moderate ($\gamma = 0.1$) (**Figure 2C**). When the number of networks is five, the performance of `BioNetStat` based on the degree centrality is similar to `GSCA` for the two evaluated values of γ (**Figures 2C,D**). When the number of networks is >10 and $\gamma \geq 0.2$, the power of `BioNetStat` based on PDFG becomes greater than `GSCA`. Furthermore, we observe that the empirical power of `GSCA` decreases faster than `BioNetStat` with the increase of the number of networks.

Besides the statistical power, other criteria are relevant in the choice of the method to be used. In the following steps, we further analyze the glioma dataset.

We applied `BioNetStat` based on PDFG and `GSCA` in the glioma dataset comparing gene co-expression networks across the glioma types. We defined gene sets according to the canonical pathways obtained from Molecular signature Database Collection v5 (Subramanian et al., 2005). That database contains 1,329 canonical pathways. We performed the tests only with the subsets that presented at least 10 genes. Then, we tested the 1,289 gene sets.

We show the results of the tests, each one based on 1,000 permutation tests, for all gene sets in **Supplementary Data Sheet 1**. For the significance values (α) equal to 0.05 and 0.1, the total number of gene sets, which has at least one network statistically different from each other, were 490 and 801, respectively. One hundred and twenty-two, and 305 gene sets were co-identified by both methods considering a q-value of 0.05 and 0.1, respectively. For $\alpha = 0.05$ and $\alpha = 0.1$, `BioNetStat` identified, respectively, 62 and 79 gene sets that were not identified by `GSCA`. The latter identified, respectively, 306 and 417 gene sets that were not identified by `BioNetStat`. Thus, these results suggest that `BioNetStat` obtains results complementary to `GSCA`.

This complementarity is already expected, because `GSCA` and `BioNetStat` present different statistical tests. `GSCA` compares the Euclidean distances among matrices. It performs the pairwise comparison, edge by edge, being more sensitive to localized changes (few edges modifications) in networks, while `BioNetStat` is more adequate for differences spread across the correlation matrix. On the other hand, methods such as `CoGA` (Santos et al., 2015) and `GSCNA` (Rahmatallah et al., 2014) compare networks based on their overall structures, such as eigenvector centrality and spectral distributions. These strategies do not detect local changes in the network, since structural properties may remain unaffected. Rahmatallah et al. (2014) stated that `GSCNA` detects alterations when the major players such as genes of signaling pathways change

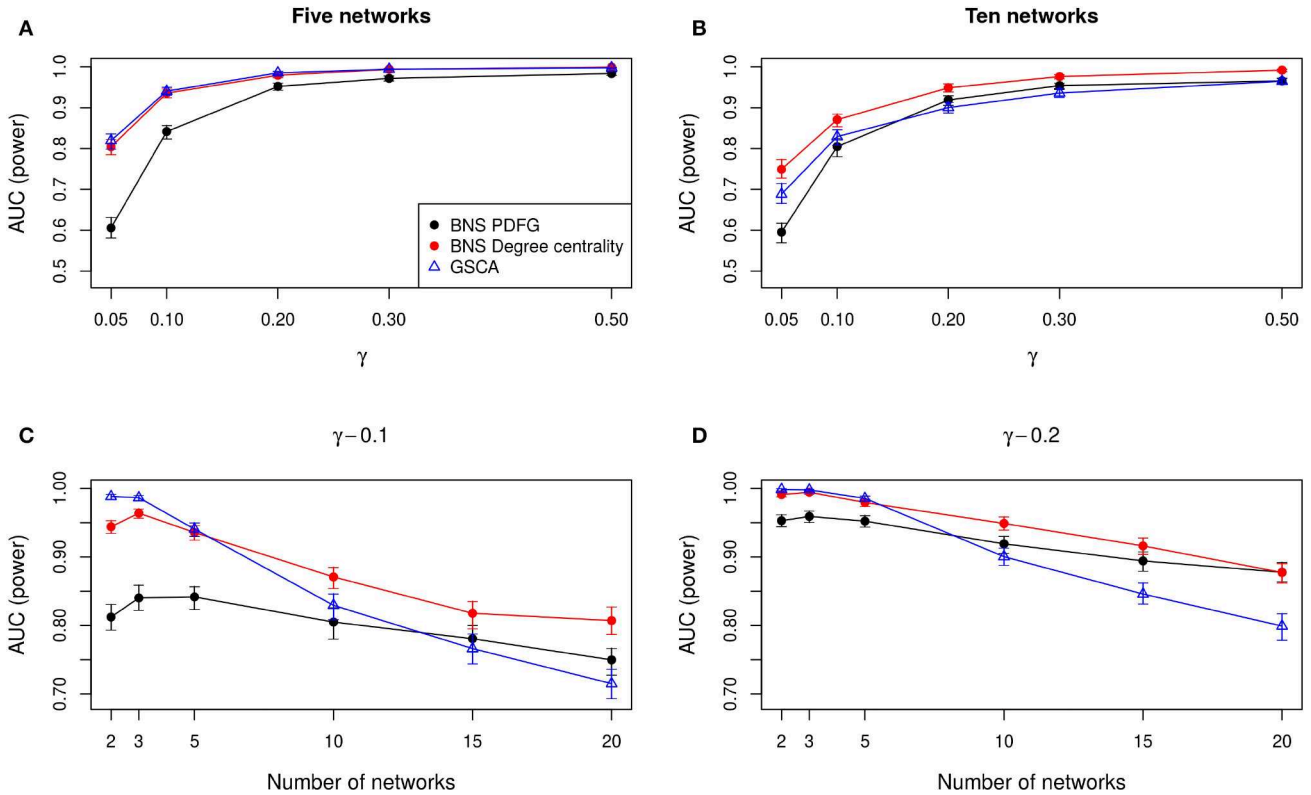


FIGURE 2 | Comparison of the statistical power of BioNetStat based on PDFG (black circles) and degree centrality (red circles), and GSCA (blue triangles). The values in the *y*-axis represent the areas under the ROC curves, considering the confidence interval of 95%. In **(A,B)**, the *x*-axis represents the ratio of perturbed genes (γ), for the comparison of 5 and 10 networks, respectively. In **(C,D)**, the *x*-axis represents the number of compared networks, which varies from two to 20, by fixing the γ to 0.1 and 0.2, respectively. Observe in **(A,B)** that both BioNetStat and GSCA statistical power increases proportionally to the increase of γ. We also observed that for a fixed γ, the empirical power decreases with the increase of the number of networks, as shown in **(C,D)**. Furthermore, we observe that the empirical power of GSCA decreases faster than BioNetStat with the increase of the number of networks.

across the different biological states, whereas `GSCA` detects these modifications when the average correlation changes (Rahmatallah et al., 2014), such as in pathways related to metabolism. As `BioNetStat` is based on topological features of the network, we expect that it would detect changes in signaling pathways rather than pathways related to metabolism.

To verify this hypothesis, we classified the 1,289 gene sets in *signaling* or *non signaling* pathways and compared the performance of `GSCA` against `BioNetStat`. To classify as *signaling* pathway, we searched for key terms in gene sets such as "signal," "cascade," "receptor," "activ*," "regula*," "pid," "ach," "arrestin," and the transcription factor names obtained from MsigDB website. The proportion of signaling pathways in the 1,289 gene sets is 51.2%. Only the gene sets selected by each method for a *q*-value threshold at 0.05 were considered. Our test classified 52.8% of the selected gene sets by `GSCA` as signaling pathways. Whereas, for `BioNetStat`, the test selected 59.2% out of 184 gene sets as signaling pathways. We performed the proportion method (`prop.test` R function), considering the null hypothesis that measured proportion is equal to 51.2% and the alternative that the measured proportion is greater than 51.2%. Only `BioNetStat` presented a proportion of signaling pathways statistically greater than the entire dataset ($p = 0.018$), whereas `GSCA` did not ($p = 0.269$). Therefore, as expected, `BioNetStat` detects more changes in signaling pathways than `GSCA`.

To highlight the applicability of the proposed method, we went deeper in the analysis of the 62 gene sets that were detected by `BioNetStat`, but not by `GSCA`, considering a *q*-value threshold at 0.05. Among this 62 differentially coexpressed gene sets, 38 were classified as signaling pathways. We searched for a gene set that contained NFκB gene, a transcription factor which controls more than a hundred of genes, well-known to be associated with glioma's formation (Mieczkowski et al., 2015; Kinker et al., 2016; Ferrandez et al., 2018). Then, we selected "KEGG TOLL-LIKE RECEPTOR SIGNALING PATHWAY." Also, Toll-like receptors (TLRs) is an important gene set, part of a signaling pathway gene set associated with gliomas (Ferrandez et al., 2018). TLRs are membrane-bound receptors, which serve as crucial pattern recognition receptors with central roles in the induction of innate immune responses (Kawai and

Akira, 2007). Pathogen recognition by TLRs provokes rapid activation of innate immunity by inducing production of proinflammatory cytokines and upregulation of costimulatory molecules (Ferrandez et al., 2018). Therefore, the TLR genes trigger a signaling chain reaction that leads to NFκB activation which, in turn, triggers inflammatory responses (Kawai and Akira, 2007).

Our analyses suggested that at least one network is different from the others in the TLR gene set. Then, we performed a pairwise comparison of the four cancer types to understand better how they differ from each other. **Figure 3** presents the dendrogram obtained by calculating the pairwise Jensen-Shannon divergence (a symmetric version of KL divergence to pairwise comparison) between the networks. We expected that the most aggressive cancer type, namely GBM, be in one branch and the other three types, on another branch. However, the cancer types GBM and oligoastrocytoma are in one branch and oligodendroglioma, and astrocytoma are in another branch. The unexpected closeness between GBM and oligoastrocytoma could be a consequence of a confusing clinic classification method of gliomas subtypes. The TCGA database classifies gliomas only into four types *astrocytoma*, *oligoastrocytoma*, *oligodendroglioma*, and GBM. However, there is a more aggressive type of *oligoastrocytoma*, called *anaplastic oligoastrocytoma*, that can also be classified as a glioblastoma with an oligodendroglial component (Nakamura et al., 2011). Since 2007, the World Health Organization (WHO) defines the *anaplastic oligoastrocytoma* as a Glioblastoma (Marucci, 2011). Therefore, there must be intermediate states between both types (Oligoastrocytoma and GBM), not discriminated in our data, that explain this closeness between them.

`BioNetStat` also allows us to identify in which node the connections change significantly by the *differential node analysis*. We performed this analysis by using the degree centrality. The *TLR signaling pathway* presented statistically significant changes of nodes degree centrality ($\theta = 2.88$ and $p = 0.027$). In this gene set, eight genes presented their degree centrality significantly altered (**Table 1**). Three of them are mitogen-activated protein kinase MAPK (3, 9, and 10) and are integrated into the RAS/MAPK signaling pathway. When RAS (Rat Sarcoma) genes are active, they regulate the MAPK pathway and vital processes into the cell, such as proliferation, differentiation, signal transduction, apoptosis, and tumorigenesis (Mao et al., 2013). Modifications in this pathway could lead to abnormal function of these processes. As an example, the overexpression of RAS was detected in astrocytoma and GBM (Mao et al., 2013). Other three genes differentially coexpressed are in the PIK3-PTEN-Akt-mTOR pathway. The genes PIK3 indirectly activates Akt which, in turn, activates mTOR (mammalian target of rapamycin). This gene cascade leads to an integration of upstream signals into effector actions, controlling multiple downstream targets involved in cell growth and division. Most of the genes differently coexpressed such as MAPKs, PIK3s, and AKT3 are involved into the gliomas formation (the PIK3 pathway is altered in about 70% of GBMs) (Mao et al., 2013), demonstrating the importance of gene set detected by `BioNetStat`.

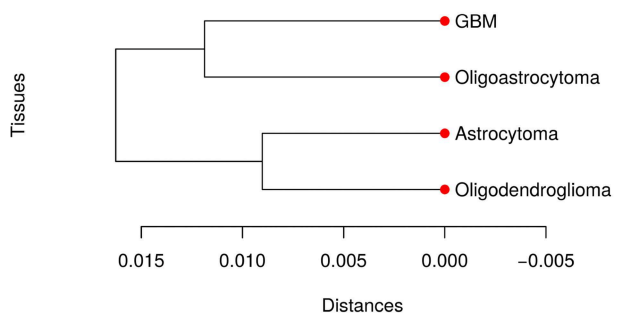


FIGURE 3 | Dendrogram of the distances among the four glioma subtypes regarding the *Toll like receptor signaling pathway*. The unexpected closeness between Oligoatrocytoma and GBM is probably due to the specific features in tissues characterizations.

3.2. Analyses Using *Sorghum bicolor*'s Data Set

In the second data set, we studied how the metabolic networks of five plant organs differ from each other. The 73 metabolites analyzed in sorghum organs (leaf, culm, root, prop root, and grains) were partitioned in five groups according to their biochemical roles: carbohydrates, amino acids, organic acids, nucleotides, and all 73 metabolites. We built one network for each organ and each metabolic group. Then we compared the networks across the organs using the PDFG, the centrality tests of `BioNetStat`, and `GSCA` method.

The grain-filling stage in plants is largely dependent on metabolic status (Schnyder, 1993). Thus, it is important to understand to what extent the metabolic networks in distinct organs differ from each other. de Souza et al. (2015) investigated whether each organ performs a specific role in plant metabolism during the grain-filling in sorghum plants. Here, we complemented their study by analyzing the same dataset based on a systemic point of view and network modeling. First, we investigated if the PDFG and degree centralities are different among the networks (organs). **Table 2** presents the results of PDFG, degree centrality, and `GSCA` tests. Comparing the metabolic networks structures, through their PDFG, it can be observed that at least one organ is different from the others, regarding the *all metabolites* and the *carbohydrates* set. According to the degree centrality analysis, the organs networks are significantly different in the five metabolites sets. `GSCA` detected the *organic acids* and the *nucleotides* sets as differentially coexpressed. Analyzing the concentrations of metabolites, de Souza et al. (2015) also found differences among organs in the four metabolites sets.

We obtained pairwise distances among the organ networks for those metabolic sets with a statistically significant difference.

TABLE 2 | Results of the PDFG and degree centrality statistical tests comparing all five organs networks.

		PDFG			Degree centrality			GSCA		
Name	**Size**	**θ Statistic**	***p*-value**	***q*-value**	**θ Statistic**	***p*-value**	***q*-value**	**θ Statistic**	***p*-value**	***q*-value**
All	73	0.017	**0.006**	**0.015**	17.167	**0.001**	**0.002**	0.329	**0.006**	**0.042**
Carbohydrate	18	0.056	**0.003**	**0.015**	3.857	**0.001**	**0.002**	0.299	0.416	0.416
Organic acid	13	0.044	0.065	0.108	3.482	**0.001**	**0.002**	0.341	**0.019**	**0.044**
Amino acid	24	0.018	0.292	0.312	5.152	**0.003**	**0.004**	0.314	0.179	0.313
Nucleotide	12	0.034	0.312	0.312	3.041	**0.006**	**0.006**	0.352	**0.019**	**0.044**

The q-values < 0.05 are in bold.

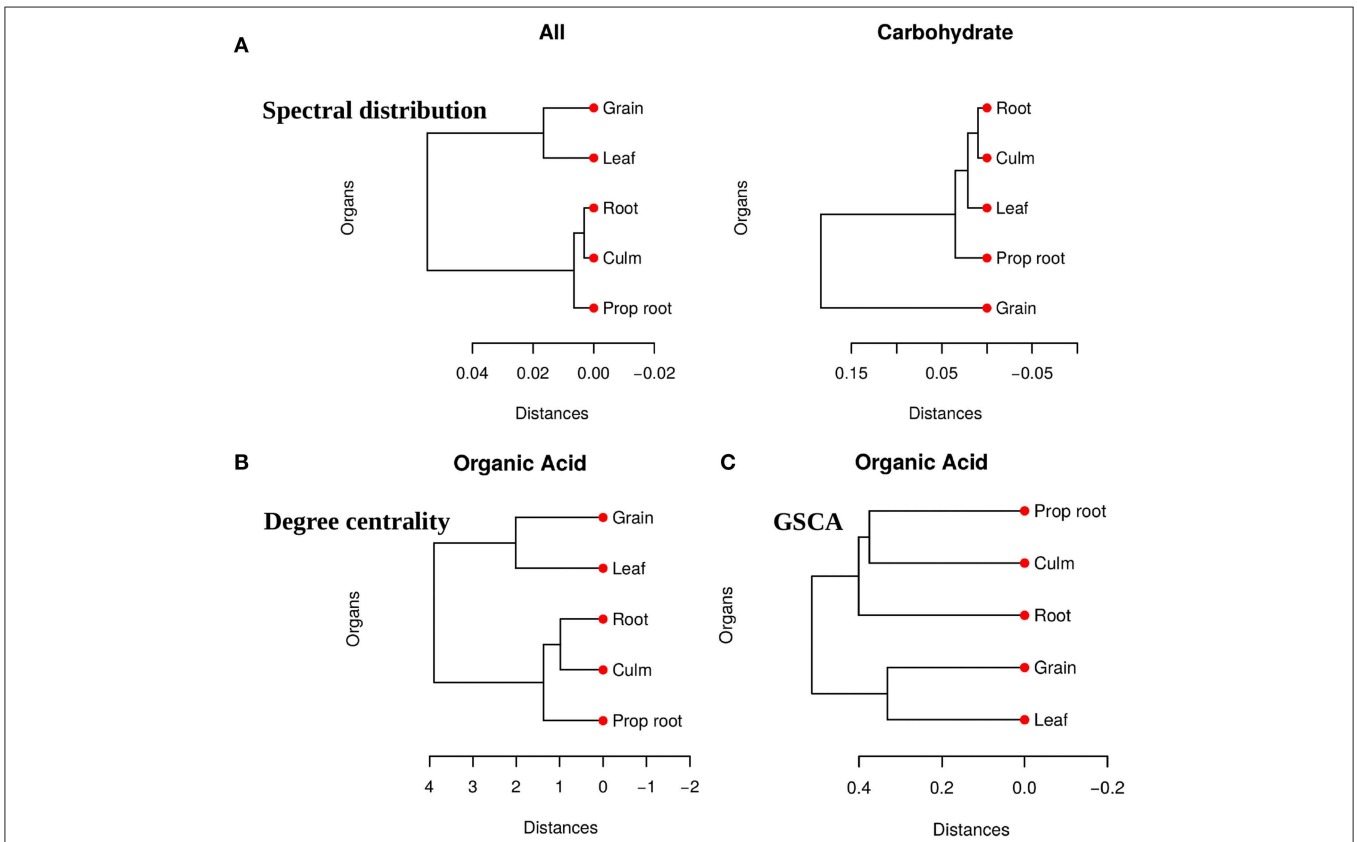


FIGURE 4 | (A) Dendrogram of distances among five network PDFG for the groups of *all metabolites* and *carbohydrates*. **(B)** Dendrogram of distances among five network degree cetrality of the organs for the *organic acid* metabolites. **(C)** Dendrogram of distances among five network of the organs for the *organic acid* metabolites measured by GSCA. We can conclude that the grain network is more distant network from the others, reinforcing that this organ has a different metabolism.

Figure 4A shows the distances among networks according to Jensen-Shannon divergence. Considering the *all metabolites* network, the grain is significantly different from the culm, prop root, and roots (**Table S1**). Additionally, according to the *carbohydrate* results, the metabolism of the grains is different from all other organs (**Table S2**). The results suggest that the grain has a specific metabolic structure and that the leaf network is more similar to the grain network than to the other organs. Considering that the grain is the main sink of the plant during the grain-filling (period of the experimental harvest of the studied data) (de Souza et al., 2015), we expected that its metabolism to be different from other organs.

For the tests performed with the degree centrality, we identified significant differences in all groups. The results suggest that even if the network structure (PDFG) does not change, the role of the metabolites and its mean correlation values in each organ can be different. The *organic acid*, identified by `BioNetStat` degree centrality network and `GSCA` can exemplify this phenomenon. According to both methods, the grain and leaf networks are the most distant (**Figures 4B,C**) and statistically different from the remaining organs (**Tables S3, S4**).

TABLE 3 | Differential node analysis based on degree centrality.

Metabolite	θ Statistic	*p*-value	*q*-value	Degree centrality				
				Leaf	Culm	Root	Prop root	Grain
Piruvate	4.219	0.001	0.003	10.328	1.906	0.765	0.821	9.579
Mevalonate	4.215	0.001	0.003	9.93	0	0.838	0	8.191
cis-Aconitate	3.582	0.001	0.003	9.361	0	0.872	0.821	6.693
AKG	3.499	0.001	0.003	9.474	2.641	0.872	5.11	10.854
2/3PGA	3.862	0.001	0.003	10.412	0.805	5.216	0	9.695
Chiquimate	3.523	0.002	0.004	8.97	0	0.913	2.517	7.994
Malate	3.029	0.003	0.005	10.374	1.917	0.765	5.206	7.376
Isocitrate	2.782	0.003	0.005	9.588	1.872	4.654	5.316	9.898
Citrate	2.631	0.009	0.013	10.019	1.857	2.702	6.104	7.156
PEP	2.205	0.064	0.083	9.393	1.863	3.583	5.111	2.529
Fumarate	2.387	0.089	0.105	4.018	1.845	3.505	0.829	10.01
trans-Aconitate	2.712	0.097	0.105	0	1.842	0.913	3.304	9.835
Succinate	2.115	0.141	0.141	8.871	1.84	4.567	6.22	7.738

Statistical tests results of comparison among five organs on the 13 organic acids. The table shows the metabolite's name, the θ statistic, the nominal p-value, the adjusted p-value for tests performed (q-value), and the degree centrality of each node in the five organs.

For this reason, we investigated which nodes changed the degree centrality value in the *organic acids* network among the organs (**Table 3**), by performing the *differential node analysis*. The `GSCA` has not implemented a similar method capable of comparing whether the importance of the nodes changes among states. Therefore, we forwarded the analysis using only `BioNetStat`.

The majority of the metabolites of the *organic acids* dataset belong to the citrate cycle (or Krebs cycle), a chain of reactions that transfer energy (by electrons transferring) from complete pyruvate oxidation to cofactors used in ATP production (Siedow and Day, 2000). The network of the *organic acid* is more connected in the leaf and grain than in the culm, prop root and root. The average degree centrality in the leaf and grain is 8.51 and 8.27, respectively, whereas in the culm, prop root, and root networks the average degree centrality is 1.41, 3.18, and 2.32, respectively (extracted from **Table 3**). The metabolites with highest degree centrality in the leaf and in the grain are the pyruvate and the AKG (α-ketoglutarate), respectively (**Table 3**). These results are in agreement with previous observations by de Souza et al. (2015) that pointed out pyruvate as a central molecule in metabolism, connecting the citrate cycle with many other pathways. Our network analysis using `BioNetStat` revealed that the AKG is also a relevant metabolite, being a precursor of many amino acids synthesis pathways (**Figure 5**) (Siedow and Day, 2000).

The analyzed data were collected between 10 a.m. and 12 a.m. when the leaf performs constant photosynthesis and mobilization of carbon. Also, the grain metabolism is geared toward storage of carbohydrate and proteins. Therefore, we have evidence to believe that the average degree centrality of metabolites are higher in the leaf and grain networks because the organic acid metabolism of these organs is more active than the organic acid metabolism of the other organs. Our findings reinforce that network analysis brings a new view to the data, since de Souza et al. (2015) did not find these molecules in comparisons among organ metabolisms, as highly concentrated in these organs. Furthermore, to highlight relevant variables in the system, `BioNetStat` performs the *differential node analysis*, a method not available in other tools considered in this work.

4. CONCLUSION

`BioNetStat` is a network analysis `Bioconductor` package, containing a Graphical User Interface, that allows the comparison of two or more correlations networks. The proposed method is an adaptation and generalization of `CoGA`, which aims to meet demand on multistate experiments. We show here that `BioNetStat` performs the *differential network analysis*, exploring networks features and highlighting the main differences among states. Moreover, it carries out statistical tests to estimate the significance of the results. We showed that all the statistical tests performed by `BioNetStat` effectively control the rate of false positives. Our simulation experiments and applications in real datasets suggest that `BioNetStat` complements and advances previous tools (`CoGA` and `GSNCA`) for differential co-expression analysis, i.e., `BioNetStat` allows the comparison of more than two networks simultaneously. We also conclude that `BioNetStat` is less sensitive to the increase in the number of networks than `GSCA`. Furthermore, it is able to identify more gene sets associated with important signaling pathways than `GSCA`, and also highlights key genes in the networks (centrality analyses). The study cases show that `BioNetStat` helps to find differences beyond the analysis of the network, highlighting features that can be biologically supported while undetected by in orthodox analyses. `BioNetStat` provides numerical results combined with visual inspection in the graphical user interface that might be helpful

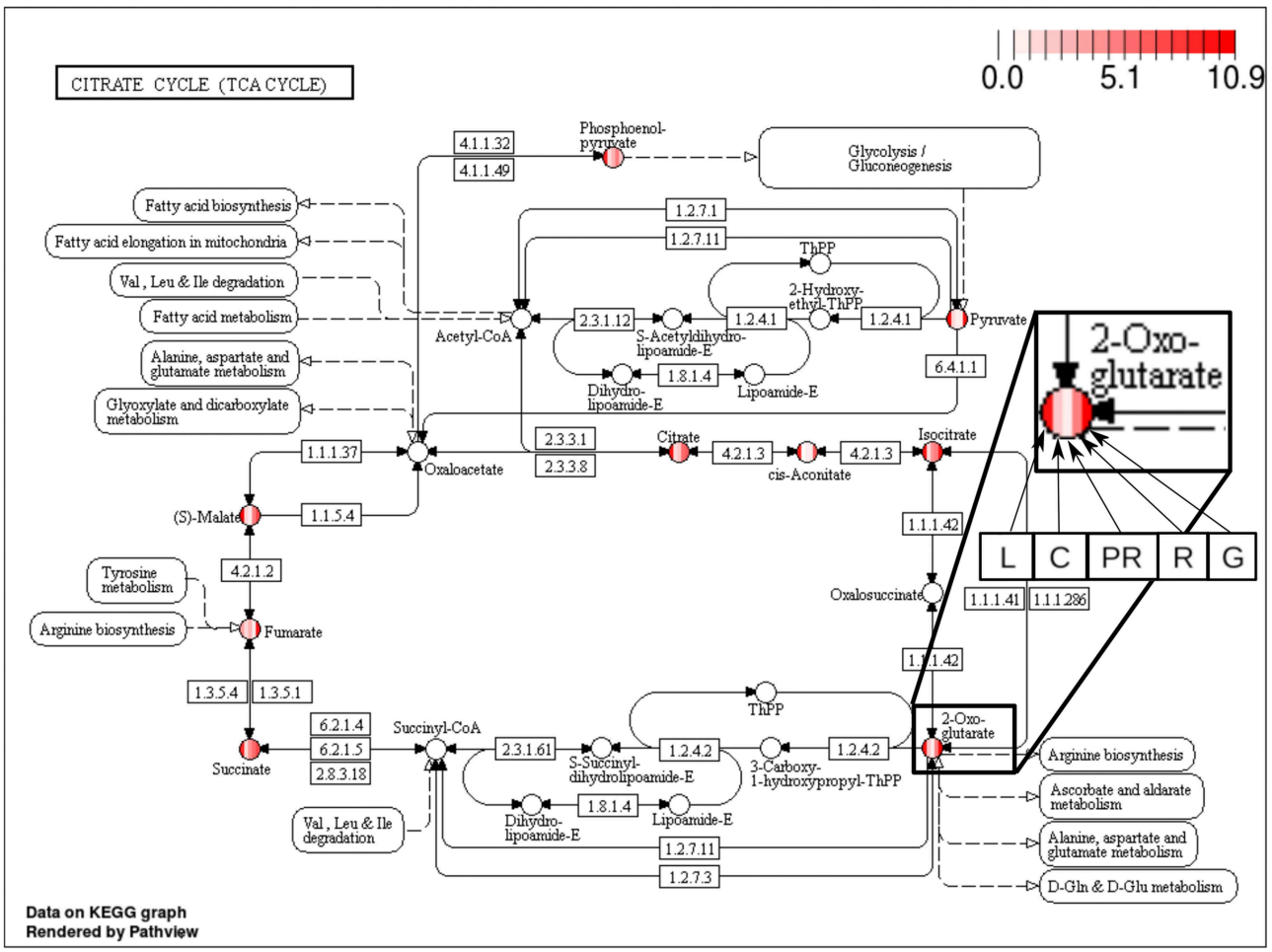


FIGURE 5 | Citrate cycle (TCA Cycle) metabolic pathway from KEGG database (Kanehisa and Goto, 2000). Each metabolite, represented by the circles, is partitioned in five columns, in which each one represents one of the organs, leaf (L), culm (C), prop root (PR), root (R), and grains (G), from the left to right, as shown in the highlighted node of 2-Oxoglutarate (AKG) at the right of the figure. The color of the columns represents the degree centrality value of that metabolite in the organ network. The values of degree centrality in this figure vary from 0 to 10.9.

in the identification of critical elements of the analyzed system. `BioNetStat` is not restricted to analyses of genes coexpression networks. Differently from other tools, `BioNetStat` can be used with different types of data sets such as the ones generated by metabolomics, proteomics, phenomics, and possibly social and economic data.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov.

AUTHOR CONTRIBUTIONS

VJ, SS, AF, and MB conceived and designed the experiments, analyzed the data, and wrote the paper. VJ performed the experiments.

FUNDING

This work has been financed by the Microsoft Research-FAPESP Institute (FAPESP 2011/52065) and the National Institute of Science and Technology of Bioethanol (INCT Bioetanol—FAPESP, grant no. 2008/57908-6; 2014/50884-5, CNPq, grant no. 574002/2008-1; 465319/2014-9). Moreover, AF was partially supported by FAPESP (2016/13422-9; 2018/21934-5), CNPq (304876/2016-0). SS was partially supported by FAPESP (2015/21162-4) fellowship. VJ was partially supported by CAPES fellowship. CAPES (Finance Code 001).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00594/full#supplementary-material

REFERENCES

Barabasi, A.-L., and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science* 286, 509–512. doi: 10.1126/science.286.5439.509

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bonacich, P. (1972). Bonacich_1972_technique for analyzing overlapping memberships. *Sociol. Methodol.* 4, 176–185. doi: 10.2307/270732

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., et al. (2011). High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. *Plant J.* 67, 869–884. doi: 10.1111/j.1365-313X.2011.04640.x

Cho, S. B., Kim, J., and Kim, J. H. (2009). Identifying set-wise differential co-expression in gene expression microarray data. *BMC Bioinformatics* 10:109. doi: 10.1186/1471-2105-10-109

Choi, Y., and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics* 25, 2780–2786. doi: 10.1093/bioinformatics/btp502

Dawson, J. A., Ye, S., and Kendziorski, C. (2012). R/ebcoexpress: an empirical bayesian framework for discovering differential co-expression. *Bioinformatics* 28, 1939–1940. doi: 10.1093/bioinformatics/bts268

de Souza, A. P., Cocuron, J.-C., Garcia, A. C., Alonso, A. P., and Buckeridge, M. S. (2015). Changes in whole-plant metabolism during grain-filling stage in *Sorghum bicolor* L. (moench) grown under elevated Co2 and drought. *Plant Physiol.* 169, 1755–1765. doi: 10.1104/pp.15.01054

de Souza, A. P., Gaspar, M., Da Silva, E. A., Ulian, E. C., Waclawovsky, A. J., Nishiyama, M. Y., et al. (2008). Elevated CO2 increases photosynthesis, biomass and productivity, and modifies gene expression in sugarcane. *Plant Cell Environ.* 31, 1116–1127. doi: 10.1111/j.1365-3040.2008.01822.x

Ding, Y., Chang, J., Ma, Q., Chen, L., Liu, S., Jin, S., et al. (2015). Network analysis of postharvest senescence process in citrus fruits revealed by transcriptomic and metabolomic profiling. *Plant Physiol.* 168, 357–376. doi: 10.1104/pp.114.255711

Ferrandez, E., Gutierrez, O., Segundo, D. S., and Fernandez-luna, J. L. (2018). NFκB activation in differentiating glioblastoma stem-like cells is promoted by hyaluronic acid signaling through TLR4. *Sci. Rep.* 8:6341. doi: 10.1038/s41598-018-24444-6

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Soc. Netw.* 1, 215–239. doi: 10.1016/0378-8733(78)90021-7

Fujita, A., Vidal, M. C., and Takahashi, D. Y. (2017). A statistical method to distinguish functional brain networks. *Front. Neurosci.* 11:66. doi: 10.3389/fnins.2017.00066

Fukushima, A. (2013). DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* 518, 209–214. doi: 10.1016/j.gene.2012.11.028

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Gysi, D. M., Fragoso, T. D. M., Almaas, E., Nowick, K., Intelligence, S., Group, C. S., et al. (2018). CoDiNA: an R package for co-expression differential network analysis in n dimensions. *arXiv [Preprint]. arXiv:1802.00828*.

Ha, M. J., Baladandayuthapani, V., and Do, K. A. (2015). DINGO: Differential network analysis in genomics. *Bioinformatics* 31, 3413–3420. doi: 10.1093/bioinformatics/btv406

Hochberg, U., Degu, A., Toubiana, D., Gendler, T., Nikoloski, Z., Rachmilevitch, S., et al. (2013). Metabolite profiling and network analysis reveal coordinated changes in grapevine water stress response. *BMC Plant Biol.* 13:184. doi: 10.1186/1471-2229-13-184

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* 407, 651–654. doi: 10.1038/35036627

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

Kawai, T., and Akira, S. (2007). TLR signaling. *Semin. Immunol.* 19, 24–32. doi: 10.1016/j.smim.2006.12.004

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.1093/biomet/30.1-2.81

Kinker, G. S., Thomas, A. M., Carvalho, V. J., Lima, F. P., and Fujita, A. (2016). Deletion and low expression of NFKBIA are associated with poor prognosis in lower-grade glioma patients. *Sci. Rep.* 6:24160. doi: 10.1038/srep24160

Li, J., Li, Y. X., and Li, Y. Y. (2016). Differential regulatory analysis based on coexpression network in cancer research. *BioMed Res. Int.* 2016, 1–8. doi: 10.1155/2016/4241293

Liu, B. H., Yu, H., Tu, K., Li, C., Li, Y. X., and Li, Y. Y. (2010). DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data. *Bioinformatics* 26, 2637–2638. doi: 10.1093/bioinformatics/btq471

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Lui, T. W. H., Tsui, N. B. Y., Chan, L. W. C., Wong, C. S. C., Siu, P. M. F., and Yung, B. Y. M. (2015). DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. *BMC Bioinformatics* 16:182. doi: 10.1186/s12859-015-0582-4

Mao, H., LeBrun, D. G., Yang, J., Zhu, V. F., and Li, M. (2013). Deregulated signaling pathways in glioblastoma multiforme: molecular mechanisms and therapeutic targets. *Cancer Invest.* 30, 48–56. doi: 10.3109/07357907.2011.630050

Marucci, G. (2011). The effect of WHO reclassification of necrotic anaplastic oligoastrocytomas on incidence and survival in glioblastoma. *J. Neuro-Oncol.* 104, 621–622. doi: 10.1007/s11060-010-0523-z

McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M., and Zhang, B. (2016). DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Syst. Biol.* 10:106. doi: 10.1186/s12918-016-0349-1

Mieczkowski, J., Kocyk, M., Nauman, P., Gabrusiewicz, K., Sielska, M., Przanowski, P., et al. (2015). Down-regulation of IKKβ expression in glioma-infiltrating microglia/macrophages is associated with defective inflammatory/immune gene responses in glioblastoma. *Oncotarget* 6, 33077–33090. doi: 10.18632/oncotarget.5310

Nakamura, H., Makino, K., and Kuratsu, J. I. (2011). Molecular and clinical analysis of glioblastoma with an oligodendroglial component (GBMO). *Brain Tumor Pathol.* 28, 185–190. doi: 10.1007/s10014-011-0039-z

Pearson, K. (1920). Notes on the history of correlation. *Biometrika* 2, 209–213. doi: 10.1093/biomet/13.1.25

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., et al. (2001). *Neuroscience, 2nd Edn*. Sunderland, MA: Sinauer Associates.

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics* 30, 360–368. doi: 10.1093/bioinformatics/btt687

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Santos, S. D. S., De Almeida Galatro, T. F., Watanabe, R. A., Oba-Shinjo, S. M., Marie, S. K. N., and Fujita, A. (2015). CoGA: An R package to identify differentially co-expressed gene sets by analyzing the graph spectra. *PLoS ONE* 10:e0135831. doi: 10.1371/journal.pone.0135831

Schnyder, H. (1993). The role of carbohydrate storage and redistribution in the source-sink relations of wheat and barley during grain filling -a review. *Nezc Phytol* 123, 233–245. doi: 10.1111/j.1469-8137.1993.tb03731.x

Siedow, J. N., and Day, D. A. (2000). "Chapter 14: Respiration and photorespiration," in *Biochemistry & Molecular Biology of Plants, 1st Edn*, eds B. B. Buchanan, W. Gruissem, and R. L. Jones (Rockville, MD: American Society of Plant Physiologists), 676–729.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis, Vol. 26*. London, UK: CRC Press.

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101.

Sturges, H. A. (1926). The choice of a class interval. *J. Am. Stat. Assoc.* 21, 65–66. doi: 10.1080/01621459.1926.10502161

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

Sun, S.-Y., Liu, Z.-P., Zeng, T., Wang, Y., and Chen, L. (2013). Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. *Sci. Rep.* 3:2268. doi: 10.1038/srep02268

Takahashi, D. Y., Sato, J. R., Ferreira, C. E., and Fujita, A. (2012). Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS ONE* 7:e49949. doi: 10.1371/journal.pone.0049949

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* 11:497. doi: 10.1186/1471-2105-11-497

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Wspolczesna Onkol.* 1A, A68–A77. doi: 10.5114/wo.2014.47136

Watson, M. (2006). CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics* 7:509. doi: 10.1186/1471-2105-7-509

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Weston, D. J., Karve, A. A., Gunter, L. E., Jawdy, S. S., Yang, X., Allen, S. M., et al. (2011). Comparative physiology and transcriptional networks underlying the heat shock response in *Populus trichocarpa*, *Arabidopsis thaliana* and *Glycine max*. *Plant Cell Environ.* 34, 1488–1506. doi: 10.1111/j.1365-3040.2011.02347.x

Wu, S., Alseekh, S., Cuadros-Inostroza, A., Fusari, C. M., Mutwil, M., Kooke, R., et al. (2016). Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in *Arabidopsis thaliana*. *PLoS Genet.* 12:e1006363. doi: 10.1371/journal.pgen.1006363

Zhang, H., and Yin, T. (2016). Identifying candidate genes for wood formation in poplar based on microarray network analysis and graph theory. *Tree Genet. Genomes* 12:61. doi: 10.1007/s11295-016-1016-9

Zhu, J., Zhang, B., and Schadt, E. E. (2008). A systems biology approach to drug discovery. *Adv. Genet.* 60, 603–635. doi: 10.1016/S0065-2660(07)00421-X

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

The handling editor declared a shared affiliation, though no other collaboration, with the authors at the time of the review.

frontiers in Genetics

TECHNOLOGY REPORT
published: 28 June 2019
doi: 10.3389/fgene.2019.00614

Pipeliner: A Nextflow-Based Framework for the Definition of Sequencing Data Processing Pipelines

Anthony Federico[1,2], Tanya Karagiannis[1], Kritika Karri[1], Dileep Kishore[1], Yusuke Koga[2], Joshua D. Campbell[1,2] and Stefano Monti[1,2]**

[1] Bioinformatics Program, Boston University, Boston, MA, United States, [2] Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA, United States

The advent of high-throughput sequencing technologies has led to the need for flexible and user-friendly data preprocessing platforms. The Pipeliner framework provides an out-of-the-box solution for processing various types of sequencing data. It combines the Nextflow scripting language and Anaconda package manager to generate modular computational workflows. We have used Pipeliner to create several pipelines for sequencing data processing including bulk RNA-sequencing (RNA-seq), single-cell RNA-seq, as well as digital gene expression data. This report highlights the design methodology behind Pipeliner that enables the development of highly flexible and reproducible pipelines that are easy to extend and maintain on multiple computing environments. We also provide a quick start user guide demonstrating how to setup and execute available pipelines with toy datasets.

Keywords: pipeline development, sequencing workflows, Nextflow, RNA-seq pipeline, scRNA-seq pipeline

OPEN ACCESS

Edited by:
Vinicius Maracaja-Coutinho, Universidad de Chile, Chile

Reviewed by:
Pao-Yang Chen, Academia Sinica, Taiwan
Ernesto Picardi, University of Bari Aldo Moro, Italy

****Correspondence:***
Anthony Federico
anfed@bu.edu
Stefano Monti
smonti@BU.EDU

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

Received: *21 November 2018*
Accepted: *13 June 2019*
Published: *28 June 2019*

Citation:
Federico A, Karagiannis T, Karri K, Kishore D, Koga Y,Campbell JD and Monti S (2019) Pipeliner: A Nextflow-Based Framework for the Definition of Sequencing Data Processing Pipelines. Front. Genet. 10:614. doi: 10.3389/fgene.2019.00614

INTRODUCTION

High-throughput sequencing (HTS) technologies are vital to the study of genomics and related fields. Breakthroughs in cost efficiency have made it common for studies to obtain millions of raw sequencing reads. However, processing these data requires a series of computationally intensive tools that can be unintuitive to use, difficult to combine into stable workflows that can handle large number of samples, and challenging to maintain over long periods of time in different environments. The effort to simplify this process has resulted in the development of sequencing pipelines such as RseqFlow (Wang et al., 2011), PRADA (Torres-García et al., 2014), and Galaxy (Goecks et al., 2010), among others. Some of these pipelines are open-source and either available for download or on publicly available servers. However, some drawbacks include difficulty when deploying on existing computational resources, limited selection of computational tools, and unintuitive or limited ability to make modifications. While other frameworks may be more flexible, they often require the user to install each needed tool separately, which may be challenging and reduce reproducibility.

Pipeliner is a framework for the definition of sequencing data processing pipelines that aims to solve these issues. Pipelines developed within the framework are platform independent and fully reproducible and inherit automated job parallelization and failure recovery. Their flexibility and modular architecture allows users to easily customize and modify processes

based on their needs. Pipeliner also provides additional resources that allow developers to rapidly build and test their own pipelines in an efficient and scalable manner. Pipeliner is a complete and user-friendly solution to meet the demands of processing large amounts and various types of sequencing data.

MATERIALS AND METHODS

Design and Features

Pipeliner is a suite of tools and methods for defining sequencing pipelines. It uses Nextflow, a portable, scalable, and parallelizable domain-specific language, to define data workflows (Di Tommaso et al., 2017). Using Nextflow, each pipeline is modularized, consisting of a configuration file as well as a series of processes. These processes define the major steps in each pipeline and can be written in Linux-executable scripting languages such as Bash, Python, Ruby, etc. Nextflow processes are connected through channels—asynchronous first in, first out queues—which allow data to be passed between the different steps in each pipeline using a dataflow programming model. Using this architecture, pipelines developed within the Pipeliner framework inherit multiple features that contribute to their flexibility, reproducibility, and extensibility (**Figure 1**).

Pipeline Flexibility

Pipeliner enables flexible customization of pipeline options and parameters. Pipeliner currently offers three pipelines to demonstrate its applicability in processing different types of data, including bulk RNA-seq, single-cell RNA-seq (scRNA-seq), as well as digital gene expression (DGE) data (Soumillon et al., 2014). For the RNA-seq pipeline, sequencing reads are checked for quality with FastQC (Andrews, 2010), trimmed with TrimGalore (Krueger, 2016), mapped to a reference genome with either STAR (Dobin et al., 2012) or HISAT2 (Kim et al., 2015), and quantified with either StringTie (Pertea et al., 2015), HTSeq (Anders et al., 2015), or featureCounts (Liao et al., 2014). After alignment, mapping quality is checked with RSeQC (Wang et al., 2012), and a comprehensive summary report of all processes is generated with MultiQC (Ewels et al., 2016). The scRNA-seq and DGE pipelines adopt a similar methodology, and the development of additional pipelines for microRNA-seq (miRNA-seq) and RNA-seq Variant Calling is currently underway.

Parameter Configuration

All pipeline options and process parameters are set from a single configuration file (**Figure 2**). Users have the option to select and skip various steps as well as customize parameters and allocate computing resources for specific processes. This flexibility gives rise to many different use cases. For example, a user may opt to provide a pre-indexed reference genome or start the pipeline after the mapping step with saved alignment files or output an ExpressionSet data structure with count and phenotypic data. Thus, each pipeline is multipurpose and allows users to frequently tweak settings without adding complexity or sacrificing reproducibility.

The default configuration file defines variables for common parameters of third-party software tools used in each pipeline. These tools are wrapped into templates—one for

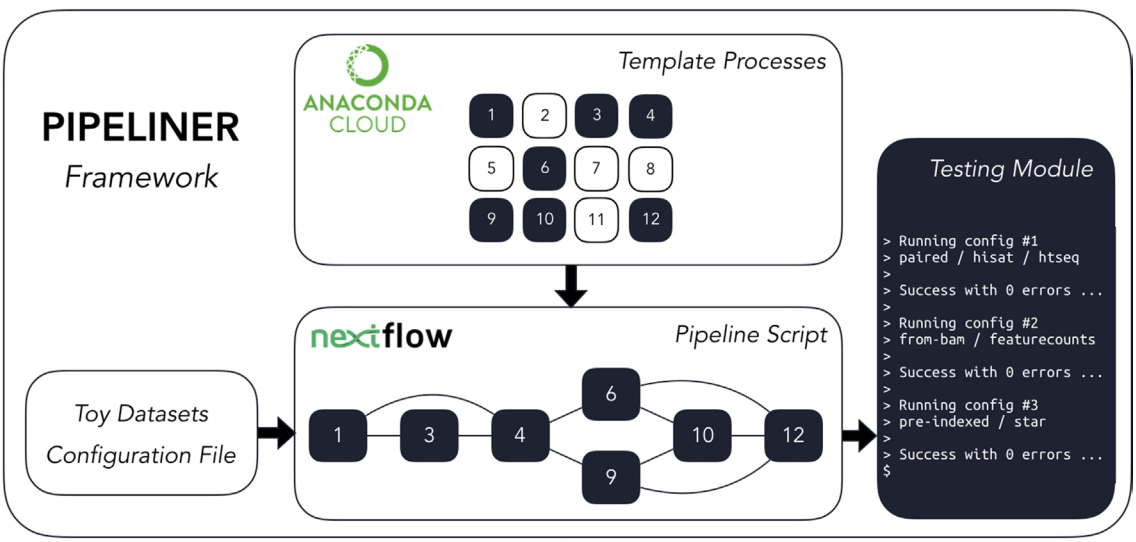


FIGURE 1 | The Pipeliner framework employs reusable template processes strung together *via* Nextflow's scripting language to create workflows in addition to developer tools such as toy datasets and testing modules.

```
process {
  executor = 'sge'

  mapping.clusterOptions  = "-P project -pe omp 16 -l mem_total=94G"
  counting.clusterOptions = "-P project -pe omp 8  -l mem_total=16G"

  indir  = "/Users/pipeliner/pipelines/toy_data/rna-seq"
  outdir = "/Users/pipeliner/pipelines/rna-seq-results"
  fasta  = "${params.indir}/genome_reference.fa"
  gtf    = "${params.indir}/genome_annotation.gtf"

  // General pipeline parameters
  paired = true
  aligner = "hisat"
  quantifier = "featurecounts"

  index.use_existing = true
  index.path = "${params.indir}/alignment_indices/hisat_index/index/part"

  // Process-specific parameters
  feature_counts.cpus   = 8
  feature_counts.type   = "exon"
  feature_counts.id     = "gene_id"
  feature_counts.xargs  = ""
  feature_counts.ainj   = ""
}
```

FIGURE 2 | A shortened example of a configuration file, highlighting the key components. This configuration includes resource allocations for cluster executions, input and output paths to data, general pipeline parameters, as well as process-specific parameters.

```
featureCounts \\

# Common flags directly defined by the user
-T ${params.feature_counts.cpus} \\
-t ${params.feature_counts.type} \\
-g ${params.feature_counts.id} \\

# Flags handled by the pipeline
-a ${gtf} \\
-o "counts.raw.txt" \\

# Arguments indirectly defined by the user
${feature_counts_sargs} \\

# Extra arguments
${params.feature_counts.xargs} \\

# Input data
${bamfiles};

# After injection
${params.feature_counts.ainj}
```

FIGURE 3 | A code example of a template defined for the software tool featureCounts. The template wraps user-defined parameters, paths to data, as well as code injections into an executable bash script used in one of the pipeline steps.

each process—which are executed sequentially within the pipeline script. Because some software tools have hundreds of arguments, users have the option to insert code injections from the configuration file. These code injections can be used to pass uncommon keyword arguments or to append *ad hoc* processing steps (**Figure 3**). These features provide unrestricted control over each step in the execution of a pipeline. Furthermore, since all modifications are made within the configuration file—which is copied with each run—the pipeline script is left intact, preserving the reproducibility of each run regardless of any execution-specific changes the user may make.

Workflow Reproducibility

Pipeliner is designed to create reproducible workflows. An abstraction layer between Nextflow and Pipeliner logic enables platform independence and seamless compatibility with high-performance cloud computing executors such as Amazon Web Services. Pipeliner also uses Anaconda—a multi-platform package and environment manager—to manage all third-party software dependencies and handle pre-compilation of all required tools before a pipeline is executed (Continuum Analytics, 2016).

Pipeliner is bundled with a prepackaged environment hosted on Anaconda Cloud that contains all software packages necessary to run any of the three pipelines available. This virtual environment ensures consistent versioning of all software tools used during each pipeline execution. Additionally, all file paths, pipeline options, and process parameters are recorded, time stamped, and copied into a new configuration file with each run, ensuring pipelines are fully reproducible regardless of where and when they are executed.

Extensibility

Pipeliner makes the development of bioinformatics pipelines more efficient. The configuration file and processes that makeup each pipeline are inherited from shared blocks of code called template processes. For example, if a major update to an alignment tool requires modification to its template process, these changes propagate to all pipelines inheriting it (**Figure 4**). This property also minimizes the amount of code introduced as new pipelines are created, making them quicker to develop and easier to maintain. If a pipeline can inherit all of its processes with predefined templates, the user is only required to link these processes *via* Nextflow's scripting language and create a basic configuration file.

Rapid Development and Testing

Users can rapidly develop pipelines by using the toy datasets conveniently included with Pipeliner, enabling developers to test modifications made to their pipeline in minutes rather than hours. When testing, each execution covers only one configuration of parameters, meaning some processes may be skipped or partially executed depending on the configuration file. Therefore, to increase decision coverage, that is, the amount of tested reachable code, Pipeliner includes a custom testing module that automatically executes and logs a series of independent tests and configuration files (**Figure 5**). With these tools, users can efficiently build, test, and maintain multiple sequencing pipelines.

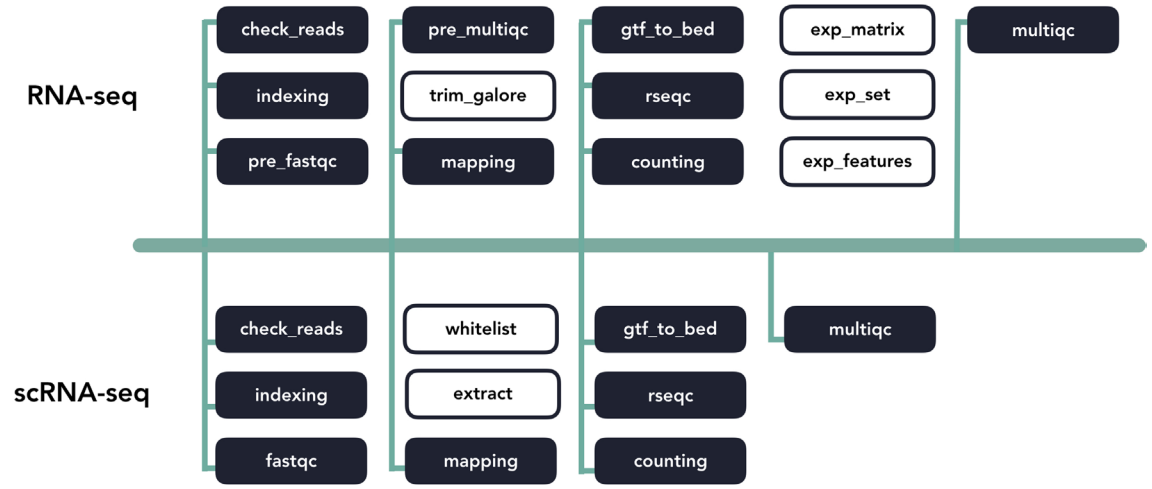


FIGURE 4 | A diagram of template code sharing between the RNA-seq and scRNA-seq pipelines. Each block represents an individual workflow step. Shaded blocks share template code, while unshaded blocks are unique.

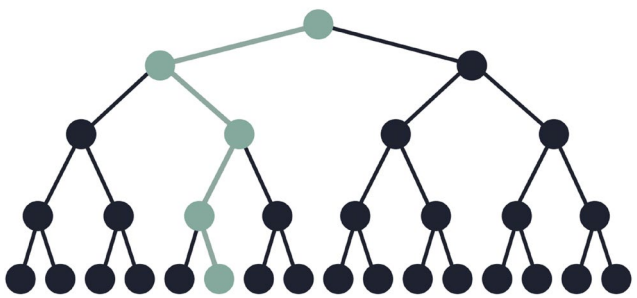


FIGURE 5 | A decision tree of the digital gene expression (DGE) pipeline based on available options in the configuration file. When testing, one execution of the pipeline will test one path in the decision tree (highlighted in green). The testing module automates the execution and logging of all possible paths.

Comparisons with Other Available Tools

Pipeliner has several characteristics that distinguish it from existing sequencing data workflows, such as RseqFlow (Wang et al., 2011), PRADA (Torres-García et al., 2014), or Galaxy (Goecks et al., 2010) (**Table 1**). These tools and their dependencies can be difficult to install and setup, lack sufficient documentation, and are rigid in their design, making customization challenging. Downloading and setting up Pipeliner is simple, and all dependencies are automatically installed through a virtual environment, ensuring data reproducibility and compatibility across various computing environments. Pipeliner is designed to be modular and flexible; therefore, workflow steps can be modified, skipped, removed, or extended. Pipeliner provides comprehensive documentation for general use as well as for developers who wish to extend the framework.

Usage Guide

In addition to comprehensive documentation of the framework and to demonstrate its ease of use, we provide a tutorial for processing the toy datasets available for each pipeline.

Processing Toy Datasets

The Pipeliner framework requires Nextflow and Anaconda. Nextflow requires Java 8 (or higher) to be installed and can be used on Linux and OS X machines. Third-party software tools will be installed and managed through an Anaconda virtual environment. Once the prerequisites are installed, the repository can be cloned from GitHub to any location through the following command:

```
$ git clone https://github.com/montilab/pipeliner
```

The next step is to clone and activate the virtual environment. The easiest method is to recreate the environment through the yml files provided in the repository. There is a single yml file for both Linux and OS X operating systems, containing all dependencies for all available pipelines.

```
$ conda env create -f pipeliner/envs/linux_env.yml
# Linux
$ conda env create -f pipeliner/envs/osx_env.yml #
OS X
$ source activate pipeliner
```

Pipeliner requires configuration of paths to input data such as fastq reads, bam alignments, references files, etc. When cloning Pipeliner to a new machine, all paths must be reconfigured. This process can be automated by running a script that will reconfigure any paths to the same directory of your clone.

```
$ python pipeliner/scripts/paths.py
```

The final step is to download a Nextflow executable package in the same directory as the available pipelines.

TABLE 1 | Comparison of Pipeliner with common sequencing data workflows.

	Pipeliner	RseqFlow	PRADA	Galaxy
Flexibility	Any computational tool or script can be incorporated into existing workflows. Workflow steps can be skipped, modified, or removed.	Computational tools are determined and difficult to remove or modify.	Computational tools are determined and difficult to remove or modify.	Users are limited to existing tools supported by the platform.
Extensibility	Extensible by design. Includes modular workflows, testing capabilities, and documentation specific to extending the framework.	Requires extensive Python experience to be extended by users. Limited documentation for making changes.	Requires extensive Python experience to be extended by users. Limited documentation for making changes.	The platform is not designed to be modified by its users.
Reproducibility	Dependencies are installable from Anaconda cloud or environment files for Linux or OS X. Workflow configuration files are recorded and reusable.	No virtualization methods used; correct dependency versions must be installed manually. Workflow steps are not logged.	No virtualization methods used; correct dependency versions must be installed manually. Workflow steps are not logged.	Workflows can be saved, shared, and reproduced on the platform.
Installation	A few simple steps to configure environment and install dependencies.	Must install dependencies and configure environment manually.	Must install dependencies and configure environment manually.	N/A
Ease of Use	Simple config file. Provides small example datasets for local machines. One command to run entire workflow. Extensive documentation.	Simple config file. Workflow steps must be run individually. Provides example datasets. Limited documentation.	Simple config file. Workflow steps must be run individually. Provides example datasets. Limited documentation	Click and drag interface. Extensive documentation.
Data Types	RNA-seq scRNA-seq digital gene expression	RNA-seq	RNA-seq	RNA-seq ChIP-seq Mass Spec 16S
Interface Type	Command Line	Command Line	Command Line	Web-based
Available for Download	GitHub repository	SourceForge Tarball	Google Code Tarball	GitHub repository
Publicly Available	Yes	Yes	Yes	Yes

```
$ cd pipeliner/pipelines
$ curl -s https://get.nextflow.io | bash
```

With the setup complete, any of the available pipelines can be executed with their respective toy datasets with the following commands.

```
$./nextflow rnaseq.nf -c rnaseq.config
$./nextflow scrnaseq.nf -c scrnaseq.config
$./nextflow dge.nf -c dge.config
```

Proof of Concept

To showcase the applicability of Pipeliner to real-world datasets, we reprocessed 48 RNA-seq-paired read files for the lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) cohort from The Cancer Genome Atlas (TCGA). For each cohort, the TCGA uses a standardized pipeline where reads are mapped to a reference genome with STAR and quantified by HTSeq. While the TCGA provides open access to the count matrix, some researchers have opted to use alignment and quantification algorithms specific to their research interests (Rahman et al., 2015). For this reason, the TCGA also provides raw sequencing data; however, its large size requires parallelization on a high-performance computing platform. We argue that Pipeliner is a suitable choice for users looking for alternative reprocessing of TCGA datasets with minimal pipeline development.

Pipeliner makes alternative processing of TCGA and other publicly available data straightforward. In processing raw RNA-seq data for DLBC, paired fastq reads were downloaded from the Genomic Data Commons Data Portal. For each sample, Pipeliner requires an absolute file path to reads. After specifying this information, Pipeliner was able to successfully process all data with HISAT2, featureCounts, and the remaining settings left to default. Data processing methods can have subtle effects on downstream analysis of sequencing data. This is exemplified by an increase in assigned features and decrease in multi-mapping when using HISAT2/featureCounts instead of STAR/HTSeq (**Figure 6**). The ability for researchers to reprocess publicly available datasets to suit their specific interests is important, and Pipeliner is a useful software tool that meets those needs. The flexibility provided by Pipeliner is ideal for users experimenting with different tools and parameters. For example, because Pipeline is capable of taking aligned bam files as input and skipping preceding steps, we were able to rapidly try all three quantification options without rerunning unrelated processes. This level of control is critical for downstream analysis of the processed data. To help researchers extend this example to other datasets, we provide the scripts used to obtain and organize TCGA data from the Genomic Data Commons as well as the configuration file used by Pipeliner to process the data in the supplementary information.

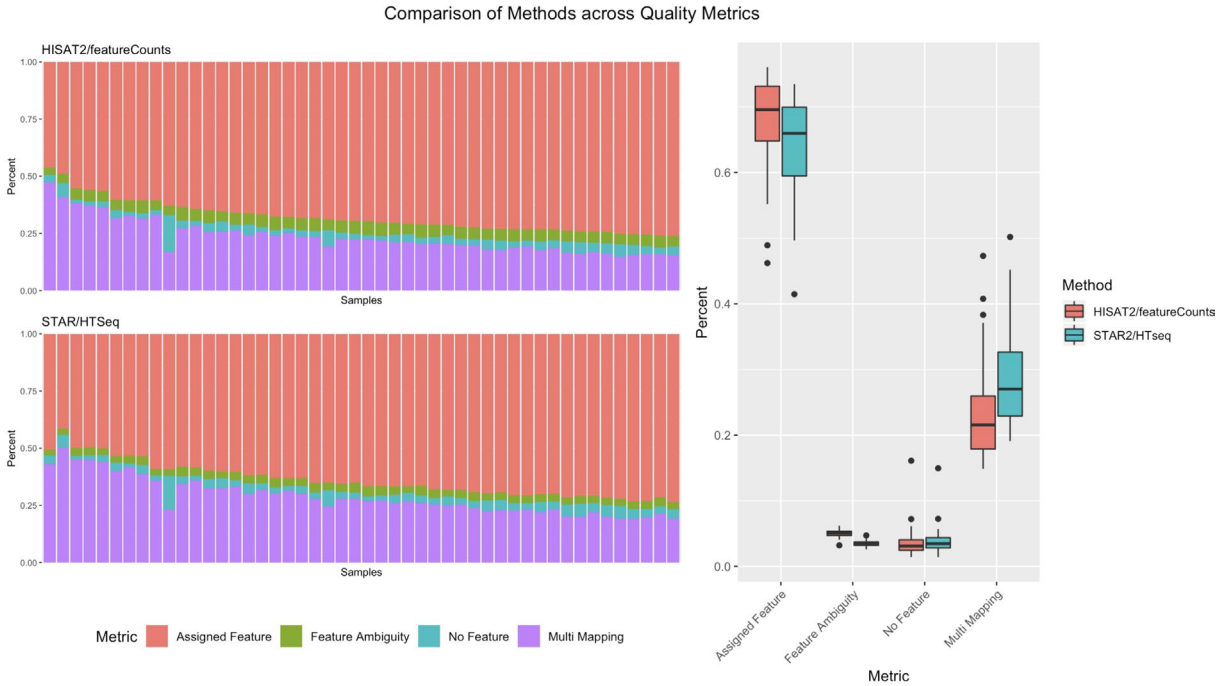


FIGURE 6 | A summary of counting and mapping quality metrics when processing with Pipeliner using HISAT2 and featureCounts compared with processed counts from The Cancer Genome Atlas (TCGA)-diffuse large B-cell lymphoma (DLBC) using STAR and HTSeq.

CONCLUSIONS

Together with Nextflow and Anaconda, Pipeliner enables users to process large and complex sequencing datasets with pipelines that are customizable, reproducible, and extensible. The framework provides a set of user-friendly tools for rapidly developing and testing new pipelines for various types of sequencing data that will inherit valuable design features of existing pipelines. We apply the RNA-seq pipeline to real-word data by processing raw sequencing reads from the DLBC cohort provided by the TCGA and provide supplementary files that can be used to repeat the analysis or serve as a template for applying Pipeliner to other publicly available datasets.

AVAILABILITY AND FUTURE DIRECTIONS

Pipeliner is implemented in Nextflow, Python, R, and Bash and released under a General Public License 3.0 license. It is publicly available at https://github.com/montilab/pipeliner and supports Linux and OS X operating systems. Comprehensive documentation is generated with Sphinx and hosted by Read the Docs at https://pipeliner.readthedocs.io/. We will continue to develop the Pipeliner framework as the Nextflow programming language matures, and we plan to provide additional pipelines for other types of sequencing data and analysis workflows in the future.

AUTHOR CONTRIBUTIONS

AF—Developed the current version of Pipeliner, wrote the manuscript, and generated the figures; TK—Initiated the project and developed early versions of Pipeliner; KK—Initiated the project and developed early versions of Pipeliner; DK—Initiated the project and developed early versions of Pipeliner; YK—Assisted in development of individual sequencing pipelines; JC—Initiated, oversaw, and guided the project as well as helped in writing the manuscript; SM—Initiated, oversaw, and guided the project as well as helped in writing the manuscript.

FUNDING

This work was supported by a Superfund Research Program grant P42ES007381 (SM) and the LUNGevity Career Development Award (JC).

ACKNOWLEDGMENTS

The authors would like to thank P. Di Tommaso for his assistance with Nextflow-related inquiries and A. Gower for his advice for improving and testing Pipeliner.

REFERENCES

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31 (2), 166–169. doi: 10.1093/bioinformatics/btu638

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinforma.* doi: 10.1016/S1048-9843(02)00144-3

Continuum Analytics. (2016). Anaconda Software Distribution: Version 2-2.4.0. Available online at: https://continuum.io.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35 (4), 316–319. doi: 10.1038/nbt.3820

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1), 15–21. doi: 10.1093/bioinformatics/bts635

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32 (19), 3047–3048. doi: 10.1093/bioinformatics/btw354

Goecks, J., Nekrutenko, A., Taylor, J., Afgan, E., Ananda, G., Baker, D., et al. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11. doi: 10.1186/gb-2010-11-8-r86

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357–360. doi: 10.1038/nmeth.3317

Krueger, F. (2016). Trim Galore. *Babraham Bioinforma*.

Liao, Y., Smyth, G. K., and Shi, W. (2014). Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30 (7), 923–930. doi: 10.1093/bioinformatics/btt656

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33 (3), 290–295. doi: 10.1038/nbt.3122

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., and Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in the Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* 31 (22), 3666–3672. doi: 10.1093/bioinformatics/btv377

Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., and Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-seq. *BioRxiv.* doi: 10.1101/003236

Torres-García, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., et al. (2014). PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* 30, 2224–2226. doi: 10.1093/bioinformatics/btu169

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28 (16), 2184–2185. doi: 10.1093/bioinformatics/bts356

Wang, Y., Mehta, G., Mayani, R., Lu, J., Souaiaia, T., Chen, Y., et al. (2011). RseqFlow: workflows for RNA-Seq data analysis. *Bioinformatics* 27, 2598–2600. doi: 10.1093/bioinformatics/btr441

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

TECHNOLOGY AND CODE
published: 04 July 2019
doi: 10.3389/fgene.2019.00633

FindTargetsWEB: A User-Friendly Tool for Identification of Potential Therapeutic Targets in Metabolic Networks of Bacteria

Thiago Castanheira Merigueti[1], Marcia Weber Carneiro[4], Ana Paula D'A. Carvalho-Assef[3], Floriano Paes Silva-Jr[2] and Fabricio Alves Barbosa da Silva[1]**

[1] Scientific Computing Program–Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [2] Laboratory of Experimental and Computational Biochemistry of Drugs (LaBECFar), Oswaldo Cruz Institute–Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [3] Research Laboratory in Hospital Infection (LAPIH), Oswaldo Cruz Institute–Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [4] Graduate Program in Biotechnology for Health and Investigative Medicine–Oswaldo Cruz Foundation (FIOCRUZ), Bahia, Brazil

OPEN ACCESS

Edited by:
Helder Nakaya, University of São Paulo, Brazil

Reviewed by:
Priyanka Baloni, Institute for Systems Biology (ISB), United States
Leandro Marcio Moreira, Universidade Federal de Ouro Preto, Brazil

***Correspondence:**
Floriano Paes Silva-Jr
floriano@ioc.fiocruz.br
Fabricio Alves Barbosa da Silva
fabricio.silva@fiocruz.br

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

Received: *27 March 2019*
Accepted: *17 June 2019*
Published: *04 July 2019*

Citation:
Merigueti TC, Carneiro MW, Carvalho-Assef APD'A, Silva-Jr FP and Silva FAB (2019) FindTargetsWEB: A User-Friendly Tool for Identification of Potential Therapeutic Targets in Metabolic Networks of Bacteria. Front. Genet. 10:633. doi: 10.3389/fgene.2019.00633

Background: Healthcare-associated infections (HAIs) are a serious public health problem. They can be associated with morbidity and mortality and are responsible for the increase in patient hospitalization. Antimicrobial resistance among pathogens causing HAI has increased at alarming levels. In this paper, a robust method for analyzing genome-scale metabolic networks of bacteria is proposed in order to identify potential therapeutic targets, along with its corresponding web implementation, dubbed FindTargetsWEB. The proposed method assumes that every metabolic network presents fragile genes whose blockade will impair one or more metabolic functions, such as biomass accumulation. FindTargetsWEB automates the process of identification of such fragile genes using flux balance analysis (FBA), flux variability analysis (FVA), extended Systems Biology Markup Language (SBML) file parsing, and queries to three public repositories, i.e., KEGG, UniProt, and DrugBank. The web application was developed in Python using COBRApy and Django.

Results: The proposed method was demonstrated to be robust enough to process even non-curated, incomplete, or imprecise metabolic networks, in addition to integrated host-pathogen models. A list of potential therapeutic targets and their putative inhibitors was generated as a result of the analysis of *Pseudomonas aeruginosa* metabolic networks available in the literature and a curated version of the metabolic network of a multidrug-resistant *P. aeruginosa* strain belonging to a clone endemic in Brazil (*P. aeruginosa* ST277). Genome-scale metabolic networks of other gram-positive and gram-negative bacteria, such as *Staphylococcus aureus*, *Klebsiella pneumoniae*, and *Haemophilus influenzae*, were also analyzed using FindTargetsWEB. Multiple potential targets have been found using the proposed method in all metabolic networks, including some overlapping between two or more pathogens. Among the potential targets, several have been previously reported in the literature as targets for antimicrobial development, and many targets have approved drugs. Despite similarities in the metabolic network structure

for closely related bacteria, we show that the method is able to selectively identify targets in pathogenic *versus* non-pathogenic organisms.

Conclusions: This new computational system can give insights into the identification of new candidate therapeutic targets for pathogenic bacteria and discovery of new antimicrobial drugs through genome-scale metabolic network analysis and heterogeneous data integration, even for non-curated or incomplete networks.

Keywords: systems biology, flux balance analysis, metabolic network, COBRA analysis, Python (programming language)

BACKGROUND

Healthcare-associated infections (HAIs), previously called hospital infections, are a serious public health problem and can develop either as a direct result of medical or surgical treatment or from being in contact with a healthcare setting. These infections include central line-associated bloodstream infections, catheter-associated urinary tract infections, ventilator-associated pneumonia (VAP), and surgical site infections. Among the pathogens related to HAI, the group of bacteria is the one that stands out. More than 2 million HAIs occur each year in the USA (Stone et al., 2005), with 50–60% being caused by antimicrobial resistant bacteria. In 2014, the World Health Organization (WHO) published the report "Antimicrobial resistance: global report on surveillance" (WHO, 2014) warning of the growing increase in antimicrobial resistance in the world. Antimicrobial resistance among hospital pathogens has increased at alarming levels, both in developed and developing countries. It is estimated that there will be a worldwide spread of untreatable infections both inside and outside hospitals. According to a bulletin published in 2017 by WHO (WHO, 2017), there are 12 major antibiotic-resistant bacteria that deserve attention and urgently need more research and development (R&D) of new and effective antibiotic treatments. Gram-negative bacteria are the most involved in HAI (carbapenem-resistant *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacteriaceae* family), and R&D on new antibiotics against these is considered to be of critical priority (WHO, 2017). In humans, *P. aeruginosa* is an opportunistic pathogen that causes severe infections in immunocompromised individuals. This pathogen is the main cause of morbi-mortality in patients with cystic fibrosis (Kerr and Snelling, 2009) and is a major cause of VAP.

Given the potential severity of multidrug-resistant bacteria and the lack of treatment options, the identification and implementation of effective strategies to prevent such infections are urgent priorities.

The integration of mathematical, statistical, and computational methods for biological data analysis to enable the discovery of new therapeutic targets for any bacteria is extremely relevant. The combination of bioinformatics, system modeling, and heterogeneous data integration can be a powerful tool for this purpose.

Several strategies have been proposed to search for drug targets from genome-scale models of bacterial metabolism. More often, essential genes are identified from single virtual knockouts where flux balance analysis (FBA) (Orth et al., 2010) is used to assess if this gene deletion is able to halt a selected function of bacterial metabolism. Usually, such function is biomass production (Rienksma et al., 2014). Other criteria can be combined to prioritize genes among candidate drug targets, such as existence of druggable pockets (Kozakov et al., 2015) or specificity to the bacteria as compared to the host proteins.

The construction of genome-scale metabolic network is a laborious endeavor. It combines automated steps with manual curation. The most used protocol, proposed by Thiele and Palsson (2010), lists a total of 94 steps. Nevertheless, the process is error-prone, and normally the resulting network may correctly predict some phenomena while disregarding others, which are less relevant to the study related to the reconstructed metabolic network.

The BioCyc database (Caspi et al., 2015) classifies pathway/genome databases (PGDB), each containing the full genome and predicted metabolic network of one organism, into three tiers. *Tier 1* corresponds to PGDBs that have received at least 1 year of manual curation and are updated continuously. *Tier 2* includes PGDBs that have received moderate (less than a year) amounts of review and are usually not updated on an ongoing basis. Finally, *Tier 3* refers to PGDBs that were created computationally and received no subsequent manual review or updating.

In this work, the same classification for genome-scale metabolic network models is adopted. The focus here is on metabolic network models that can be classified as Tier 2 and Tier 3, according to the BioCyc database classification. In this manuscript, *draft* metabolic reconstructions are considered Tier 3 models. Published curated metabolic models are classified as Tier 2, unless the model is identified in the literature as Tier 1.

Herein, a method for analyzing genome-scale metabolic networks of bacteria is proposed in order to identify potential therapeutic targets, along with its corresponding web implementation, dubbed *FindTargetsWEB*. The proposed method is computationally efficient, user-friendly, and robust to errors in reconstructed genome-scale metabolic networks, which are more frequent in Tier 3 (*draft*) metabolic networks. The web interface of the application is straightforward, and results are sent directly to an email address informed by the user. To demonstrate the flexibility of FindTargetsWEB, 10 genomic-scale metabolic networks of bacterial strains are analyzed in this paper. Nine of the 10 networks are available in the literature, all classified as Tier 2 models in this work: *P. aeruginosa* PAO1—version 2008 (Oberhardt et al., 2008), *P. aeruginosa* PAO1—version 2017

(Bartell et al., 2017), *P. aeruginosa* PA14 (Bartell et al., 2017), *Klebsiella pneumoniae* (Liao et al., 2011), *Haemophilus influenzae* (Schilling and Palsson, 2000), a host-pathogen genome-scale reconstruction based on the *Mycobacterium tuberculosis* metabolic network (Bordbar et al., 2010), *Staphylococcus aureus* (Becker and Palsson, 2005), and *Pseudomonas putida* (Puchałka et al., 2008). Results are also presented for two metabolic networks of *P. aeruginosa* CCBH4851, which is a multi-drug resistant strain belonging to a clone endemic in Brazil (*P. aeruginosa* ST277) (Silveira et al., 2014). Both reconstructions of *P. aeruginosa* CCBH4851 were made by our group. One reconstruction can be classified as Tier 3, and the other is the corresponding curated version, classified as Tier 2.

The web application proposed in this work combines FBA, flux variability analysis (FVA) (Gudmundsson and Thiele, 2010), extended Systems Biology Markup Language (SBML) parsing, and heterogeneous data integration in order to identify the most promising therapeutic targets. All SBML files processed in this work are available as Supplementary Material. The underlying hypothesis related to FVA is that reactions which the maximum flux is equal to the minimum flux (i.e., flux range equal to zero), given the optimal biomass production, are less robust to potential perturbations. Indeed, a high rigidity for a given reaction flux (i.e., flux range equal to zero) may indicate that the flux through this reaction is crucial for sustaining optimal growth, while a lower rigidity (i.e., flux range greater than zero) indicates that there might be alternate pathways to carry the reaction flux (Oberhardt et al., 2010). Flux ranges fell into three categories: i) inflexible fluxes (flux range equal to zero), ii) fluxes with bounded flexibility (flux range greater than zero, but bounded), and iii) infinitely flexible fluxes (flux range greater than zero, unbounded). The FVA analysis carried out by FindTargetsWEB aims to identify potential targets associated with inflexible fluxes, i.e., flux range equal to zero. The genome-scale metabolic network analysis is combined with several queries to multiple public repositories, such as KEGG (Ogata et al., 1999), UniProt (UniProt, 2018), and DrugBank (Wishart et al., 2008), to assess the druggability and toxicology of potential targets. FindTargetsWEB has identified potential targets for all networks. Several of the potential targets have been described in the literature. Other targets are candidates for future experimental investigation.

IMPLEMENTATION

Some of the main requirements related to the implementation of the general method described in this work, dubbed FindTargetsWEB, were ease of use, availability, robustness, and performance. After careful consideration, Python was selected as the implementation language. Python is a high-level, interpreted, scripted, imperative, object-oriented, dynamic, and strongly typed programming language created by Van Rossum and Drake (2003). Its many advantages favor the fulfillment of the main requirements of the application. Another advantage is the availability of the COBRApy package. *COnstraint-Based Reconstruction and Analysis Toolbox* (COBRA) (Hyduke et al. 2011) methods are widely used for genome-scale modeling of metabolic networks in prokaryotes and eukaryotes. The COBRA Toolbox for MATLAB is a leading software package for analyzing metabolism on a genomic scale. On the other hand, COBRApy (Ebrahim et al., 2013) is a Python module that provides support for basic COBRA methods. COBRApy is designed in an object-oriented way, which facilitates the representation of the complex biological processes of metabolism. COBRApy does not require MATLAB to work; however, it includes an interface to the COBRA Toolbox for MATLAB to facilitate the use of legacy codes. To improve performance, COBRApy includes parallel processing support for computationally intensive processes. FindTargetsWEB is implemented as a web application. Therefore, the user only needs a web browser to access the system. The system interface is intuitive: the user needs to provide the SBML file describing the metabolic network reconstruction, the organism species associated with the metabolic network reconstruction, which defines a filter to KEGG queries, and information such as name and e-mail address (**Figure 1**). It should be emphasized that the FindTargetsWEB list of analyzable species is easily expandable and can include both gram-negative bacteria, gram-positive bacteria, and bacteria that cannot be classified as either gram-positive or gram-negative. In the following screen, the user decides if he/she wants to analyze the network using the FBA method alone or a combination of the FBA+FVA methods (**Figure 2**). The FBA+FVA method pinpoints reactions and associated genes in which knockout completely stops (zeroes) biomass generation and has an FVA range of zero. Therefore, the FBA+FVA method is more restrictive than the FBA-only option. It should be highlighted that the targets found by the FBA+FVA method compose a proper subset of the set of targets found by the FBA-only method. Robustness is provided by the design of the method itself, as described in the following paragraphs.

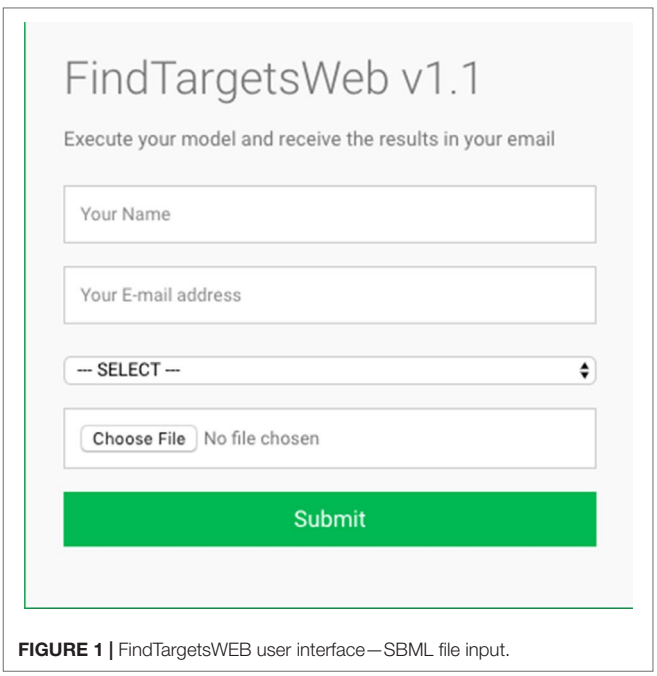


FIGURE 1 | FindTargetsWEB user interface—SBML file input.

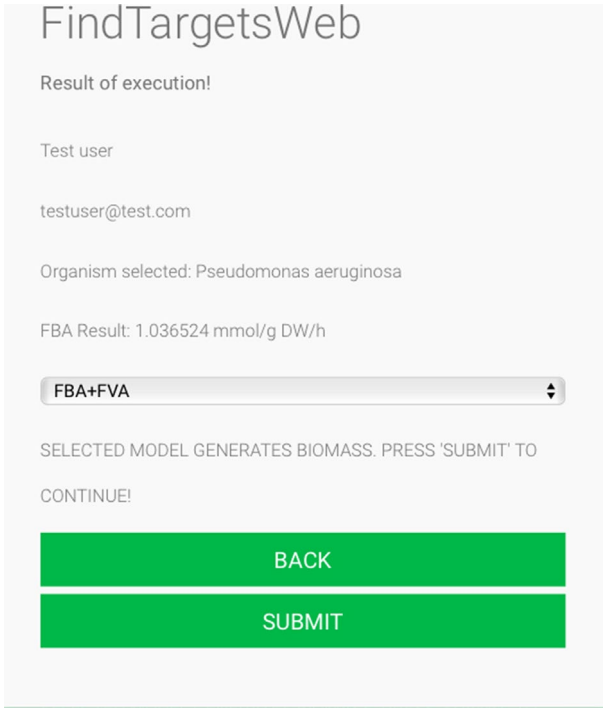


FIGURE 2 | FindTargetsWEB user interface—choice of analysis method.

Target identification is carried out through a computational workflow that runs the metabolic network analysis and pinpoints genes whose virtual knockout interrupts the generation of biomass. Therefore, the minimum level of curation required for a metabolic network model to be processed by FindTargetsWEB is to have a biomass reaction flux greater than zero. The list of potential targets is filtered using FVA (if the user decides to do so), and the workflow retrieves possible inhibitors for the identified genes, verify if such inhibitors are available as approved drugs, and evaluate their toxicity to humans by querying several repositories.

The workflow was implemented using the Python programming language, version 3.6.3, and the COBRApy framework version 0.9.0. This framework has the necessary methods for reading the SBML (Hucka et al., 2015) file that describes the genome-scale metabolic network of the bacterium under analysis. The solver used for FBA and FVA analysis is GLPK (https://www.gnu.org/software/glpk/), which is the COBRApy default solver that is easily deployable on Linux platforms. The system is deployed in an Ubuntu v18.04 server with 64GB RAM. Prior to processing, when needed, SBML files were converted to the SBML level 3 format using the command `cobra.io.sbml3.write_sbml_model` from COBRA. The SBML files processed in this manuscript were retrieved from the BioModels repository (Glont et al., 2017) or directly from the supplementary material of the associated reference. The main steps of the method are described below. The whole method is depicted in **Figure 3**.

1. **Validation of the SBML file describing the genome-scale metabolic network**—In this step, the system first creates a table containing gene/reaction/metabolite data obtained from the SBML file and then checks if the metabolic network reconstruction generates biomass. This is done through the FBA method, considering the biomass reaction as the target for maximization. If the biomass value is zero, the system outputs an error to the user and halts processing. If the maximum flux of the biomass reaction is greater than zero, the workflow proceeds to the next step.
2. **Use of FVA to filter reactions—**After validating the metabolic network, reactions are filtered using the FVA method, if the user has decided to analyze the metabolic network using a combination of the FBA+FVA methods. The objective is to consider, in the following processing steps, those reactions which the range of possible flux values is equal to zero, given the optimal biomass generation value determined in the previous step. The underlying assumption is that reactions with a range equal to zero are less robust, i.e., more susceptible to perturbations, as stated in the introduction. Note that the FVA method can be implemented in a computationally efficient way (Gudmundsson and Thiele, 2010), and the cost of FVA analysis on the overall execution time of FindTargetsWEB is negligible.
3. **Simulation of reaction knockout**—In this step, single reaction knockouts are performed. The process is done by zeroing the maximum and minimum reaction flux constraints and running FBA again, for each reaction in the network. If biomass generation is zeroed when knocking out a given reaction, its information is stored in a list for further processing. If gene IDs are available in the SBML file, the workflow proceeds to step 4. Otherwise, it jumps directly to step 6b.
4. **Simulation of gene knockout**—In this step, the system performs single knockouts for each gene described in the model, where the COBRApy framework queries the reactions that are linked to the selected gene and zeroes the minimum and maximum value of each reaction bound to the gene, taking into account gene-protein-reaction (GPR) relations. In the same way as the previous step, if the value of the generation of biomass has zeroed, the corresponding gene information is stored in a second list. It is worth noting that one gene can be associated with more than one reaction, and one reaction may require the expression of several genes.
5. **Consolidation/unification of knockouts results—**In this step, both lists generated in the previous steps are unified, i.e., the list of reactions generated in step 3 and the gene list generated in step 4. In order to a gene to be included in the final list, it should be included in the list of step 4 and be associated with at least one reaction stored in step 3 (see Algorithm 1). These are the candidate genes that the workflow is going to consider in the following steps. It should be highlighted that the final list is filtered according to the FVA processing performed in step 2, if the option FBA+FVA is selected by the user.

Algorithm 1: Consolidation of knockout results (SBML with mapped genes)

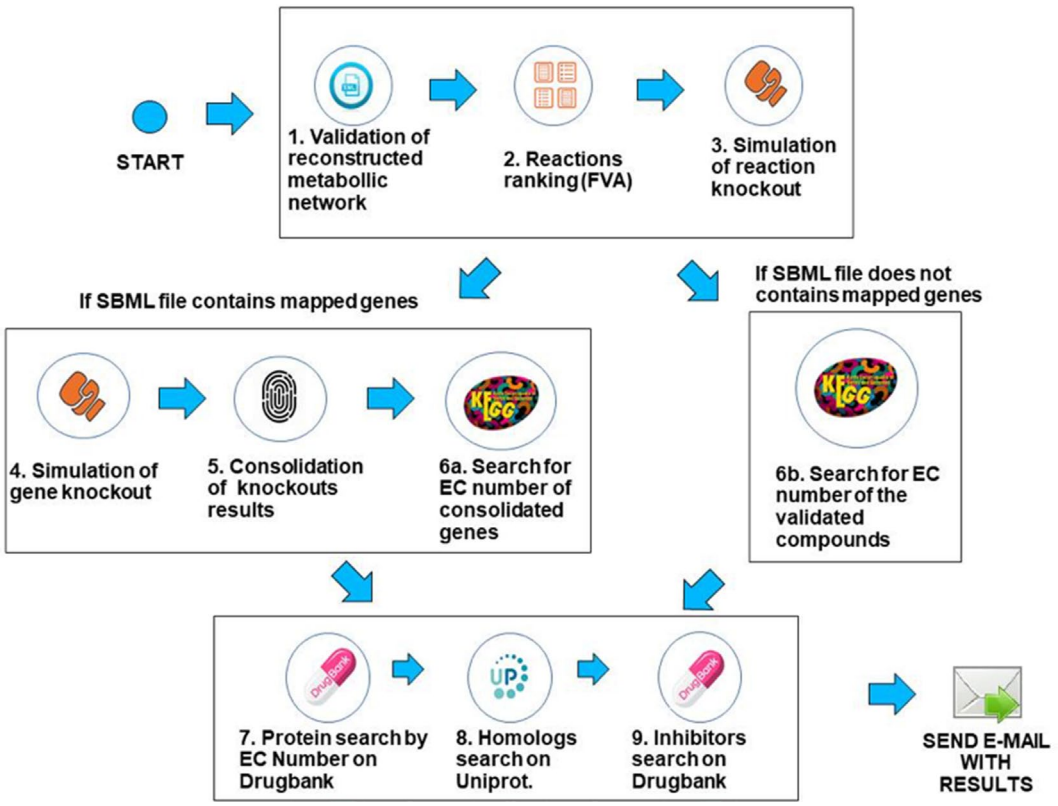


FIGURE 3 | Description of the computational workflow.

Input: List of genes from knocked-out reactions/list of knocked-out genes

Output: Unified list of target genes in a text file

```
1: procedure UnificationTargetsList(targetGeneListFromReact, targetGeneList)
2:     read targetGeneListFromReact
3:     read targetGeneList
4:     open file "targetgenes.txt"
5:     for all targetgene in targetGeneListFromReact do
6:         if targetgene in targetGeneList then
7:         write targetgene in file "targetgenes.txt"
8:         end if
9:     end for
10:    close file "targetgenes.txt"
11: end procedure
```

6a. Search for EC numbers of consolidated genes. In this step, the system queries the KEGG repository to obtain the EC number of each gene included in the final gene list obtained in the previous step (file "targetgenes.txt"). KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information (Ogata et al., 1999). This step is important because drug retrieval in DrugBank requires the associated EC number. The result of this step is a list of EC numbers associated to their respective genes. The workflow then proceeds to step 7.

6b. Search for EC number using reaction information. If gene IDs are not available in the SBML file, which may be the case in draft (Tier 3) metabolic network models, EC numbers are retrieved from KEGG based on reaction information. This step is particularly important for incomplete metabolic reconstructions that do not include GPR relations and is directly related to the application's requirement of robustness to incompleteness on metabolic network data. The KEGG search is performed using all the compounds involved in the corresponding reaction. See Algorithm 2 for a detailed description of the processing related to this step. It is worth emphasizing that this step is executed only for incomplete descriptions of genome-scale metabolic networks. The complexity of Algorithm 2 is $O(C)$, where C is the number of compounds included in the SBML file.

Algorithm 2: Search for EC numbers using reaction information (SBML without mapped genes)

Input: List of chemical compounds of reaction

Output: List of EC numbers found

```
1: procedure alternativeStepToGetECNumberWithoutGenes (listCompoundFromSBML)
2: # file with all compounds in SBML.
3: read listCompoundFromSBML
4:
5: # Instance of biomodels python module
6: k <- KEGG instance
```

```
# setting timeout in seconds
k.timeout <- 200000

# All compounds in SBML file. Ex.: 2 A –> B
for all compound in listCompoundFromSBML do

    # find all stoichiometric values with regex method in compound. Ex.: A –> B
    compound_no_stoich <- remove all stoichiometric values in compound
    param_splt <- empty

    # Verify if reaction is reversible or irreversible
    if "< = >" in compound_no_stoich then
            param_splt <- "< = >"
    else
            param_splt <- "–>"
    end if

    # Separate compounds in reactant and product. Ex.: compound_splt = ['A', 'B']
    compound_splt <- compound_no_stoich.split(param_splt)

    # If compound belongs to a transport reaction (influx or eflux), jump to next iteration
    if compound_splt.length < 2 then
            continue
    end if

    list_ec_number_0 <- initialize empty list
    list_ec_number_1 <- initialize empty list

    # Start iterating compound_splt list with reactant and product
    for (x = 0,1) do

            # Get reactant or product in this variable
            item_compound <- compound_splt[x] without spaces
    list_id_cpd_KEGG <- initialize empty list

    # If reactant or product contains "+", find ID in KEGG for components.
    # Else, find ID in KEGG for only one component.
    if " + " in item_compound then
            item_compound_splt <- item_compound.split(" + ")
            for all cpd_item in item_compound_splt do
                    # find id compound in KEGG for cpd_item
                    # and insert in ids list
                    result_id_cpd<-k.find("compound", cpd_item)
                    insert result_id_cpd in list_id_cpd_KEGG
            end for
    else
            result_id_cpd <- k.find("compound", item_compound)
            insert item_compound in list_id_cpd_KEGG
    end if

    # Here, all list_id_cpd_KEGG are concatenated
    # found to search the reaction in KEGG.
    # In Python, if list_id_cpd_KEGG length is less than 2,
    # don't put the "+" in end of string.
    str_item_compound_in_cpd <- list_id_cpd_KEGG concat with "+"

    # find all reactions in KEGG with IDs of compounds
    result_link_reactions_cpd <- k.link("reaction", str_item_compound_in_cpd)

    # All results of result_link_reactions_cpd are inserted here
    set_id_reaction_KEGG <- insert all reactions found in KEGG.

    # find all EC numbers in KEGG with reactions IDs
    # in set_id_reaction_KEGG and insert in result_list_ec
    result_list_ec = k.link("enzyme", set_id_reaction_KEGG)
    if x = 0 then
            insert result_list_ec in list_ec_number_0
    else
            insert result_list_ec in list_ec_number_1
    end if

end for

list_ec_number_intersect <- initialize empty list
txt_file <- initialize txt file

# Starts to iterate the list of ECs to identify intersections
# If found, related EC numbers are written in a text file
for all ec_number_0 in list_ec_number_0 do
    if ec_number_0 in list_ec_number_1 then
            record ec_number_0 in a txt_file
    end if
end for

end for

end procedure
```

7. **Search for EC numbers on DrugBank**—With the EC numbers obtained in the previous steps, the system queries the DrugBank repository to verify if this database has any record of the listed EC numbers. The DrugBank database is a repository that combines detailed drug data with comprehensive drug target information (Wishart et al., 2008). If an exact match is found, the system retrieves the values of the name of the protein, organism, and UniProt ID.

   When executing this query, the protein retrieved can be mapped in another organism, distinct from the target

bacterium. Thus, the next step (step 8) is necessary to confirm whether the protein retrieved has a homologue in the target bacterium. Clearly, exact matches are also possible. In any case, the retrieved data is validated in the next step.

8. **Search for homologues on UniProt**—Finally, the system searches for sequence similarity between the proteins described by UniProt IDs retrieved in the last step and the proteins encoded by the genome of the target bacterium using the BLAST (basic local alignment search tool) (Altschul et al., 1990) application deployed in the UniProt server. If there is a hit (i.e., sequence similarity above 30%), all corresponding data concerning the homologue found is stored.

   In this step, the homology between the target protein and human proteins is also considered. If the sequence similarity with a human protein is greater than the similarity with the target bacterium, the protein under analysis is discarded, since the inhibition of that protein could be harmful to the host. Otherwise, several data are stored, such as metabolic pathway, function, and catalytic activity, among others. This step of the workflow is the most time-consuming, since BLAST is executed for all proteins identified in the previous step.
9. **Search for existing inhibitors**—The last step is to query the DrugBank repository, using the stored UniProt IDs, in order to retrieve known inhibitors, if available. After this last step, the system generates spreadsheets containing all results that are sent to the user in a compressed file.

This method presents as results candidate genes that, when knocked-out, will cease the biomass production of the microorganism. Candidate genes must be associated with potential drug targets in DrugBank, and their sequence similarity to human proteins is also checked. The application then identifies available ligands, most often inhibitors, to the selected genes.

System Output

Results of FindTargetsWEB's analysis are sent to the user as a compressed file, to the e-mail address informed at the start of execution. Five spreadsheets are included in the compressed file:

- *08-filter_ECNumbers_DrugBank*—This spreadsheet contains the EC number of putative targets, along with product, organism name, UniProt ID, and DrugBank ID
- *11- hits_Uniprot*—This spreadsheet contains additional information related to UniProt queries, such as percentage of sequence similarity, BLAST e-value, gene name, pathway, function, and catalytic activity.
- *13-list_inhibitors_per_target*—This spreadsheet lists all inhibitors found for all targets. Included information are drug name, drug group (e.g. experimental, approved, investigational), and drug action.
- *14-list_inhibitors_approved*—This spreadsheet lists all inhibitors with approved drugs found for all targets. Included information are drug name, drug group (approved), and drug action.
- *model_data*—This spreadsheet lists data related to the input SBML file, such as gene IDs and associated reactions. The complete information of which reactions are associated with each gene in the metabolic network model is included in this file.
- *summary_results*—This spreadsheet contains a summary of data included in the previous files. Included fields are EC numbers, product, organism name, gene name, pathway, function, catalytic activity, drug name, drug group, and drug action.

RESULTS

In this section, analysis results for several strains of *P. aeruginosa*, *K. pneumoniae*, *H. influenzae, S. aureus, P. putida*, and a host-pathogen genome-scale reconstruction based on the *M. tuberculosis* metabolic network are presented. It should be highlighted that FindTargetsWEB can carry out analysis for other bacterial species, as indicated by the list box on the initial web page of the application. Indeed, even this list can be easily expanded to include additional species of interest, through a user request to FindTargetsWEB support team.

Analysis of Metabolic Network Models of *P. aeruginosa*

To evaluate the accuracy of results for several metabolic networks, initially, the analysis of four metabolic networks of *P. aeruginosa* is discussed. A survey of the literature is also presented to confirm the feasibility of the candidate genes as antibacterial drug targets. Gene function and related pathways are also considered in the evaluation of results.

The four metabolic networks of *P. aeruginosa* strains analyzed by FindTargetWEB were: PAO1 version 2008—iMO1056 (BioModels ID 1507180020) (Oberhardt et al., 2008), PAO1 version 2017—iPAE1146 (Bartell et al., 2017), PA14—iPAU1129 (Bartell et al., 2017), and a curated version (Tier 2) of the metabolic network of *P. aeruginosa* CCBH4851 (Silveira et al., 2014). The SBML level 3 file describing the Tier 2 *P. aeruginosa* CCBH4851 network is available as supplementary material, as well as the SBML files of the other networks considered in this paper. It is worth noting that each metabolic network model presents a different value for the growth rate after validation of biomass generation by FBA; for PA01 version 2008, the growth rate corresponds to 1.047929 h^{-1}; PA01 version 2017 has a growth rate of 15.509635 h^{-1}; for the PA14 model, the growth rate is 15.508373 h^{-1}, and the Tier 2 CCBH4851 model has a growth rate of 1.036524 h^{-1}. Differences in growth rate among metabolic network models are due to the distinct biomass equations, as well as variation in the number of genes, reactions, and metabolites in each of the metabolic network models.

It should be mentioned that the growth rates associated with the PA14 and PAO1-2017 (Bartell et al., 2017) models depart by far from the observed growth rates of *P. aeruginosa* spp., which may vary between 0.3 and 0.8 h^{-1}, depending on cultivation conditions (Brown, 1957) (Seto and Noda, 1982) (Yang et al., 2008). Nevertheless, FindTargetsWEB can still process those networks. The only requirement is to have a growth rate greater than zero.

Description of Common Targets for *P. aeruginosa* Networks

In this subsection, common targets for all Tier 2 *P. aeruginosa* networks are listed. The metabolic network models of *P. aeruginosa* analyzed in this subsection are described at Oberhardt et al., 2008 (PAO1) and Bartell et al., 2017 (PAO1 and PA14). The *P. aeruginosa* CCBH4851 metabolic network is being modeled by our group and represents a bacterium found in a catheter of a patient hospitalized at the Brazilian state of Goiás (Silveira et al., 2014). It is worth highlighting that the Bartell et al. (2017) networks focused on modeling virulence factors. Due to this fact, the biomass equation received less attention and the growth rate is not inside the range observed for *Pseudomonas* spp. Nevertheless, the workflow was able to process both networks and found several targets common to other metabolic reconstructions. The number of unique targets found in each network, for both FAB+FVA and FBA-only methods, are listed in **Table 1**. The spreadsheet detailing all targets found is available as supplementary material.

For the FBA-only method, 25 targets are common to all four networks. For the FBA+FVA method, 11 targets are common to all four networks.

It is important to highlight some of the genes identified as common targets for all four metabolic network models of *P. aeruginosa* (**Table 2**). The *murA* (EC 2.5.1.7) and *murB* (EC 1.3.1.98) genes encode enzymes involved in bacterial cell wall synthesis and have been identified as essential in both *Pseudomonas* spp. and *Escherichia coli* (Benson et al., 1996). The *folP* gene product (EC 2.5.1.15) is important for folic acid biosynthesis, which is fundamental for bacterial growth and reproduction (Dallas et al., 1992). The *folA* gene product (EC 1.5.1.3) is related to the biosynthesis of cofactors, being an important intermediary of folate metabolism. It is considered the key enzyme of this process and essential for microbial growth (Myllykallio et al., 2003). Another target worth mentioning is the *aroE* gene (EC 1.1.1.25), which has been described as a potential therapeutic target of both *P. putida* and *E. coli* (Peek et al., 2014).

Table 3 shows common targets with approved drugs. It is worth mentioning that several approved drugs have been identified; some of them are potential candidates for drug repositioning. Another relevant remark is the fact that most targets are also associated with experimental drugs.

Another noteworthy observation is that a considerable number of approved drugs in **Table 3** are most probably artifacts from the DrugBank database. For instance, flavin adenine dinucleotide (FAD), listed as an approved drug related to gene *murB*, is in fact approved for use in Japan under the trade name adeflavin as an ophthalmic treatment for vitamin B2 deficiency, it is just a cofactor for the product of gene *murB*, the enzyme UDP-N-acetylenolpyruvoylglucosamine reductase. All similar cases are highlighted with double asterisks in **Table 3**. This observation only reinforces a known limitation of all computational methods relying on databases at least partially annotated using automated workflows.

TABLE 1 | Number of unique targets found in the Tier 2 metabolic networks of *P. aeruginosa*.

	FBA-Only	FBA+FVA
PAO1-2008	53	50
PAO1-2017	50	42
PA14	44	42
CCBH4851	50	17

TABLE 2 | Potential targets common to all Tier 2 *P. aeruginosa* metabolic network models. Common targets identified by both FBA-only and FBA+FVA methods are marked with asterisks (*). The other targets were identified by the FBA-only method but not by the FBA+FVA method.

EC Number	Gene Name	Product	DrugBank Inhibitor
1.1.1.100	*fabG*	3-oxoacyl-[acyl-carrier-protein] reductase FabG	E
1.1.1.25	*aroE*	Shikimate dehydrogenase	E
1.17.1.8	*dapB*	4-hydroxy-tetrahydrodipicolinate reductase	E
1.3.1.98	*murB*	UDP-N-acetylenolpyruvoylglucosamine reductase	A/E
1.5.1.3	*folA*	Dihydrofolate reductase	A/E
2.1.1.45	*thyA**	Thymidylate synthase	E
2.3.1.41	*fabB*	3-oxoacyl-[acyl-carrier-protein] synthase 1	A/E
2.4.1.227	murG	UDP-N-acetylglucosamine–N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	E
2.4.2.14	*purF**	Amidophosphoribosyltransferase	E
2.5.1.15	*folP**	Dihydropteroate synthase	A
2.5.1.6	*metK**	S-adenosylmethionine synthase	E
2.5.1.7	*murA*	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	A/E
2.6.1.16	*glmS*	Glutamine—fructose-6-phosphate aminotransferase [isomerizing]	E
2.6.1.85	*pabB*	Para-aminobenzoate synthase component 1	A
2.7.4.25	*cmk*	Cytidylate kinase	E
2.7.7.23	*glmU**	Bifunctional protein GlmU	E
3.1.3.1	*phoA**	Alkaline phosphatase	E
4.1.3.38	*pabC**	Aminodeoxychorismate lyase	E
4.2.1.24	*hemB**	Delta-aminolevulinic acid dehydratase	A/E
4.2.3.5	*aroC*	Chorismate synthase	A
5.3.1.1	*tpiA*	Triosephosphate isomerase	E
5.3.1.6	*rpiA*	Ribose-5-phosphate isomerase A	A/E
6.3.2.13	*murE**	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate–2,6-diaminopimelate ligase	E
6.3.2.8	*murC**	UDP-N-acetylmuramate–L-alanine ligase	E
6.3.2.9	*murD**	UDP-N-acetylmuramoylalanine–D-glutamate ligase	E

The EC (Enzyme Commission) numbers represent the classification of P. aeruginosa enzymes according to the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Gene, product, and DrugBank Inhibitor Status were retrieved from UniProt and DrugBank databases, respectively. Abbreviations: experimental (E) and approved (A).

Analysis of the Metabolic Network Model of the Multidrug-Resistant Strain *P. aeruginosa* CCBH4851

Considering the curated version of the metabolic network of multidrug-resistant strain *P. aeruginosa* CCBH4851, 17 unique

TABLE 3 | Putative targets with approved drugs common to all Tier 2 metabolic network models of *P. aeruginosa*. Targets marked with asterisks are also associated with drugs in the experimental stage. Drugs marked with double asterisks are most probably artifacts inherited from DrugBank.

EC number	Gene name	Approved drug
1.3.1.98	*murB**	Flavin adenine dinucleotide**
1.5.1.3	*folA**	Levoleucovorin
1.5.1.3	*folA**	Isoniazid
2.3.1.41	*fabB**	Cerulenin
2.5.1.15	*folP*	Sulfacytine
2.5.1.15	*folP*	Sulfaphenazole
2.5.1.15	*folP*	Sulfamethoxazole
2.5.1.15	*folP*	Sulfanilamide
2.5.1.15	*folP*	Sulfacetamide
2.5.1.15	*folP*	Sulfamethazine
2.5.1.15	*folP*	Sulfamethizole
2.5.1.15	*folP*	Sulfisoxazole
2.5.1.15	*folP*	Sulfamerazine
2.5.1.7	*murA**	Fosfomycin
2.6.1.85	*pabB*	Formic acid**
4.2.1.24	*hemB**	Formic acid**
4.2.3.5	*aroC*	Riboflavin monophosphate**
5.3.1.6	*rpiA**	Citric acid**

The EC (enzyme commission) numbers represent the classification of P. aeruginosa enzymes according to the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB).Gene and associated drug names were retrieved from UniProt and DrugBank databases, respectively.

targets were identified using the FBA+FVA method, while the FBA-only method returned 50 unique potential targets. Among those results, it is important to highlight four potential targets: *asd*, *ispE*, *fabA*, and *dapA*. Both *asd* and *dapA* are involved in the L-lysine biosynthesis *via* DAP pathway, which synthesizes L-lysine from aspartate and pyruvate. In bacteria, the lysine biosynthesis pathway yields the important metabolites meso-2,6-diaminopimelate (meso-DAP) and lysine. Lysine is utilized for protein synthesis in bacteria and forms part of the peptidoglycan cross-link structure in the cell wall of most gram-positive species, whilst meso-DAP is the peptidoglycan cross-linking moiety in the cell wall of gram-negative bacteria (Dogovski et al., 2012). This pathway is utilized by most bacteria, some archaea, some fungi, some algae, and plants (Liu et al., 2010b), and therefore are suitable candidates for therapeutic targets. Only experimental drugs are available to both targets.

ispE encodes a cytoplasmic kinase of the MEP pathway that is involved in the biosynthesis of the isoprenoids used by many gram-negative bacteria (including *P. aeruginosa*) (Heuston et al., 2012). Because isoprenoids are involved in a wide variety of vital biological functions, the seven enzymes without close human homologs that participate in their metabolism (encoded by *dxr, ispC, ispD, ispE, ispF, ispG, ispH* genes) are favorable candidate drug targets and several inhibitors have been already reported (Masini and Hirsch, 2014). Specifically for *ispE*, only experimental drugs are available.

fabA participates in fatty acid synthesis (FAS) processes, which includes also *fabB, fabD, fabI*, and *fabH*. The proteins encoded by these genes have an essential role during the synthesis of bacterial phospholipid membranes, lipopolysaccharide (LPS), and lipoproteins, thus representing attractive targets due to the structural differences between the human and bacterial proteins and the essentiality of FAS (Zhang et al., 2006; Leibundgut et al., 2008). Only experimental drugs are available to this target.

All four potential targets described above are reported to be overexpressed in *K. pneumoniae* when the pathogen is exposed to polymyxin B (Ramos et al., 2018), which is considered as a "last resort" antibiotic for infections caused by Carbapenem-resistant *Enterobacteriaceae*. Indeed, it has been shown that *P. aeruginosa* CCBH4851 is sensible only to polymyxin B (Silveira et al., 2014). This observation can be of interest in a combination therapy perspective when dealing with resistant *P. aeruginosa* infections, possibly acting synergistically with other drugs. An interesting observation is that the same target may be associated with similar reactions in both Tier 2 P. *aeruginosa* CCBH4851 and *K. pneumoniae*. For instance, *asd* is associated with the aspartate-semialdehyde dehydrogenase reaction in both metabolic networks, but reactants, products, and directionality differ. On the other hand, reactions associated with *fabA* differ in both metabolic network models. The gene *fabA* is associated to 13 reactions in *K. pneumoniae* and nine reactions in Tier 2 *P. aeruginosa* CCBH4851.

Another interesting observation is that the above targets have been identified by the FBA-only method. Only *dapA* is included in FBA+FVA results. One possible inference from this fact is that *dapA* should be prioritized over the other targets. Nevertheless, it also highlights the importance of considering both methods when looking for new potential targets.

A fifth target worth mentioning is *algC*, which encodes a highly reversible phosphoryltransferase. The phosphomannomutase activity produces a precursor for alginate polymerization; the alginate layer causes a mucoid phenotype and provides a protective barrier against host immune defenses and antibiotics. It is involved in core LPS biosynthesis due to its phosphoglucomutase activity and is essential for rhamnolipid production, an exoproduct correlated with pathogenicity (Olvera et al., 1999). It is also required for biofilm production (Davies and Geesey, 1995). This particular target was identified using the FBA-only method. Only experimental drugs are available to *algC*.

Analysis of the Tier 3 *P. aeruginosa* CCBH4851 Metabolic Network

To evaluate the robustness of FindTargetsWEB regarding Tier 3 networks, which generally are networks generated automatically without manual curation, FindTargetsWEB processed a preliminary version of the metabolic network model of *P. aeruginosa* CCBH4851, which precedes the Tier 2 network described previously. This network is the only one in this paper which was processed using step 6b (algorithm 2) of the overall method. The growth rate of the Tier 3 version of the *P. aeruginosa* CCBH4851 network is 1.757 h^{-1}, which is less consistent to the biology of *P. aeruginosa* spp. than the growth rate obtained by the Tier 2 version of the network. The processing of this Tier 3 network generated 32 targets in the FBA+FVA analysis, and

48 targets using the FBA-only method. It is remarkable that this less curated version of *P. aeruginosa* CCBH4851 network generated more potential targets in the FBA+FVA analysis than the corresponding Tier 2 network.

Among targets identified using the FVA+FBA method, 10 targets are common between the Tier 2 and Tier 3 networks. For the FBA-only analysis, 21 targets are common between the two versions. It is worth mentioning that many targets found in Tier 2 networks are present in the analysis of the CCBH4851 Tier 3 network, which corroborates the relevance of the targets found even in draft versions of metabolic networks. This comparison also highlights the importance of careful curation of automatically generated metabolic networks. For instance, from the targets discussed in the previous subsection, only *dapA* is present as a potential target in the Tier 3 network.

Analysis of Metabolic Network Models of *K. pneumoniae* and *H. influenzae*

Metabolic networks of bacteria other than *P. aeruginosa* were also processed using FindTargetsWEB. In the previous subsections, results for *P. aeruginosa* metabolic network models were presented, but it is also possible to analyze networks of other species of bacteria. In this subsection, FindTargetsWEB results for a metabolic network reconstruction of *K. pneumoniae* MGH78578—iYL1228 (BioModels ID 1507180054) (Liao et al., 2011) and *H. influenzae*—iCS400 (BioModels ID 1507180053) (Schilling and Palsson, 2000) are presented (**Table 4**).

For *K. pneumoniae,* a total of 45 unique potential targets were found using the FBA+FVA method and also 45 for the FBA-only method. Some of the more representative targets are listed in **Table 4** (complete results are available as **Supplementary Material**).

Several targets identified in **Table 4** are worth mentioning. For instance, the cytoplasmic enzyme encoded by *lpxA* gene is involved in the initial steps of lipid A production through the Raetz pathway. As stated in the previous subsection, *fabA*, *fabB*, and *fabF* participate in FAS processes and represent attractive targets due to the structural differences between the human and bacterial proteins and the essentiality of FAS. The cytoplasmic protein N-acetylglutamate (NAG) kinase (encoded by *argB*), which promotes phosphorylation of NAG in a rate-limiting step of bacterial L-arginine production, occurs through acetylated intermediates, unlike mammals which use non-acetylated intermediates, and for this reason, it was previously considered a candidate drug target (Marcos et al., 2010). Indeed, Ramos et al. (2018) identified several potential targets found by FindTargetsWEB as priority targets for *K. pneumoniae*. Examples are *dapD, lpxA, fabA, fabB, tmk, murE,* and *murD*. Their analysis included a reconstruction of the metabolic network model of *K. pneumoniae* Kp13 and an essentiality analysis based on literature search. A target prioritization pipeline was proposed that takes into account gene essentiality, topological measures, literature information, and gene expression data. It is worth noting that neither FBA nor FVA were used in their analysis.

For the metabolic network model of *H. influenzae*, 16 unique potential targets were found by FindTargetsWEB for both FBA+FVA and FBA-only methods (**Table 4**). Complete results are available as **Supplementary Material**.

It is worth mentioning that the genes *folA, tmk, kdsB, metG, thrS*, and *guaA* were identified as essential for *H. influenzae* growth and survival by Akerley and colleagues (2002), using a high-density

TABLE 4 | List of EC numbers, product, and DrugBank inhibitor status for putative targets for metabolic network models of *K. pneumoniae* and *H. influenzae*. All targets listed in this table are included in the results of both FBA+FVA and FBA-only methods.

EC number	Gene name	Product	DrugBank inhibitor	Species
1.3.1.98	*murB*	UDP-N-acetylenolpyruvoylglucosamine reductase	A/E	*K. pneumoniae*
2.3.1.117	*dapD*	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	E	*K. pneumoniae*
2.3.1.129	*lpxA*	Acyl-[acyl-carrier-protein]–UDP-N-acetylglucosamine O-acyltransferase	E	*K. pneumoniae*
2.3.1.179	*fabF*	3-oxoacyl-[acyl-carrier-protein] synthase 2	A/E	*K. pneumoniae*
2.3.1.41	*fabB*	3-oxoacyl-[acyl-carrier-protein] synthase 1	A/E	*K. pneumoniae*
2.7.2.8	*argB*	Acetylglutamate kinase	E	*K. pneumoniae*
2.7.4.9	*tmk*	Thymidylate kinase	E	*K. pneumoniae*
4.2.1.59	*fabA*	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase	E	*K. pneumoniae*
6.3.2.13	*murE*	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate–2,6-diaminopimelate ligase	E	*K. pneumoniae*
6.3.2.8	*murC*	UDP-N-acetylmuramate–L-alanine ligase	E	*K. pneumoniae*
6.3.2.9	*murD*	UDP-N-acetylmuramoylalanine–D-glutamate ligase	E	*K. pneumoniae*
1.5.1.3	*folA*	Dihydrofolate reductase	A/E	*H. influenzae*
2.7.4.9	*tmk*	Thymidylate kinase	E	*H. influenzae*
2.7.7.38	*kdsB*	3-deoxy-manno-octulosonate cytidylyltransferase	E	*H. influenzae*
6.1.1.10	*metG*	Methionine–tRNA ligase	E	*H. influenzae*
6.1.1.2	*trpS*	Tryptophan–tRNA ligase	A/E	*H. influenzae*
6.1.1.21	*hisS*	Histidine–tRNA ligase	E	*H. influenzae*
6.1.1.3	*thrS*	Threonine–tRNA ligase	E	*H. influenzae*
6.3.5.2	*guaA*	GMP synthase [glutamine-hydrolyzing]	A	*H. influenzae*

The EC (Enzyme Commission) numbers represent the classification of K. pneumoniae and H. influenzae enzymes according to the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Gene, product, and DrugBank Inhibitor Status were retrieved from UniProt and DrugBank databases, respectively. Abbreviations: experimental (E) and approved (A).

transposon mutagenesis strategy. Another relevant observation is the presence of potential targets common to *K. pneumoniae* (*tmk*) and *P. aeruginosa* (*folA*). Both methods, FBA+FVA and FBA-only, generate exactly the same results. Therefore, the FVA ranges for all targets in **Table 4** are equal to zero.

Analysis of a Host-Pathogen Integrated Metabolic Network Model

FindTargetsWEB is also capable of processing integrated metabolic network models. The analysis presented in this subsection used a host-pathogen genome-scale reconstruction, iAB-AMØ-1410-Mt-661 (BIOMODELS ID 1011090001), which integrates a cell-specific alveolar macrophage model, iAB-AMØ-1410, from the global human metabolic reconstruction, with an *M. tuberculosis* H37Rv model, iNJ661 (Bordbar et al., 2010). The integrated host-pathogen network enables simulation of the metabolic changes during infection.

A total of 35 unique potential targets was identified by FindTargetsWEB on the integrated model by both the FBA+FVA and FBA-only methods (complete results are available as **Supplementary Material**). Several potential targets found by FindTargetsWEB in the host-pathogen integrated model have been previously reported in the literature as essential to *M. tuberculosis* survival (Bordbar et al., 2010; Sassetti and Rubin, 2003). Examples are *nrdE*, *mmaA2*, *mmaA3, aroQ,* and *ahcY,* from which only *mmaA2* and *mmaA3* have approved drugs. It is worth highlighting that the selection of potential targets of FindTargetsWEB depends not only on network analysis, but also on data retrieved from DrugBank and additional filters, such as a low level of similarity with human proteins.

Analysis of the Metabolic Network Model of a Gram-Positive Bacterium

None of the results presented in the previous subsections include gram-positive bacteria. To emphasize FindTargetsWEB flexibility, in this subsection, we present results from the metabolic network model analysis of a gram-positive pathogen. *S. aureus* is a pathogenic gram-positive bacterium that causes a variety of disease conditions both in hospital settings and in the community at large. The metabolic model iSB619 (BIOMODELS ID 1507180070) (Becker and Palsson, 2005), reconstructed from the strain N315, was processed using FindTargetsWEB. Complete results for both FBA-only and FVA+FBA are available as **Supplementary Material**.

A total of 27 unique potential targets were generated using the FBA-only method. The FBA+FVA analysis returned 22 unique targets. Some potential targets are common to gram-negative bacteria (such as *murB*, *aroC*), while others such as *mvaA* (locus tag SA2333 for the N315 strain, SAOUHSC_02859 for the NCTC8325 strain), *tkt* (SA1177, SAOUHSC_01337), and *dfrA* (SA1259, SAOUHSC_01434) are defined as essential for *S. aureus* in both minimal and rich medias (Becker and Palsson, 2005). Regarding the metabolic network models analyzed in this manuscript, the potential targets *mvaA*, *tkt*, and *dfrA* only appear in the *S. aureus* metabolic network model.

Analysis of the Metabolic Network Model of a Non-Pathogenic Bacteria

The pseudomonads include a diverse set of bacteria whose metabolic versatility and genetic plasticity have enabled their survival in a broad range of environments. Many members of this family are able to either degrade toxic compounds or to efficiently produce high value compounds and are therefore of interest for both bioremediation and bulk chemical production. *P. putida* is a representative of those industrially relevant pseudomonads. In this subsection, an analysis of the metabolic network model of the *P. putida* KT2440 (Puchałka et al., 2008), named iJP815 (BIOMODELS ID 1507180044), is compared to the previous analysis of a pathogenic member of the family, *P. aeruginosa*. Complete results for the analysis of the *P. putida* metabolic network model is available as supplementary material.

A first comparison between *P. putida* e *P. aeruginosa* metabolic network models is the number of potential targets. The analysis of the metabolic network model of *P. putida* returned a comparable number of potential targets: 52 for FBA-only, 50 for the FBA+FVA method (see **Table 1**). Indeed, the size of the metabolic network model iJP815 is comparable with other *P. aeruginosa* metabolic networks: 824 intracellular and 62 extracellular metabolites connected by 877 reactions. Other interesting observation is that some targets present in the multidrug-resistant *P. aeruginosa* CCBH4851 are absent in *P. putida*, despite the comparable number of potential targets. Remarkable examples are *asd*, *ispE*, *fabA*, *dapA*, and *algC*. Indeed, from the 25 targets common to all Tier 2 *P. aeruginosa* metabolic network model displayed in **Table 2** (FBA-only method), only 18 are also potential targets for the *P. putida* KT2440 metabolic network model.

DISCUSSION

Several advantages of the proposed method can be highlighted: first the robustness of the system, which can identify potential targets even for draft (Tier 3) networks, pointing out that such metabolic network models are very common and are the only models available for some organisms. The system is deployed as a web application and is asynchronous: the user is notified when results are available. The performance of the system is optimized, since the COBRApy framework can make use of multiple cores available in the host machine, and it is able to process the metabolic network of various bacteria, as described in the previous section. The only requirement is the availability of an SBML level 3 file describing the corresponding genome-scale metabolic network. The user interface is straightforward (see **Figures 1** and **2**), and the user should only provide a name, an e-mail address, and the corresponding SBML file. The user should also indicate the species of bacterium associated with the metabolic network model. FindTargetsWEB is a highly flexible tool, capable of processing genome-scale metabolic network models of gram-negative bacteria, gram-positive bacteria, bacteria not classified as either gram-positive or gram-negative, and even integrated host-pathogen genome-scale metabolic network models.

Other proposals for the analysis of metabolic networks at genomic scale are available in the literature. Chavali et al. (2012) used FBA and FVA for identification of potential targets, but their application does not propose any drugs for the targets found neither describes the potential targets in detail. The procedure reported in (Oberhardt et al., 2010) describes a processing similar to the one proposed in this work up to the EC number mapping step, and then uses graphical tools to identify the potential targets for *E. coli* and *Bacillus subtilis*, without pinpointing any potential drug. Ramos et al. (2018) propose a method to identify drug targets in metabolic network model of *K. pneumoniae*. However, their method is not automated, and it was not applied to other species of bacteria. None of these works go as far as FindTargetsWEB, which can process metabolic network models of several species of bacteria, identify potential targets, confirm homology with the analyzed gene, and identify all available drugs for the potential target in a fully automated manner.

Regarding the options to identify potential targets, i.e., *FBA+FVA* and *FBA-only*, one can conclude that the FBA+FVA method represents a way to prioritize the targets identified by the FBA-only method, since the set of targets identified by FBA+FVA is a proper subset of the set of targets identified by FBA-only. However, as stated in the detailed description of the targets of the Results section, potential targets that are associated with the FBA-only method and do not appear as results of the FBA+FVA method should not be disconsidered. Many important targets described in the literature have a FVA range greater than zero, and a careful analysis of both sets of potential targets is advised.

Several of the approved drugs identified by FindTargetsWEB are already used against *P. aeruginosa* and other bacteria and can be effective against non multidrug-resistant strains. As expected, for the multidrug-resistant strain, most of the approved drugs are not effective. For instance, it is known that *P. aeruginosa* can be resistant to both to trimethoprim and sulfamethoxazole (see **Table 3**) due to the MexAB-OprM multidrug efflux system (Köhler et al., 1996). Nevertheless, FindTargetsWEB also pinpoints a large number of experimental drugs that can be effective. Actually, most of the targets identified by FindTargetsWEB for all strains are associated to experimental drugs and may represent new therapeutic options. Clearly, additional *in vitro* and *in vivo* testings are needed in order to confirm the experimental drugs as new therapeutic options.

Additional information provided by FindTargetsWEB can also be considered in the definition of new strategies to fight multidrug-resistant bacteria. Information such as pathway, target function, and catalytic activity can be considered in order to devise a multi-target strategy, which can be very effective in some scenarios. As an example of a multi-target strategy, in bacteremia caused by *P. aeruginosa*, the combination of efflux pump inhibitors and iron chelators has been proposed to control the infection process in view of the overexpression of the MexAB-OprM efflux system during iron deprivation (Liu et al., 2010a). Indeed, several targets in the analysis of results for *P. aeruginosa* are related to different cellular functions. Targeting several cellular functions and processes at the same time can be a more promising strategy than considering only one isolated target. For instance, it is known that inhibiting bacterial growth can accelerate the process of biofilm formation (Xu et al., 2013). Therefore, the pathogen can form a biofilm before it is eliminated. Multi-target therapies are already commonplace in treating bacteria infections, and the wealth of information provided by FindTargetsWEB can be used to define new multi-target treatments not considered before. For instance, *algC* (*P. aeruginosa* CCBH4851, PA14, and PAO1-2017 metabolic networks) is both essential to metabolic growth and biofilm formation, according to the FUNCTION field returned by FindTargetsWEB and literature sources (Davies and Geesey, 1995). Therefore, a targeting strategy based on other genes may consider also targeting *algC* to prevent biofilm formation.

CONCLUDING REMARKS

FindTargetsWEB is a user-friendly web application that combines bioinformatics and systems biology, providing insights of new therapeutic targets for multidrug-resistant bacteria, increasing the available therapeutic options. By identifying more effectively potential targets along with candidate active compounds for posterior experimental confirmation, this tool prevents exhaustive bacterial drug screening. Importantly, FindTargetWEB can also be applied to the study of other bacteria due to the flexibility proposed by computational modeling, serving as a base for other relevant studies. In addition, it will serve as a starting point for the creation of even more complete applications in a web environment, such as one capable of processing integrated computational models and retrieving data from more databases.

AVAILABILITY AND REQUIREMENTS

Project name: FindTargetsWEB

Project home page: http://pseudomonas.procc.fiocruz.br/FindTargetsWEB

Operating system: e.g. Web-based, Platform independent

Programming language: Python 3.6

Other requirements: An updated web browser (e.g. Google Chrome, Mozilla Firefox, Apple Safari, Microsoft Edge)

License: Not Applicable

Any restriction to use by non-academics: Not Applicable

The user must provide a SBML level 3 file describing the metabolic network reconstruction and an e-mail address to which the results will be forwarded.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript and the supplementary files.

AUTHOR CONTRIBUTIONS

TCM, FPSJ, and FABS designed the system. TCM was the main programmer. MWC and ADC-A tested the system and evaluated its correctness. All authors have equally participated in the writing of this paper.

FUNDING

The authors would like to thank CAPES, FAPERJ, CNPq and FIOCRUZ (INOVA-FIOCRUZ VPPCB-007-FIO-18-2-29) for financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00633/full#supplementary-material

REFERENCES

Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., Judson, N., and Mekalanos J. J. (2002). A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci.* 99 (2), 966–971. doi: 10.1073/pnas.012602299

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Bartell, J. A., Blazier, A. S., Yen, P., Thøgersen, J. C., Jelsbak, L., Goldberg, J. B., et al. (2017). Reconstruction of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. *Nat. Commun* 8, 14631. doi: 10.1038/ncomms14631

Becker, S. A., and Palsson, BØ. (2005). Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* 5 (1), 8. doi: 10.1186/1471-2180-5-8

Benson, T. E., Walsh, C. T., and Hogle, J. M. (1996). The structure of the substrate-free form of MurB, an essential enzyme for the synthesis of bacterial cell walls. *Structure* 4 (1), 47–54. doi: 10.1016/S0969-2126(96)00008-1

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, BØ., and Jamshidi, N. (2010). Insight into human alveolar macrophage and M. tuberculosis interactions *via* metabolic reconstructions. *Mol. Syst. Biol.* 6 (1), 422. doi: 10.1038/msb.2010.68

Brown, A. D. (1957). Some general properties of a psychrophilic pseudomonad: the effects of temperature on some of these properties and the utilization of glucose by this organism and *Pseudomonas aeruginosa Microbiology* 17 (3), 640–648. doi: 10.1099/00221287-17-3-640

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., et al. (2015). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 44 (D1), D471–D480. doi: 10.1093/nar/gkv1164

Chavali, A. K., D'Auria, K. M., Hewlett, E. L., Pearson, R. D., and Papin, J. A. (2012). A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends Microbiol.* 20, 113–123. doi: 10.1016/j.tim.2011.12.004

Dallas, W. S., Gowen, J. E., Ray, P. H., Cox, M. J., and Dev, I. K. (1992). Cloning, sequencing, and enhanced expression of the dihydropteroate synthase gene of *Escherichia coli* MC4100. *J. Bacteriol.* 174 (18), 5961–5970. doi: 10.1128/jb.174.18.5961-5970.1992

Davies, D. G., and Geesey, G. G. (1995). Regulation of the alginate biosynthesis gene algC in *Pseudomonas aeruginosa* during biofilm development in continuous culture. *Appl. Environ. Microbiol.* 61 (3), 860–867.

Dogovski, C., Atkinson, S. C., Dommaraju, S. R., Downton, M., Hor, L., Moore, S., et al. (2012). Enzymology of bacterial lysine biosynthesis. *Biochemistry* 1, 225–262. doi: 10.5772/34121

Ebrahim, A., Lerman, J. A., Palsson, BØ, and Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* 7, 74–79. doi: 10.1186/1752-0509-7-74

Glont, M., Nguyen, T. V. N., Graesslin, M., Hälke, R., Ali, R., Schramm, J., et al. (2017). BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Res.* 46 (D1), D1248–D1253. doi: 10.1093/nar/gkx1023

Gudmundsson, S., and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinf.* 11(1), 489. doi: 10.1186/1471-2105-11-489

Heuston, S., Begley, M., Gahan, C. G., and Hill, C. (2012). Isoprenoid biosynthesis in bacterial pathogens. *Microbiology* 158 (6), 1389–1401. doi: 10.1099/mic.0.051599-0

Hucka, M., Bergmann, F. T., Hoops, S., Keating, S. M., Sahle, S., Schaff, J. C., et al. (2015). The Systems Biology Markup Language (SBML): language specification for level 3 version 1 core. *J. Integrat. Bioinf.* 12 (2), 382–549. doi: 10.1515/jib-2015-266

Hyduke, D., Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth J., et al. (2011). COBRA Toolbox 2.0. *Protoc. Exch.* 22. doi: 10.1038/protex.2011.234

Kerr, K. G., and Snelling, A. M. (2009). *Pseudomonas aeruginosa*: a formidable and ever-present adversary. *J. Hosp. Infect.* 73, 338–344. doi: 10.1016/j.jhin.2009.04.020

Köhler, T., Kok, M., Michea-Hamzehpour, M., Plesiat, P., Gotoh, N., Nishino, T., et al. (1996). Multidrug efflux in intrinsic resistance to trimethoprim and sulfamethoxazole in *Pseudomonas aeruginosa Antimicrob. Agents Chemother.* 40 (10), 2288–2290. doi: 10.1128/AAC.40.10.2288

Kozakov, D., Hall, D. R., Napoleon, R. L., Yueh, C., Whitty, A., and Vajda, S. (2015). New frontiers in druggability. *J. Med. Chem.* 58 (23), 9063–9088. doi: 10.1021/acs.jmedchem.5b00586

Leibundgut, M., Maier, T., Jenni, S., and Ban, N. (2008). The multienzyme architecture of eukaryotic fatty acid synthases. *Curr. Opin. Struct. Biol.* 18 (6), 714–725. doi: 10.1016/j.sbi.2008.09.008

Liao, Y.-C., Huang, T.-W., Chen, F.-C., Charusanti, P., Hong, J. S. J., Chang, H.-Y., et al. (2011). An experimentally validated genome-scale metabolic reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *J. Bacteriol.* 193, 1710–1717. doi: 10.1128/JB.01218-10

Liu, Y., Yang, L., and Molin, S. (2010a). Synergistic activities of an efflux pump inhibitor and iron chelators against *Pseudomonas aeruginosa* growth and biofilm formation. *Antimicrob. Agents Chemother.* 54, 3960–3963. doi: 10.1128/AAC.00463-10

Liu, Y., White, R. H., and Whitman, W. B. (2010b). Methanococci use the diaminopimelate aminotransferase (DapL) pathway for lysine biosynthesis. *J. Bacteriol.* 192 (13), 3304–3310. doi: 10.1128/JB.00172-10

Marcos, E., Ramon, C., and Ivet, B. (2010). On the conservation of the slow conformational dynamics within the amino acid kinase family: NAGK the paradigm. *PLoS Computat. Biol* 6 (4), e1000738. doi: 10.1371/journal.pcbi.1000738

Masini, T., and Hirsch, A. K. (2014). Development of inhibitors of the 2 C-Methyl-d-erythritol 4-phosphate (MEP) pathway enzymes as potential anti-infective agents. *J. Med. Chem.* 57 (23), 9740–9763. doi: 10.1021/jm5010978

Myllykallio, H., Leduc, D., Filee, J., and Liebl, U. (2003). Life without dihydrofolate reductase FolA. *Trends Microbiol.* 11, 220–223. doi: 10.1016/S0966-842X(03)00101-X

Oberhardt, M. A., Puchałka, J., Fryer, K. E., Martins Dos Santos, V. A. P., and Papin, J. A. (2008). Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* 190, 2790–2803. doi: 10.1128/JB.01583-07

Oberhardt, M. A., Goldberg, J. B., Hogardt, M., and Papin, J. A. (2010). Metabolic network analysis of *Pseudomonas aeruginosa* during chronic cystic fibrosis lung infection. *J. Bacteriol.* 192 (20), 5534–5548. doi: 10.1128/JB.00900-10

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Olvera, C., Goldberg, J. B., Sánchez, R., and Soberón-Chávez, G. (1999). The *Pseudomonas aeruginosa* algC gene product participates in rhamnolipid biosynthesis. *FEMS Microbiol. Lett.* 179 (1), 85–90. doi: 10.1111/j.1574-6968.1999.tb08712.x

Orth, J. D., Thiele, I., and Palsson, BØ (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614

Peek, J., Shi, T., and Christendat, D. (2014). Identification of novel polyphenolic inhibitors of shikimate dehydrogenase (AroE). *J. Biomol. Screen* 19 (7), 1090–1098. doi: 10.1177/1087057114527127

Puchałka, J., Oberhardt, M. A., Godinho, M., Bielecka, A., Regenhardt, D., Timmis, K. N., et al. (2008). Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology. *PLoS Computat. Biol.* 4 (10), e1000210. doi: 10.1371/journal.pcbi.1000210

Ramos, P. I. P., Do Porto, D. F., Lanzarotti, E., Sosa, E. J., Burguener, G., Pardo, A. M., et al. (2018). An integrative, multi-omics approach towards the prioritization of *Klebsiella pneumoniae* drug targets. *Sci. Rep.* 8 (1), 10755. doi: 10.1038/s41598-018-28916-7

Rienksma, R. A., Suarez-Diez, M., Spina, L., Schaap, P. J., and Martins Dos Santos, V. A. P. (2014). Systems-level modeling of mycobacterial metabolism for the identification of new (multi-)drug targets. *Semin. Immunol.* 26, 610–622. doi: 10.1016/j.smim.2014.09.013

Sassetti, C. M., and Rubin, E. J. (2003). Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12989–12994. doi: 10.1073/pnas.2134250100

Schilling, C. H., and Palsson, BØ. (2000). Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.* 203, 249–283. doi: 10.1006/jtbi.2000.1088

Seto, M., and Noda, M. (1982). Growth rate, biomass production and carbon balance of *Pseudomonas aeruginosa* at pH extremes in a carbon-limited medium. *Jap. J. Limnol.* 43(4), 263–271. doi: 10.3739/rikusui.43.263

Silva, F. A. B., Medeiros Filho, F., Merigueti, T. C., Giannini, T., Brum, R., de Faria, L. M. et al., (2018). "Computational Modeling of Multidrug-Resistant Bacteria," in *Theoretical and Applied Aspects of Systems Biology*. Eds. F. A. B. Silva, N. Carels, and F. Silva-Jr, Cham, Switzerland: Springer International Publishing AG, 195–220. doi: 10.1007/978-3-319-74974-7_11

Silveira, M., Albano, R., Asensi, M., and Assef, A. P. C. (2014). The draft genome sequence of multidrug-resistant *Pseudomonas aeruginosa* strain CCBH4851, a nosocomial isolate belonging to clone SP (ST277) that is prevalent in Brazil. *Mem. Inst. Oswaldo Cruz* 190, 1086–1087. doi: 10.1590/0074-0276140336

Stone, P. W., Braccia, D., and Larson, E. (2005). Systematic review of economic analyses of health care-associated infections. *Am. J. Infect. Control* 33, 501–509. doi: 10.1016/j.ajic.2005.04.246

Thiele, I., and Palsson, BØ. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi: 10.1038/nprot.2009.203

UniProt Consortium (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46 (5), 2699. doi: 10.1093/nar/gky092

Van Rossum, G., and Drake, F. L. (2003). An introduction to Python. Bristol: Network Theory Ltd. 115 pp.

WHO (World Health Organization) (2014). *Antimicrobial resistance: global report on surveillance*. Geneva: World Health Organization.

WHO (World Health Organization) (2017). *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics*. Geneva: World Health Organization.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906. doi: 10.1093/nar/gkm958

Xu, Z., Fang, X., and Wood, T. K. (2013). Huang ZJ. A system-level approach for investigating *Pseudomonas aeruginosa* biofilm formation. *PLoS One* 8 (2), e57050. doi: 10.1371/journal.pone.0057050

Yang, L., Haagensen, J. A., Jelsbak, L., Johansen, H. K., Sternberg, C., Høiby, N., et al. (2008). In situ growth rates and biofilm development of *Pseudomonas aeruginosa* populations in chronic lung infections. *J. Bacteriol.* 190 (8), 2767–2776. doi: 10.1128/JB.01581-07

Zhang, Y. M., White, S. W., and Rock, C. O. (2006). Inhibiting bacterial fatty acid synthesis. *J. Biol. Chem.* 281 (26), 17541–17544. doi: 10.1074/jbc.R600004200

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

TECHNOLOGY AND CODE
published: 26 August 2019
doi: 10.3389/fgene.2019.00725

Gene Tags Assessment by Comparative Genomics (GTACG): A User-Friendly Framework for Bacterial Comparative Genomics

Caio Rafael do Nascimento Santiago[1,2], Renata de Almeida Barbosa Assis[3], Leandro Marcio Moreira[3,4]* and Luciano Antonio Digiampietri[1,5]*

[1] Bioinformatics Graduate Program, University of Sao Paulo, Sao Paulo, Brazil, [2] Adventist University of Sao Paulo, Sao Paulo, Brazil, [3] Biotecnology Graduate Program, Núcleo de Pesquisas em Ciências Biológicas, Federal University of Ouro Preto, Ouro Preto, Brazil, [4] Department of Biological Sciences, Federal University of Ouro Preto, Ouro Preto, Brazil, [5] School of Arts, Science, and Humanities, University of Sao Paulo, Sao Paulo, Brazil

Genomics research has produced an exponential amount of data. However, the genetic knowledge pertaining to certain phenotypic characteristics is lacking. Also, a considerable part of these genomes have coding sequences (CDSs) with unknown functions, posing additional challenges to researchers. Phylogenetically close microorganisms share much of their CDSs, and certain phenotypes unique to a set of microorganisms may be the result of the genes found exclusively in those microorganisms. This study presents the GTACG framework, an easy-to-use tool for identifying in the subgroups of bacterial genomes whose microorganisms have common phenotypic characteristics, to find data that differentiates them from other associated genomes in a simple and fast way. The GTACG analysis is based on the formation of homologous CDS clusters from local alignments. The front-end is easy to use, and the installation packages have been developed to enable users lacking knowledge of programming languages or bioinformatics analyze high-throughput data using the tool. The validation of the GTACG framework has been carried out based on a case report involving a set of 161 genomes from the Xanthomonadaceae family, in which 19 families of orthologous proteins were found in 90% of the plant-associated genomes, allowing the identification of the proteins potentially associated with adaptation and virulence in plant tissue. The results show the potential use of GTACG in the search for new targets for molecular studies, and GTACG can be used as a research tool by biologists who lack advanced knowledge in the use of computational tools for bacterial comparative genomics.

Keywords: user-friendly tools, systems biology, comparative genomics, orthologs, gene families

OPEN ACCESS

Edited by:
Vinicius Maracaja-Coutinho, University of Chile, Chile

Reviewed by:
George Colin DiCenzo, University of Florence, Italy
Haruo Suzuki, Keio University, Japan

****Correspondence:***
Caio Rafael do Nascimento Santiago
caio.santiago@usp.br
Leandro Marcio Moreira
lmmorei@ufop.edu.br

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

Received: *29 March 2019*
Accepted: *10 July 2019*
Published: *26 August 2019*

Citation:
Santiago CRdN, Assis RdAB, Moreira LM and Digiampietri LA (2019) Gene Tags Assessment by Comparative Genomics (GTACG): A User-Friendly Framework for Bacterial Comparative Genomics. Front. Genet. 10:725. doi: 10.3389/fgene.2019.00725

INTRODUCTION

Systems biology seeks to study the interaction between the components of a biological system holistically, mediated by several analytical tools, aiming the search for information capable of supporting the discovery of phenomena or complex biological processes (Chuang et al., 2010). Over the past years, such approaches, which have always developed from a multidisciplinary perspective,

have made possible great discoveries involving new biomarkers of selection and diseases, targets for drug development, among others, all concurrently with the development of the robust platforms and computational tools for analyzing high-throughput data (Berg, 2014).

Despite the advances mentioned above, some challenges still exist. Among these, the search for specific genes that may be associated with certain phenotypes stands out. Such a search is a non-trivial task because it consists of solving a multifactorial problem (Casadesús and Low, 2006). In microbiology, this challenge is even more pronounced, as the functional characteristics of a gene may be directly associated with the biological processes of biotechnological interest or that allow a better understanding of the host's immune response in the case of pathogenic microorganisms (Zamioudis and Pieterse, 2012; Campbell et al., 2017).

The development of new sequencing platforms in association with the set of "omics" sciences that seek to functionally analyze sequenced genes and genomes has substantially increased the volume of biological data available over the past years (Field et al., 2009). However, the understanding of genes' specific functions has advanced modestly, despite the efforts of the scientific community (Chervitz et al., 2011; Berger et al., 2013). This is justified by numerous factors that hinder gaining such understanding. Some of them are inherent to the limitations and constraints of molecular techniques (Tierney and Lamour, 2005). However, some of them arise from two factors: 1) the lack of robust data analysis tools for different biological questions, many of which are specific to a particular type of biological knowledge, or 2) the existence of data analysis tools that make interpreting the processing mechanism or displaying the results generated by such tools challenging (Hillmer, 2015).

To make experimental validation more assertive, scientists from different fields have developed computational tools that allow integrating biological data using complex algorithms and enabling user interaction through user-friendly interfaces. It is in such a context that the need for user-friendly tools applied to systems biology arises, developed with an intuitive interface that allows biologist users to perform complex analyses, guiding them to answer biological questions.

In this study, we present a new user-friendly tool named Gene-Tag Assessment by Comparative Genomics (GTACG) applied to genetics or systems biology and developed for the comparative analysis of bacterial genomes, aiming the selection of genes for studying correlation of presence or absence of genes with lifestyle, virulence, among other biological questions.

GTACG allows interactive analysis and data visualization, always considering the comparison of phenotypic groups. Different characteristics are considered in this process, such as the composition of gene families as well as their individual alignments and phylogeny, producing more robust data than binary metrics. The result of the execution pipeline is a static website, which allows gaining easy-to-share data and specific results through URLs.

GTACG produces phylogenies based on different characteristics, which allows for a more detailed analysis of phylogenetic relationships, particularly when phylogenetically closely related organisms are being analyzed. Also, the framework presents a methodology for the discovery of genetic characteristics highly related to phenotypic characteristics in pangenomes. The genomes from the previous manual annotation are divided into groups to identify characteristics unique or more related to a particular group of interest. These characteristics have the potential to explain the different phenotypes among the genomes and may be the key for different kinds of research, such as the identification of biotechnological targets for disease control, the development of vaccines, among others.

The validation of GTACG's functionality is established from the following biological question: is it possible to identify the potential genes that would justify the fact that some bacteria have the ability to survive in association with plants while others do not have such an adaptive characteristic? To answer this complex question, we analyzed a set of 161 genomes from the Xanthomonadaceae family using GTACG. This family is considered for analysis because it comprises genera of strictly phytopathogenic bacteria as well as those with distinct lifestyles not associated with plants. After the processing and presentation of the results, GTACG has proven efficient in answering the established question, allowing the identification of the potential gene families for the molecular studies of the plant–pathogen interaction in pathosystems of agricultural interest. In conclusion, therefore, GTACG can be used to answer similar questions at different levels of complexity, using any set of genomes previously established by users.

MATERIALS AND METHODS

The environment as a whole can be divided into back-end and front-end. The back-end is developed in Java, which is the stage when the preprocessing of the genomic data provided by users occurs. Users provide data such as the complete genome sequence (in the FASTA format and multiple files if necessary), manual annotation of these genomes (in plain text files), and annotation of CDSs (preferably automatic annotation of sequences in the formats FASTA, gb, gbf, and gff). The GTACG execution pipeline is schematically described in **Figure 1** and has three main pillars: 1) identification of homologous genes, 2) comparison of complete genomes, and 3) genome visualization. In order to avoid inconsistencies between the annotations of the different genomes, all the genomes used were automatically reannotated using a RASTtk-based tool available at the PATRIC web service (Wattam et al., 2016).

Identification of Homologous Genes

The first step is to calculate the local alignments of all CDSs against all CDSs using blastp to obtain the alignment length and E-value metrics. Then, a threshold of a minimum size of alignment associated with the degree of separability of the families is set by users. The E-value is automatically chosen to maximize the clustering coefficient of the graph which represented the relationships among CDSs and, therefore, maximizing the transitivity of the homologous correlations (Santiago et al., 2018). The result of this process generates layers of thresholds

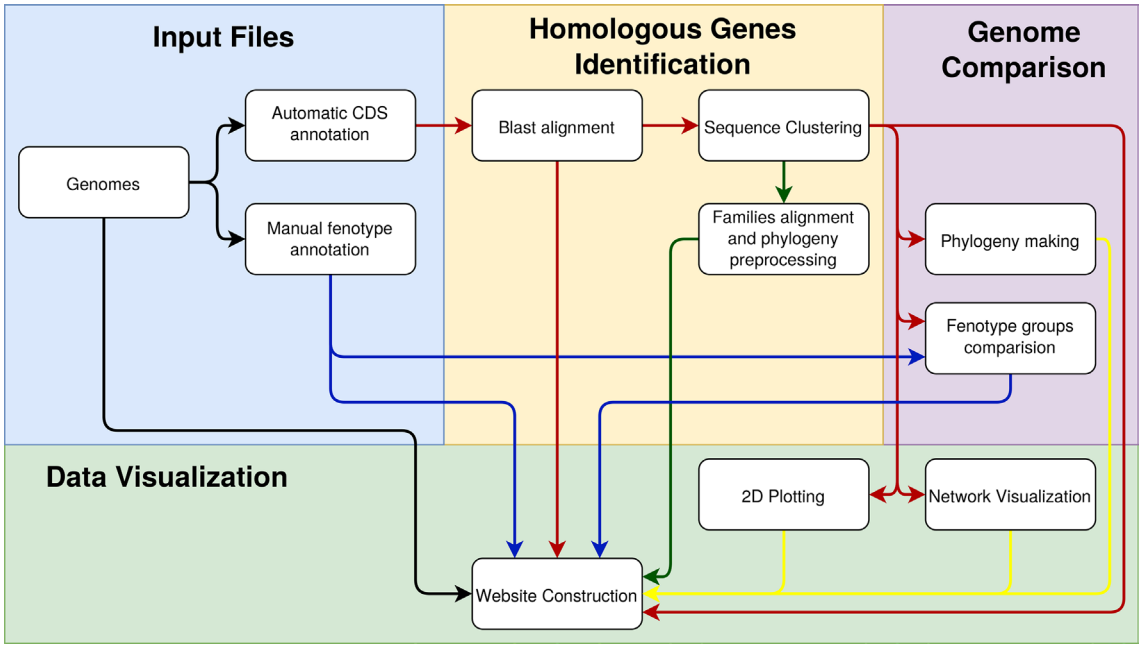


FIGURE 1 | Steps involved in constructing the GTACG pipeline separated into three stages of pre-processing: the identification of homologous genes, the comparison of genomes, and the visualization of data. To facilitate the visualization of the relationships that the data have in each of the activities, the arrows were colored as follows: in black is the general data on genomes; in blue is the data about groups of genomes; in red is the data on the sequences; in yellow are the graphical results for visualization.

that indicates the decisions needed to identify the homologous gene families. These layers allow users to use different levels of trust to build gene families that can be chosen according to the goals of their research.

Also, two other steps were established for the subdivision of homologous CDS families. In the first step, a simple phylogenetic analysis is used, in which branches longer than a certain threshold are excluded, producing the division of a potentially homologous family into two or more orthologous families (Ding et al., 2017). Finally, a search for multidomain proteins is made, taking advantage of the asymmetry in the alignment graphs of each of the previously established families. A family with multidomain proteins, binding two or more CDS groups, is then subdivided into these groups. Unlike homology and orthology, this step resulted in intersecting subdivisions.

For each family, from the three depth levels (homology, orthology, and 102 domains), multiple alignments of the CDSs are done, and the generation of phylogenies is established using Clustal Omega (Sievers et al., 2011) and FastTree (Price et al., 2010), respectively. These data are preprocessed to generate a unified phylogeny, to calculate the metrics related to group phenotypes and for visualization in a graphical environment.

Comparison of Complete Genomes

Using different approaches, three phylogenetic profiles are constructed from the families of homologous CDSs identified in the previous step. The first considers the presence or absence of each genome in the homologous gene families. From these data, a binary vector of characteristics is constructed, in which each characteristic represents a family and assumes the value 1 (one) if the respective genome has one or more CDS in the corresponding family and 0 (zero) otherwise. The junction of all these vectors is then presented to an algorithm for phylogeny inference. The second approach uses a distance matrix for phylogenetic inference constructed by the Euclidean distance between the binary vectors of characteristics. The third approach is based on the concept of supertree (Creevey and McInerney, 2004) and corresponds to a summary of the phylogenetic relationships among several taxa fed by a set of phylogenies. The set of phylogenies chosen is the set of phylogenies of each of the gene families (generated from the alignment of their sequences).

Regarding the investigation of genetic traits based on genome annotations, three categories of characteristics of the families are considered. However, most of the approaches comprised finding characteristics that are common to a certain group of genomes (genomes that share some characteristic of interest set up by users) and simultaneously uncommon to the others. For this investigation, the following categories of characteristics are considered: 1) The conformation of families, defined by families (individually or in combination) unique to a particular group of genomes or families more considerably present in a particular group of genomes. In this way, metrics are presented to indicate how many CDSs are present in the family that belongs to the genomes of a given group. This data is also presented in percentages, indicating how much these CDSs are representative of the total family size and how many genomes of the group are represented by the family. 2) The alignment of the sequences of

the families, identifying specific amino acids variations more common to a certain group of genomes. To express this concept numerically, we developed a metric of dissimilarity that assigns a correlation weight to a given group for each base. 3) The phylogeny of the families, analyzing the grouping or separation of a certain group of genomes in the phylogeny in relation to the others. The Most Isolated SubTree (MIST) metric was developed to express this concept that shows the size of the largest subtree found of the phylogeny that has sequences only related to the group under analysis.

Genome Visualization

Similar to the comparison of genomes, the visualization is also quite dependent on the conformation of the families. The homologous gene identification algorithm utilizes a graph-based algorithm, in which the sequences are represented as nodes and the alignments as edges. Given this data structure, the pangenome is then presented as a gene network, where each homologous family is represented as a connected component, providing a comprehensive notion of the pangenome situation. A force-directed algorithm (Kobourov, 2012) is applied to approximate or separate the sequences according to their edges.

A bidimensional mapping of the genomes is also performed using the same distance matrix constructed from the characteristic vectors described for the phylogeny construction. Using a Multidimensional Scaling algorithm (Borg and Groenen, 2005), the distance matrix is approximated to a bidimensional plane, proportionally preserving the distances in the plane from the distances present in the matrix, resulting in an overview of the proximity/distance between the analyzed genomes.

In this step, the data from all previous steps is consolidated in a static website, so it is unnecessary to use complex server configurations to take advantage of most system functions. This is justified by the fact that the system uses data produced by pre-processing. The website also does not require the installation of a database management system because the data is written as JavaScript scripts. Although the data is related to each other, these relationships are managed internally and not through a database, thus not requiring computational background by users, which makes the GTACG a typical user-friendly tool in genetic analysis.

The website format was chosen due to qualities such as the ease in publishing results, the flexibility to change the environment, and the reusability of the data in other programs or systems. On the other hand, it allows different filters on the data as well as the creation of different data groups, allowing a rich interaction and the visualization or analysis of only the information of interest set by users. Another advantage is the possibility of sharing, through URLs, pages, and search results, which makes the data generated accessible for collaboration between researchers.

Case Studies: Validation of GTACG Functionality

To present the potentialities of this framework, we implemented a pilot study. The case study contains 161 genomes from the *Xanthomonadaceae* family, belonging to the genera *Pseudoxanthomonas* (3), *Stenotrophomonas* (19), *Xanthomonas* (125), and *Xylella* (14) (**Supplementary Table 1**). The choice was made because the first two genera are not associated with plants, while the latter two are strictly phytopathogenic (except one species), thus allowing the re-evaluation of the preliminary results pre-generated by our team (Assis et al., 2017).

RESULTS

Through a single package of compressed files containing source code and shell scripts, users can easily install all the tools to run GTACG on a Linux desktop or server. Once installed, users can load the genomes of interest, and automatically the GTACG will perform an automatic reannotation as a way to standardize the data to be compared.

The searches are flexible to meet users' needs by providing several metrics that can be combined in a variety of ways and shared through URLs. The customization of all visualization data (alignments, phylogenies, and graphs) is also available, which can be exported in ready-to-publish formats such as SVG and high-quality PNG.

The data visualization process has different levels of detail. In the initial screen of GTACG are the more macroscopic data that approach the visualization and interaction with genomes (**Figure 2**). In this screen, users can access the next level of detail regarding family's search using basic settings in the Settings or Filters sections. In the second section, it is possible to define filters on the visualization of families in the results, based on genomes or groups. Families can be filtered on the basis of whether or not they require a particular genome to be present in the listed results, and information related to a particular genome can be ignored. It is also possible to easily find all the families that are shared or not shared by a certain group. In the following section named Statistics, graphs are built through the Google Charts library based on the metrics related to families, sequences, and local alignments. Finally, sections 2D Plot and Phylogeny present the chosen methods for visualization of genomes. Moreover, these two sections can be customized based on the groups of annotated genomes, in addition to several additional configurations. The phylogenies presented in GTACG use the Phylocanvas library for visualization.

The next level of detail concerns families. At this level, families can be found through statistical data, the sequences that compose them, and their base pairs respectively, available through buttons in the Settings section on the home screen. Families' statistical data contain metrics such as the number of genomes shared by a family, the number of sequences, sequence length distribution, annotated function, the metrics discussed above for groups of genomes, and others based on the graphs constructed for the identification of families, distribution of amino acids in the alignment, and data on phylogeny. The statistical data refers to the degree of subdivision chosen for the families (homology, orthology, and domains, previously discussed in the *Materials and Methods* section), which can be changed in the initial screen of the system. These data are also available for download in formats that can be used to construct phylogenies (a distance

FIGURE 2 | Continued

matrix, for example) or in the Roary output format (Page et al., 2015) making use of a wide range of functions for the analysis and visualization of data already developed. In the sequence data, families are found according to the metrics present in each of the sequences that compose them, such as their annotated function, length, or position in the genome. In case there is a minimum server configuration (the execution of a script written in Node.js), it is possible to find families by Blast search against all sequences of the pangenome, with filters and results that are already the characteristics of this tool. These approaches have been structured as dynamic tables built with the Tabulator library, so users have at their disposal dynamic and complex filters adapted to work with mathematical and logical expressions as well as data grouping functions.

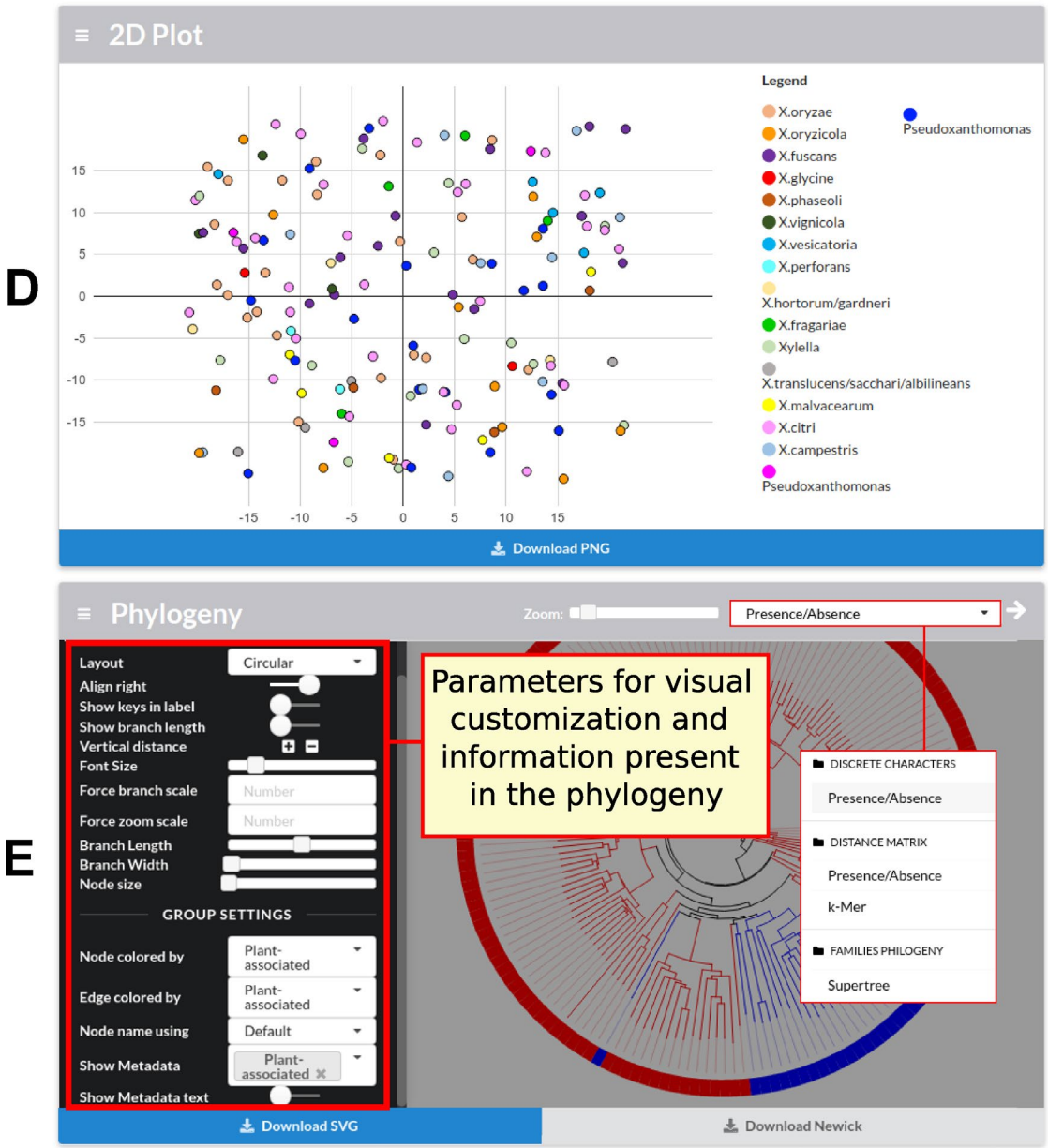


FIGURE 2 | GTACG home screen. These results are divided into five sections: Settings, Filters, Statistics, 2D Plot, and Phylogeny. The first two sections are related to the subsequent family's searches; the others are related to genome data. **(A)** The first allows the navigation between the different levels of clustering (homology, orthology, and domains). **(B)** The second allows filtering the presence/absence of the genomes or according to groups of genomes; this section also shows the number of genomes which are being filtered (label 1 in the figure) and the number of families after applying the filters (label 2 in the figure). **(C)** The third, Statistics, presents the graphs for the metrics related to families, sequences, and local alignments. **(D)** The fourth, 2D Plot, presents a bidimensional projection of the genomes. **(E)** Finally, Phylogeny presents the built phylogenies and customization options. Most sections fit users' screen size.

The last and lowest level of detail pertains to families. At this level, each family has its own page with its respective data (**Figure 3**). These pages have a total of five sections. In the first section, sequence data (annotation, length, among others) are combined with genome data (genome identification and annotated groups). Also, for each sequence, a link to the NCBI website to perform a Blast search is present. In case the server (a script written in Node.js) is configured, it is also possible to visualize the desired sequence and its synteny in the genome, due to the igv.js library. In the next two sections are phylogeny and sequence alignment respectively, using the Phylocanvas and MSAViewer (Yachdav et al., 2016) tools, and even when results are already pre-processed in the back-end, new results can be processed using FastTree (Price et al., 2010), PhyML (Guindon et al., 2010), RaxML (Stamatakis, 2014), Clustal Omega (Sievers et al., 2011) and MUSCLE (Edgar, 2004). The fourth section is

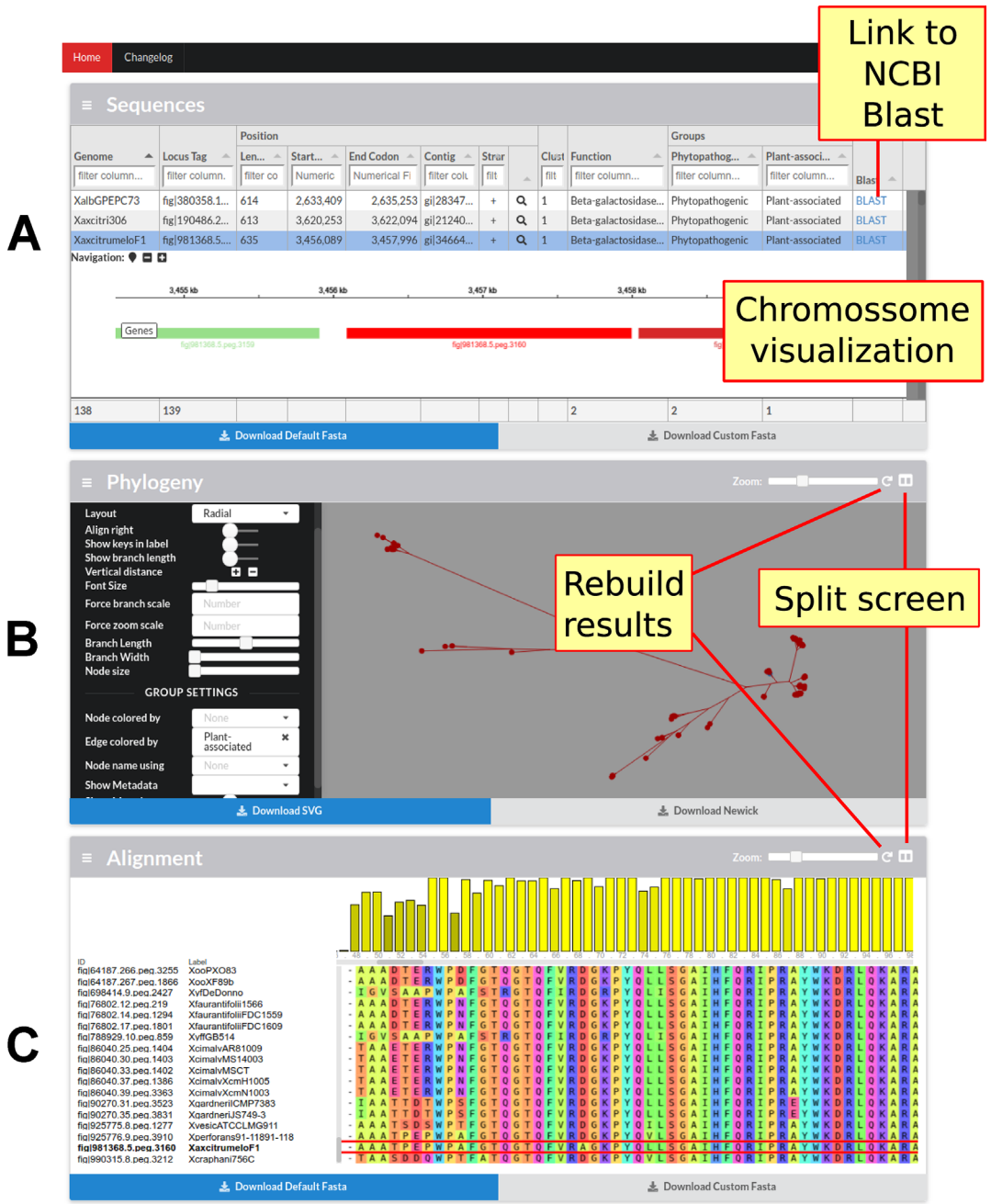


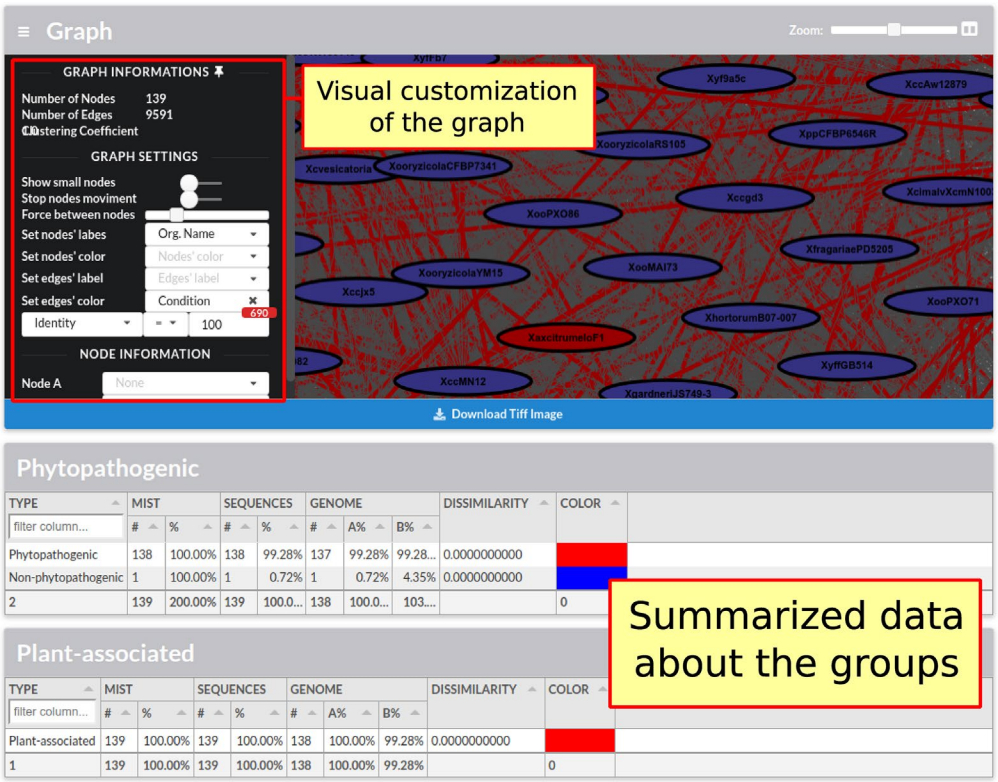
FIGURE 3 | Continued

devoted to the graph that generates the family, in the process of identifying families, representing the sequences as vertices and local alignments as edges. All this data is available for viewing and can be used to highlight edges by defining a condition, for example, highlighting the local alignments where the identity is less than 80%. Finally, the last section presents a statistical summary of the genome groups limited to family data.

Owing to all these possibilities, users are able to structure a research based on a top-down approach, first trimming with genomic data (such as phenotype annotation, phylogenetic data or exclusive genes statistics, for example) and then delving deeper to the point of better understanding the genetic mechanisms that can justify the initial data. The reverse is also possible, as users can find the orthologous family by having the amino acid sequence.

The Case Study Validated by GTACG

The 161 genomes from the Xanthomonadaceae family employed in this study ranged in size from 2.5 to 5.5 million base pairs, with an average of 4,480 CDSs. The 743,920 CDSs were grouped



TYPE	MIST		SEQUENCES		GENOME			DISSIMILARITY	COLOR
filter column...	#	%	#	%	#	A%	B%		
Phytopathogenic	138	100.00%	138	99.28%	137	99.28%	99.28...	0.0000000000	
Non-phytopathogenic	1	100.00%	1	0.72%	1	0.72%	4.35%	0.0000000000	
2	139	200.00%	139	100.0...	138	100.0...	103....		0

TYPE	MIST		SEQUENCES		GENOME			DISSIMILARITY	COLOR
filter column...	#	%	#	%	#	A%	B%		
Plant-associated	139	100.00%	139	100.00%	138	100.00%	99.28%	0.0000000000	
1	139	100.00%	139	100.00%	138	100.00%	99.28%		0

FIGURE 3 | Screen containing the results of only one family. This screen contains four main sections followed by sections summarizing the family data in relation to the groups of genomes presented in the framework. **(A)** The first section has the sequence data and the data of their respective genomes; it is also possible to graphically visualize the position of each sequence in the genome as well as its vicinity. **(B)** In the next section, Phylogeny, it is possible to visualize, customize, and reconstruct (with different parameters) the phylogeny of the sequences. **(C)** The following section shows the alignment of all the sequences; it is possible to view, customize, and rebuild the alignment. **(D)** The fourth section presents the graph constructed to identify families, in which sequences are represented as vertices and local alignments as edges. The graph can be customized to highlight the alignments in accordance with some specific metrics. In this figure, the local alignments with identity equal to 100% are highlighted. **(E)** Finally, the last section summarizes the statistical data from each group of genomes with metrics about the number of genomes, dissimilarity, and MIST.".

into 48,477 homologous families, of which 4,287 were subdivided into 13,528 orthologous families, resulting in a total of 57,718 orthological families. This number of orthological families can be considered acceptable for this large and complex set of genomes. To obtain these results, two parameters were specified: 1) a maximum E-value threshold of 10-10 and 2) a minimum size of 45% for the alignments.

The main phenotype of interest evaluated in the proposal of GTACG validation is associated with the fact that some microorganisms from specific genera within the *Xanthomonadaceae* family have an adaptive association with plants, either as phytopathogens or not. It is important to emphasize that this characteristic was not mandatory for all the genomes investigated. This is justified by the fact that with this phenotypic characteristic, 139 genomes belonging to the genera *Xanthomonas* and *Xylella* and without this characteristic, 22 genomes belonging to the genera *Pseudoxanthomonas* and *Stenotrophomonas* were previously selected.

As can be seen in **Figure 4**, the sets of associated and not associated with plants genomes are well separated from each other, which is reiterated in the literature (Sharma and Patil, 2011). In the tree constructed based on the binary vectors (**Figure 4A**) and in the tree constructed based on the distance matrix (**Figure 4B**), it is possible to clearly observe the separation of non-plant-associated microorganisms. Two exceptions can be observed in both trees, the clustering of *P. spadix* BD-a59 to plant-associated genomes and the clustering of *X. mangiferaeindicae* genomes into the cluster of non-plant-associated genomes. Moreover, the supertree (**Figure 4C**) presented a clustering with a more recent hypothetical ancestor for the non-plant-associated group, thus excluding *Xylella* (in discordance with the two phylogenies discussed above). This result corroborates with that of other studies that show that *Stenotrophomonas* is phylogenetically closer to *X. campestris* than to *Xylella* (Ramos et al., 2011; Naushad and Gupta, 2013).

No orthologous family presented the ideal behavior of being present in all genomes associated with plants and absent in all others. Nevertheless, very interesting results have been found that are consistent with the phylogenies constructed. It was found that 19 families of genes identified in 90% of the genomes associated with plants but were absent in genomes not associated with plants. Interestingly, the absent genomes are the same ones that were identified as separate groups in the phylogeny. In none of these 19 families, *X. mangiferaeindicae* is present. In three

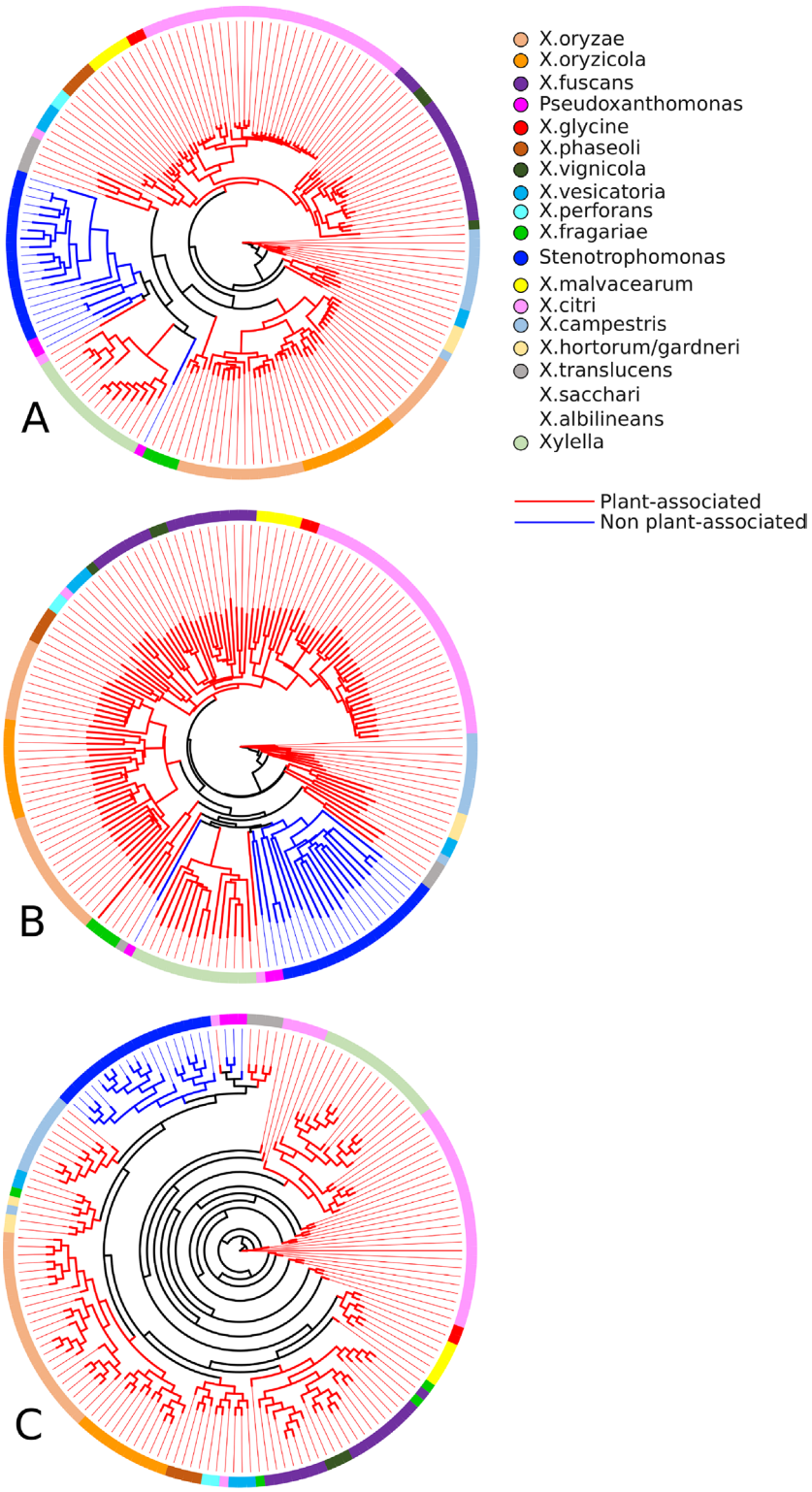


FIGURE 4 | Phylogenetic profiles established by GTACG from the input genomes. The phylogeny **(A)** was inferred using the binary vectors for each genome; the positions of the vector represent the families and are defined as 0 or 1, depending on the presence/absence of the genome in the respective family; the method of inference was the parsimony program (pars) for binary features in the Phylip package. The phylogeny **(B)** was constructed using the distance matrix (using the Euclidian distance) of the binary vectors referred to above; the inference method chosen was the neighbor-joining also available in the Phylip package. The phylogeny **(C)** was constructed using a supertree that summarizes the collection of all the phylogeny constructed for the families; the tree of each family was obtained using the Clustal Omega to make the alignments and after that the FastTree produce the trees; the supertree method was the Quartet Fit algorithm with Nearest Neighbour Interchange available in the Clann.

families, *X. albilineans* is also not present, and in two families, two strains of *X. translucens* and *X. sacchari* are also not present.

In another search, we also found nine families shared by all genomes associated with plants and less than 30% of the non-plant-associated genomes. Similarly, it should be noted that a few genomes not associated with plants have been integrated into this group and respective analysis. Interestingly, regarding these nine families, the number of non-plant-associated genomes that were included were very small (between three and six genomes). This result was partly expected, given the result presented by the supertree, as *P. spadix* BD-a59, *P. suwonensis* 11-1, and *P. suwonensis* J1 (genomes present in these families) were grouped in a branch with plant-associated genomes.

Also, nine protein families that compose the core genome have dissimilarity greater than 1% in their alignments, indicating amino acids with mutations more correlated to the genomes associated with plants. Finally, another 13 families from the core genome were isolated in a single branch of the phylogeny containing all sequences from microorganisms associated with plants.

DISCUSSION

Pangenome Analysis Tools

The analysis of pangenome date back more than a decade (Vernikos et al., 2015). Several published works and computational tools are available, some of which using a similar approach presented in GTACG to study the genomes based on the clusters of homologous families (or orthologous). However, most of these works and tools are limited to global numerical analyses such as finding the different categorizations of the core genome or counting the number of unique genes in the analyzed genomes (Laing et al., 2010; Zhao et al., 2011; Benedict et al., 2014; Page et al., 2015; Zhao et al., 2018). Another common approach of these tools is the search for a reliable phylogeny from the raw input data, with the possibility of generating a rapid alignment of the genomes and not limited to the low resolution of some phylogenetic markers (Clarridge, 2004).

However, families of sequences or homologous genes have a wide range of information to be mined. It is in this context of a more refined search for information that the number of works and tools available still have limitations. Some of them, although discussing similar problems, use manual methods, which de-characterize them as potential user-friendly tools in systems biology.

Regarding the automatic methods already developed for the analysis of pangenome and homologous/orthologous genes or sequences search (some of them listed in **Table 1**), the PGAT (Brittnacher et al., 2011), the PanX (Ding et al., 2017), and the Obolski (Obolski et al., 2018) stand out. Although the PGAT provides a wide range of possibilities for gene searches with specific interests, it is limited, as it allows such search to be established only by a particular set of genomes. Moreover, one of the main limitations of the PGAT lies in the rigidity of not allowing approximate results to be found, a limitation also shared by BPGA (Chaudhari et al., 2016) that presents searches for phenotypic characteristics, but with inflexible search formats. For example, if any phenotype has not been correctly annotated

TABLE 1 | Comparison of the main functionalities of some comparative genomics frameworks.

	GTACG	BPGA	PanX	PGAT	PanGP	PGAP	Panseq	ITEP	Get Homologues
Identification of phenotype-specific genes – list	X	X		X					X
Identification of phenotype-specific genes – metrics	X								
Distribution of core, accessory and unique genes	X	X	X						
Pangenome profile analysis	X	X			X	X			X
Size of core and pan-genome	X	X	X			X	X		X
Extraction of core, accessory and unique genes' sequence	X	X						X	
Evolutionary analysis	X	X	X			X	X	X	X
Protein/gene clustering	X	X	X	X		X	X	X	X
Multilevel perspective of the genes	X		X	X				X	
Input data from user	X	X			X	X	X		X
Easy to share results	X			X			X		
Integration with roary scripts	X								
Data preparation	C	C	C	N/A	G	C	C	C	C
User interface	W	GO	W	W	GO	GO	GO	GO	GO
References		Chaudhari et al. (2016)	Ding et al. (2017)	Brittnacher et al. (2011)	Zhao et al. (2014)	Zhao et al. (2011)	Laing et al. (2010)	Benedict et al. (2014)	Contreras-Moreira and Vinuesa (2013)

Data preparation: C, Command line; G, Graphical interface.
User interface: W, Website; GO, Graphical output.

(or expressed) by users, it will not be easily found, thus requiring many consecutive searches to solve the problem. Although the PGAT is able to present the results as a website, the specificities of the results (such as the result of a search) are not easily shared. PanX also presents the results in a website but more dynamically than PGAT. However, the search options are still limited to the basic statistical data on families such as the number of genomes present, and therefore there is a possibility of searches that support the study on phenotypes. An interesting advantage of the PanX is the visualization of family's phylogenetic trees using metadata such as phenotypes from genomes as visual support. Finally, Obolski uses a Random Forest algorithm to find the families most correlated with the invasiveness phenotype, as presented by some strains of *Streptococcus pneumoniae*.

PanSeq Laing et al. (2010), as well as PanX and PGAT, also make the results easily available (*via* URLs), but as a service which provides only files with specific results, without customization and any interaction with the user. In general, the rest of the available frameworks are quite focused on an experience restricted to text commands, such as ITEP or get_homologues, or limited interactive interfaces, such as PGAP (Zhao et al., 2011) that has been recently extended with graphical interfaces (Zhao et al., 2018).

Based on the description of the qualities and limitations of the tools mentioned above, GTACG is able to combine the main advantages of all of them, besides having its own algorithm for the identification of homologous gene families with different levels of grouping, which minimizes some of the limitations imposed by other tools. Also, GTACG stands out by facilitating data presentation and the sharing of search results, a feature that is highly desirable in a user-friendly tool for systems biology. Although it does not cover all the diversity of software that address pangenome, owing to the existence of an open and easily modifiable environment, GTACG requires less effort to program new content, thus reducing the difficulties imposed by some tools aimed at the study of systems biology (Hillmer, 2015).

The GTACG: Structural and Functional Characteristics

Some demands and difficulties imposed by the tools developed for studying systems biology guided the development of GTACG. GTACG was developed in consideration of the following:

Easy to Load the Information to be Analyzed

As it is aimed at the interdisciplinary public, the results were produced from files commonly used in genomic projects (for example, fasta, gb, and gff), easily obtained through NCBI and automatic annotation tools, and the interaction of the results with users occurs through a graphical environment. This allows users to load an unlimited number of genomes.

Minimizes the Propagation of Annotation Errors

Perhaps the most critical decision in a project on pangenomes concerns the formation of families of homologous sequences, especially if the problem is aggravated in situations where the sequence was annotated incorrectly (Devos and Valencia, 2001; Green and Karp, 2005). This leads to error propagation, and it is deterministic in the characterization of gene families incorrectly identified as homologous, thus creating false positive or false negative errors that are difficult to be identified. Therefore, the first step of GTCAG was established to standardize the CDSs' annotation through an automatic annotation, as many genomes present in the NCBI database were submitted using different methodologies and at different times (Klimke et al., 2011).

Accuracy in the Clustering Method

Once the annotations have been standardized, another parameter crucial for the quality of the tool is the identification of the gene families, which many other studies have chosen to use—Markov Cluster Algorithm (MCL) and its derivatives (Enright et al., 2002; Li et al., 2003). However, this is a general-purpose clustering method. In this work, GTACG was chosen because it was developed with the implementation of the Multilayer Clustering, which is a more stringent parameter to be used in sequences from phylogenetically closer genomes. Also, this algorithm uses global decisions, considering the influence of all sequences on the formation of families, as the relationships between the sequences in pangenome studies are much more homogeneous than more diverse sequences.

Accuracy in the Search for Families of Sequences or Homologous Genes

The identification of homologous genes is a critical step. It impacts all obtained results such as phylogeny, searches for families, genome visualization, among others. To deal with this task, GTACG uses Multilayer Clustering (Santiago et al., 2018) instead of TribeMCL or OrthoMCL, which are more commonly used among known solutions. A detailed comparison of Multilayer Clustering and TribeMCL results considering a subset of 69 genomes from the 161 of the case study can be found in Santiago et al. (2018). These algorithms achieved comparable results when multidomain proteins are not considered. But, considering multidomain proteins, Multilayer Clustering achieved better results. Moreover, the impact of the decisions made by Multilayer Clustering is easier to understand, as the basic knowledge about alignment tools is enough to understand clustering decisions. It is opposite to MCL, which does not provide a transparent picture to users concerning what decisions impact homologous identification (Santiago et al., 2018).

Dynamic and Easy-to-Use Graphic Interface

All the interface was developed together and intended for biologists. Acknowledging the interdisciplinary public, some concerns were considered. The first concern was to create an environment that do not need complex server configurations, allowing computer non-specialist users to publish their results. The second and more important concern was to develop a dynamic system and an easy-to-use interface. The interface was modeled as a website using common internet symbols and icons to facilitate user learning. The pages were divided into genomic

information (and visualization), family pre-processed metrics, and individual family information, designed as a top-down approach. Finally, the last concern was to create customizable graphics to allow users to express their ideas better. Moreover, the graphics could be exported to ready-to-publish formats (SVG, high-quality PNG, and TIFF).

Support for a Lifecycle Research Project

Considering all the features mentioned above, GTACG presents the qualities to support the work of researchers in different steps of the lifecycle of a research project. In the first step, GTACG supports researchers to obtain genomes directly from the NCBI database and, in a row, automatically reannotate them. Also, the methodology of the identification of homologous genes is covered, providing comprehensive results of clustering through the Multilayer Clustering. In the analysis step, GTACG allows researchers to test plenty of hypotheses and find data that can conduce to new hypotheses, collaboratively through URLs. Finally, the same environment of analysis serves to turn the data public and generate graphics with enough quality to support scientific publication. Thus, GTACG is able to support the full lifecycle of pangenome research without requiring computing knowledge.

Performance of the Pipeline Execution

GTACG presents fairly complete results covering different stages of pangenome research. In general, this process starts after the reannotation of the sequences and the production of local alignments, these steps are the most computationally costly.

The total time of the automatic annotation, as well as the quality and specificity of its results, is quite dependent on the choice of the tool used. This step is quite costly and some tools require a manual effort from the researchers. However, it is an inevitable step to minimize methodological errors in many pipelines of tools based on homologous gene identification.

In order to evaluate the performance of the subsequent steps, five datasets were prepared with 10, 20, 30, 40 and 50 *Xanthomonas* genomes. These genomes are presented in the **Supplementary Table 1**, and the execution times are present in the **Supplementary Table 2**.

The step of producing local alignments of all sequences against all sequences was performed using BLAST (blastp), and is currently the most costly part of the whole process, consuming between 75% and 95% of the execution time for these datasets (**Figure 5** shows the result using 20 threads). Although this result can be accelerated through multithreading, the tendency of this consumption is exponential, as in the case presented

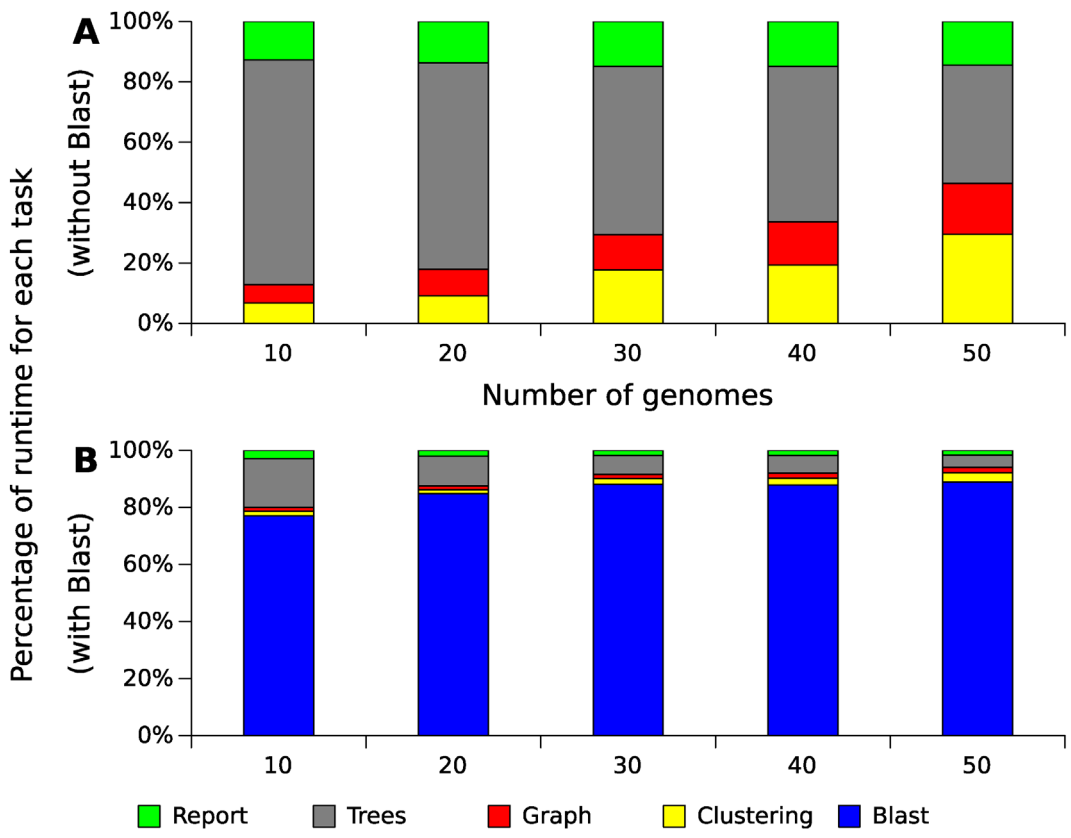


FIGURE 5 | Relative runtime for GTACG's main tasks with different datasets of Xanthomonas genomes. These results were obtained using a computer with an Intel(R) Xeon(R) CPU E5-2620. This computer has 24 cores, but only 20 of them were used. As Blast's alignments correspond to the majority of the consumed time, section **(A)** present the time spent excluding the time spent with Blast, while section **(B)** present the time including Blast.

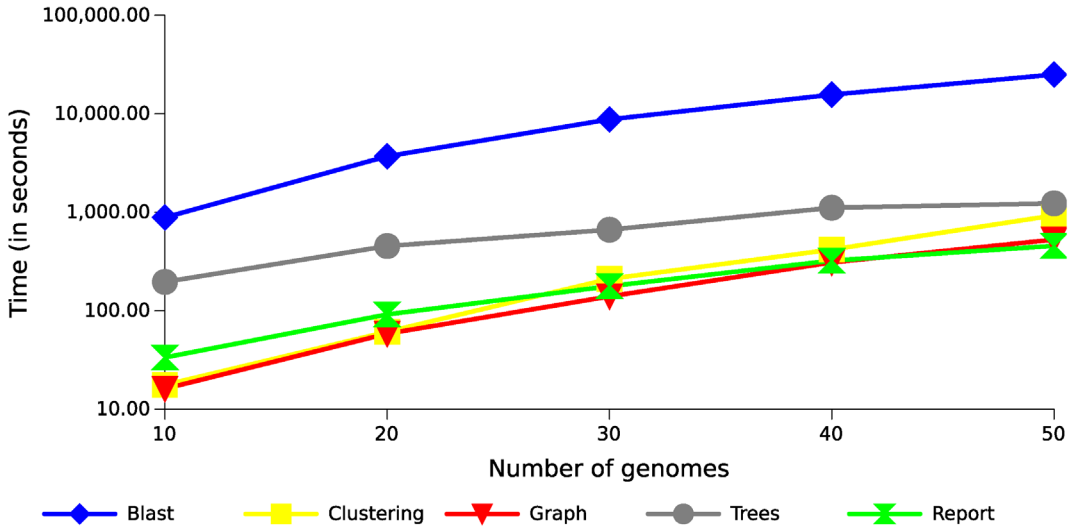


FIGURE 6 | Runtime with different datasets of *Xanthomonas* genomes. The results show the execution time growth of the GTACG's main tasks, according to the number of genomes in the datasets.

in this figure, because the number of alignments produced increase exponentially with the increase of genomes. The remaining operations also tend to be exponential following the growth of the alignments (**Figure 6**). The most costly task after the alignments is the preparation of the multiple alignments and trees for each of the families, but this step follows a more linear trend.

A very promising alternative to the use of Blast is the MMseqs2 (Steinegger and Söding, 2017) with a sensitivity of 7.5, which considerably reduced the local alignment execution time (between 30 and 35 times), while maintaining similar results both in the tests datasets and in the case study discussed below.

Although GTACG takes longer to compute than other frameworks, such as Roary (Page et al., 2015), BPGA (Chaudhari et al., 2016) or PanGP (Zhao et al., 2014), GTACG provides more information for the users, different results and more tools to help the pangenome analysis in a simple and practical way for users with no programming skills.

The Case Study Validated by GTACG

Considering the case study of 161 genomes from the Xanthomonadaceae family, all searches were done simply and efficiently, making the discovery of knowledge about phenotypes easier. Although these results are not sufficient to determine whether there is, in fact, the participation of which one of these families to express the phenotype, it is a starting point that can guide new laboratory studies.

The same behavior observed in the phylogeny is reflected in the composition of families (**Figure 7**). Even though the two groups (plant-associated and non-plant-associated) are well divided, there are branches involving few genomes in which the groups are mixed. There are 19 families unique to plant-associated genomes, and plant-associated genomes are present in at least 90% of them. *X. mangiferaeindicae* does not have genes in any of these families, and among 15 of them, it is the only one absent among plant-associated genomes. Of the four remaining families, one does not contain only *X. albilineans*, a microorganism vastly studied for being unique within this family and probably resulting from a process of genome reduction (Pieretti et al., 2009). In two other families, the same genomes grouped with non-plant-associated genomes, as described by the supertree, are absent. Considering these 19 families, most of them may be important for the metabolic interaction with plants, and therefore, *X. mangiferaeindicae* would have adapted to use an alternative strategy as well as *X. albilineans* could have adapted to using a reduced set of genes from these families. Finally, among this set of families, one of them do not contain any of the four strains of *X. fragariae* (besides *X. mangiferaeindicae*).

On the other hand, considering the families that comprise all the plant-associated genomes (but not exclusively them), there is a family that contains the same three non-plant-associated genomes grouped with the plant associated with the method of the supertree: *P. suwonensis* 11-1, *P. suwonensis* strain J1, and *P. spadix* BD-a59. Also, eight families contain, additionally to plant-associated genomes, genes from *S. nitritireducens*, *Stenotrophomonas* sp. KCTC 12332, and *S. acidaminiphila*. This can be explained by the hypothesis that perhaps the cited families are important to allow the association with plants, but some genomes potentially cannot express these genes and therefore would not express the phenotype either or the possibility that the genomes themselves were erroneously annotated.

Based on the alignments produced by the families, nine cases were found presenting amino acids with specific mutations in the plant-associated genomes with dissimilarity greater than 1%. The data below that indicates a 1% threshold does not yield very conclusive results, showing many non-exclusive mutations.

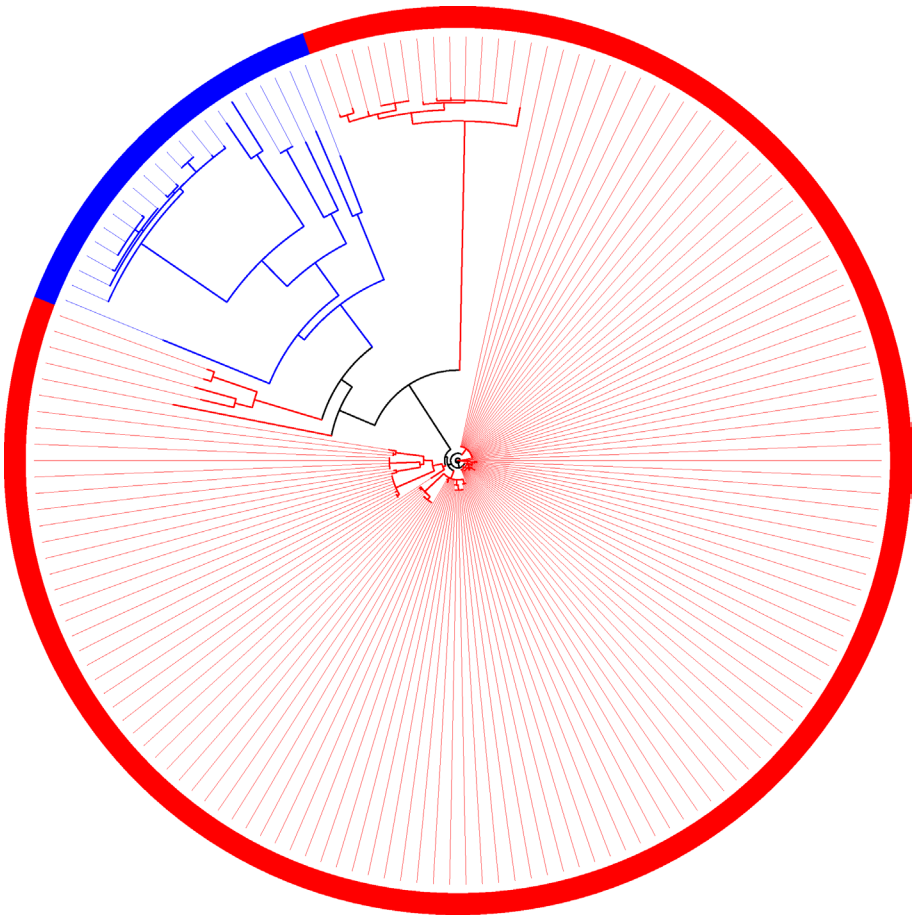


FIGURE 7 | Family of orthologous sequences in which all sequences from plant-associated genomes are isolated from the other genomes.

Besides, from the phylogenies constructed based on the alignments, it was found that 19 families can be perfectly divided into both groups, as shown in **Figure 7**. By itself, this result does not imply that this is the most appropriate phylogeny to represent the evolution of the genomes, but as the phylogeny is an analysis derived from the combination of amino acids, this result indicates a significant difference observed by that amino-acid combination.

Functional Description of Protein Families Found Exclusively in Plant-Associated Genomes

Among the 19 protein identified in at least 134 phytopathogen genomes in this study, eight protein families are involved in N-glycan degradation. Interestingly, all genes related to N-glycan degradation are located in the same genomic region constituting a cluster (nix) together with cutC (resistance to copper) and are responsible for the cleavage of the N-glycan in different glycosidic bounds (**Table 2** and **Figure 8**). Plant-pathogen interaction is driven by evolution of bacterial virulence proteins to induce virulence and modulate plant immune response alongside with evolution of plant proteins to recognize bacterial effectors and induce specialized immune response leading to resistance. Plant pattern-recognition receptors (PRR) are responsible for recognition of pathogen-associated molecular pathogens (PAMP) and activation of pathogen-triggered immunity (PTI). Häweker et al. (2010) showed that PRR require N-glycosylation to mediate plant immunity. By degrading the N-glycan associated with plant-receptors, the plant host is no longer able to recognize and activate the immune response, thus allowing greater success of colonization and adaptation of these bacteria within the host.

Additionally, other proteins identified are involved in adaptation, including two peptidases [homologous to XAC0609 (Zhou et al., 2017) and PepQ-XAC2545] and three hypothetical proteins (homologous to XAC2544, XAC4076 and XAC4164) (**Table 2**). Analysis of the sequence of XAC0501 revealed that this gene coded by LesA/LipA is a key virulence factor required for *Xylella fastidiosa* pathogenesis in Grapevines (Nascimento et al., 2016), *Xanthomonas citri* in citrus (Assis et al., 2017) and *Xanthomonas oryzae* in rice (Aparna et al., 2009). The other four genes may be related with adaptation. HspA has been described as a chaperone very important as a protective agent during the storage of proteins in *Xanthomonas campestris* (Lin et al., 2010).

TABLE 2 | Characterization of the 18 protein families exclusively identified in genomes of bacteria associated with plants.

Function	Gene name	Ref. Locus Tag	# Genomes	# Paralogs	Pathway	SP	Refs
Conserved hypothetical protein (putative lipase)	*lesA (lipA)*	XAC0501	134	27	Lipid metabolism	N	Aparna et al. (2009); Nascimento et al. (2016); Assis et al. (2017)
Peptidase M16 family/Zinc protease/ Insulinase family protein	—	XAC0609	138	1	Peptidases	Y	Zhou et al., 2017
Low molecular weight heat shock protein/ Molecular chaperone	*hspA*	XAC1151	138	1	Chaperones and folding catalysis	N	Lin et al. (2010)
Cytochrome O ubiquinol oxidase subunit IV	*cyoD*	XAC1261	138	2	Oxidative phosphoryla-tion	N	Lunak and Noel (2015)
Conserved hypothetical protein	—	XAC2544	137	2	Unknown function	Y	—
Predicted 4-hydroxyproline dipeptidase/Xaa-Pro aminopeptidase	*pepQ*	XAC2545	138	1	Metallo peptidases	N	—
Alpha-L-fucosidase	*nixE*	XAC3072	138	1	N-glycan metabolism	Y	Boulanger et al. (2014); Dupoiron et al. (2015); Assis et al. (2017)
Hypothetical protein (putative glycosyl-hydrolase)	*nixF*	XAC3073	138	1	N-glycan metabolism	Y	Boulanger et al. (2014); Dupoiron et al. (2015); Assis et al. (2017)
Beta-hexosaminidase/ Beta-N-acetylglucosaminidase	*nixG*	XAC3074	138	1	N-glycan metabolism	Y	Boulanger et al. (2014); Dupoiron et al. (2015)
Beta-mannosidase	*nixH*	XAC3075	138	3	N-glycan metabolism	Y	Boulanger et al. (2014); Dupoiron et al. (2015)
Beta-glucosidase-related glycosidases/ Gluca-beta-glucosidase	*nixI*	XAC3076	138	2	N-glycan metabolism	Y	Boulanger et al. (2014); Dupoiron et al. (2015); Assis et al. (2017)
Hypothetical protein (putative glycosyl-hydrolase)	*nixJ*	XAC3082	138	4	N-glycan metabolism	Y	Boulanger et al. (2014); Dupoiron et al. (2015)
Alpha-1,2-mannosidase	*nixK*	XAC3083	138	1	N-glycan metabolism	N	Boulanger et al. (2014); Dupoiron et al. (2015)
Beta-galactosidase	*nixL*	XAC3084	138	1	N-glycan metabolism	N	Boulanger et al. (2014); Dupoiron et al. (2015); Assis et al. (2017)
Cytoplasmic copper homeostasis protein CutC	*cutC*	XAC3091	138	2	Copper metabolism	N	—
3-isopropylmalate dehydrogenase/Isocitrate dehydrogenase	*leuB*	XAC3456	134	1	Leucine biosynthesis	N	Laia et al. (2009); Moreira et al. (2017)
Integral membrane protein	—	XAC4076	134	1	Unknown function	N	—
N-acetylglucosamine-regulated/TonB-dependent receptor	*nixD*	XAC4131/3071	138	10	TonB receptors/ N-glycan metabolism	Y	Blanvillain et al. (2007)
Conserved hypothetical protein	—	XAC4164	137	1	Unknown function	Y	Jalan (2012)

SP, signal peptide; Y, yes; N, no.

CyoD coded by a cytochrome O ubiquinol oxidase subunit IV that is a component of the aerobic respiratory chain that predominates when cells are grown at high aeration (Lunak and Noel, 2015). LeuB coded by a 3-isopropylmalate dehydrogenase that was upregulated in *Xanthomonas axonopodis* pv. *citri* (Xac) 1, 3 and 5 days after inoculation (Moreira et al., 2017), and when mutated the absence of leuB showed reduction of Xac virulence in the compatible host (Laia et al., 2009). Only homologous to XAC4076 coded by an integral membrane protein was not investigated in other studies.

Finally, the last protein family unique to plant-associated genome is coded by a TonB-dependent receptor (TBDR) homologous to XAC4131. Blanvillain et al. (2007) predicted 72 TBDR in *Xanthomonas campestris*, several of them belong to carbohydrate-utilization loci involved in the utilization of various plant carbohydrates such as sucrose, plant cell wall compounds and pectin, a major cell wall polymer in plants. Thus, the bacteria may also use the byproducts as energy source by internalizing the monomers through TBDR, an outer membrane protein mainly known for active transport of molecules. Curiously, 10 paralogous of this gene was found at investigated genomes (**Table 2**). One of this paralogous is coded by the gene XAC3071 in Xac306 genome, that corresponds to nixD, the first gene of the nix cluster previously described (**Figure 9A**). It is possible that this TBDR gene are involved with internalization of sugars derivative of N-glycan degradation, which could be used as an alternative source of carbon after suppression of the plant immunity.

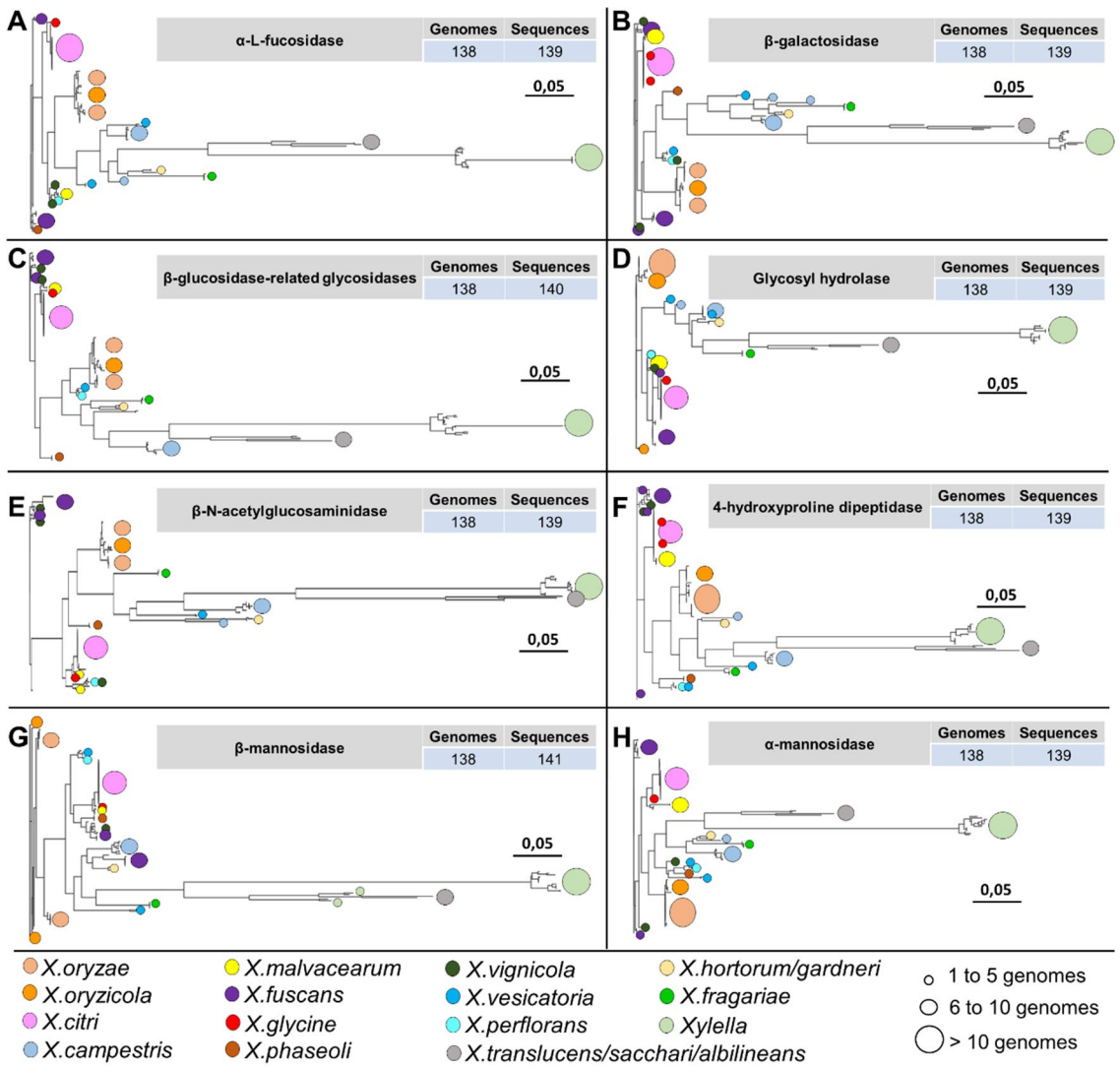


FIGURE 8 | Phylogenetic analysis of 8 out of 19 protein families identified only among the genomes associated with the plants belonging to the family *Xanthomonadaceae*. The identification of circles, colors, and sizes is not provided by the tool; they have been inserted in this context only to facilitate the description of the identifiers. It is possible to observe a pattern in the topology of the phylogenies of the hydrolases, always with larger branches for organisms of the genus *Xylella* and *Xanthomonas translucens*, *X. sacchari*, and *X. albilineans*. **(A)** alpha-L-fucosidase family. **(B)** beta-galactosidase family. **(C)** beta-glucosidase-related glycosidases family. **(D)** glycosyl hydrolase family. **(E)** beta-N-acetylglucosaminidase family. **(F)** 4-hydroxyproline dipeptidase family. **(G)** beta-mannosidase family. **(H)** alpha-mannosidase family.

This analysis of the repertoire of genes investigated allows us to infer that GTACG tool proved to be efficient in the search for a set of genetic information correlated with a phenotype of interest since the genes identified as unique to plant-associated genomes have already been described as capable of modulating bacterial adaptation to the host plant.

CONCLUSIONS

GTACG is a framework to support the research on bacterial genomes in the area of systems biology, especially the research related to the discovery of genetic knowledge pertaining to the expression of phenotypes. The searches are mainly done using the metrics for the study of pangenomes, such as the number of genomes present in a particular family, but metrics have also been used and developed to express the correlation of families with groups of genomes. GTACG structures information by a top-down model, in which the genomic data and global statistics are first presented to users, followed by the search for families of interest, and then the analysis each family in detail. GTACG encompasses the functionalities already present in some other frameworks on pangenomes, such as the automatic identification of families, identification of core/accessory genome, construction of phylogeny, and visualization of data. However, this framework presents its results in the form of a static website, which makes it easier for users lacking computational knowledge to publish their results and share searches in a simple and efficient way.

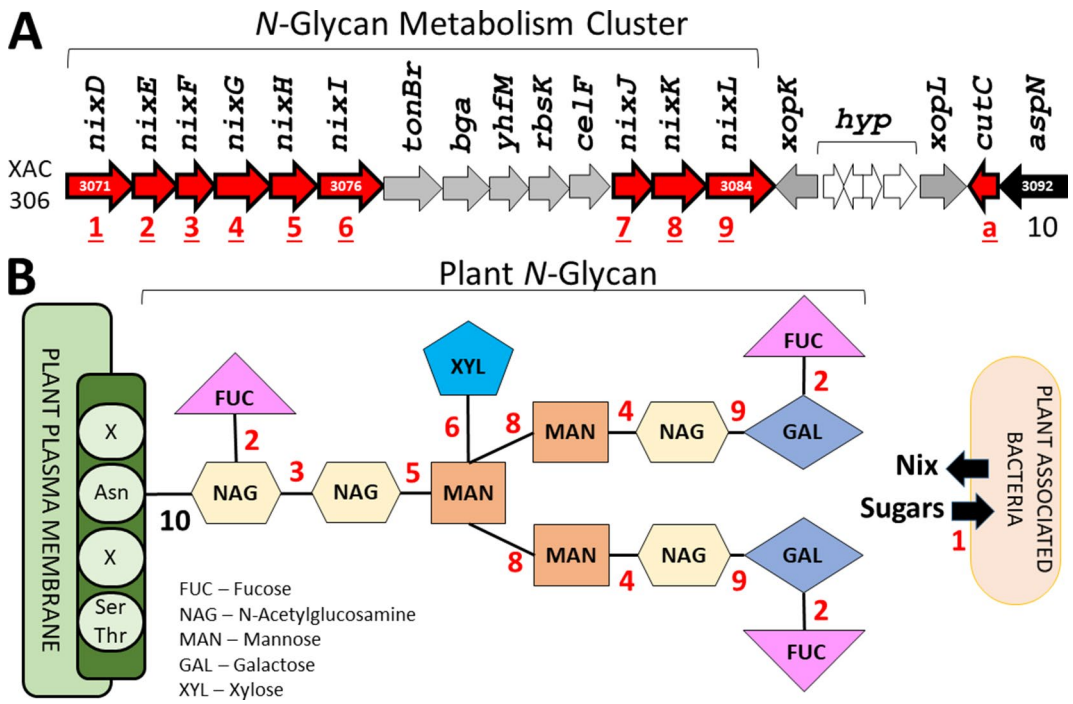


FIGURE 9 | Identification of the genes related with plant N-glycan degradation. **(A)** N-Glycan metabolism gene cluster in Xac306 genome. Red – Genes identified as exclusive of plant-associated genomes. The numbers 1 to 10 identify all genes related to N-glycan degradation. a – Non-related to N-glycan degradation. **(B)** Model of plant N-glycan structure. The numbers 1 to 10 identify the catalytic site of the proteins coded by the genes described in **(A)**. Asn – Asparagine residue. Ser/Thr – Serine and threonine residues. X – Any residue.

DATA AVAILABILITY

The datasets generated for this study can be found in the GTACG online interface at http://143.107.58.250/reportXantho161.45/. The GTACG is an open source project available at https://github.com/caiorns/GTACG-backend and https://github.com/caiorns/GTACG-frontend.

AUTHOR CONTRIBUTIONS

CS and LD designed and implemented the comparative genomics framework. CS, RA, LM and LD selected the strains.CS and LD performed the *in silico* assays. CS, RA, LM and LD analyzed the results and wrote the manuscript. CS, RA, LM and LD revised the manuscript.

FUNDING

This work was supported by the following agencies: Saõ Paulo Research Foundation – FAPESP (process 2018/03428-5), and Coordination for the Improvement of Higher Education Personnel – CAPES (the BIGA Project, CFP 51/2013, process 3385/2013). LM has a research fellowship from CNPq. CS has a PhD fellowship from CAPES.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00725/full#supplementary-material

REFERENCES

Aparna, G., Chatterjee, A., Sonti, R. V., and Sankaranarayanan, R. (2009). A cell wall-degrading esterase of xanthomonas oryzae requires a unique substrate recognition module for pathogenesis on rice. *Plant Cell* 21, 1860–1873. doi: 10.1105/tpc.109.066886

Assis, R. d. A. B., Polloni, L. C., Patané, J. S. L., Thakur, S., Felestrino, B., Diaz-Caballero, J., et al. (2017). Identification and analysis of seven effector protein families with different adaptive and evolutionary histories in plant-associated members of the xanthomonadaceae. *Sci. Rep.* 7, 16133. doi: 10.1038/s41598-017-16325-1

Benedict, M. N., Henriksen, J. R., Metcalf, W. W., Whitaker, R. J., and Price, N. D. (2014). Itep: an integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* 15, 8. doi: 10.1186/1471-2164-15-8

Berg, E. L. (2014). Systems biology in drug discovery and development. *Drug Discov. Today* 19, 113–125. doi: 10.1016/j.drudis.2013.10.003

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nat. Rev. Genet.* 14, 333. doi: 10.1038/nrg3433

Blanvillain, S., Meyer, D., Boulanger, A., Lautier, M., Guynet, C., Denancé, N., et al. (2007). Plant carbohydrate scavenging through tonb-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS One* 2, e224. doi: 10.1371/journal.pone.0000224

Borg, I., and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. New York, USA: Springer Science & Business Media.

Boulanger, A., Zischek, C., Lautier, M., Jamet, S., Rival, P., Carrère, S., et al. (2014). The plant pathogen xanthomonas campestris pv. campestris exploits n-acetylglucosamine during infection. *MBio* 5, 1–14. doi: 10.1128/mBio.01527-14

Brittnacher, M. J., Fong, C., Hayden, H., Jacobs, M., Radey, M., and Rohmer, L. (2011). Pgat: a multistrain analysis resource for microbial genomes. *Bioinformatics* 27, 2429–2430. doi: 10.1093/bioinformatics/btr418

Campbell, K., Xia, J., and Nielsen, J. (2017). The impact of systems biology on bioprocessing. *Trends Biotechnol.* 35, 1156–1168. doi: 10.1016/j.tibtech.2017.08.011

Casadesús, J., and Low, D. (2006). Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* 70, 830–856. doi: 10.1128/MMBR.00016-06

Chaudhari, N. M., Gupta, V. K., and Dutta, C. (2016). Bpga-an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* 6, 24373. doi: 10.1038/srep24373

Chervitz, S. A., Deutsch, E. W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P. et al., (2011). "Data standards for omics data: the basis of data sharing and reuse," in *Bioinformatics for Omics Data.* (New York, USA: Springer), 31–69. doi: 10.1007/978-1-61779-027-0_2

Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* 26, 721–744. doi: 10.1146/annurev-cellbio-100109-104122

Clarridge, J. E. (2004). Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17, 840–862. doi: 10.1128/CMR.17.4.840-862.2004

Contreras-Moreira, B., and Vinuesa, P. (2013). Get_homologues, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* 79, 7696–7701. doi: 10.1128/AEM.02411-13

Creevey, C., and McInerney, J. O. (2004). Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* 21, 390–392. doi: 10.1093/bioinformatics/bti020

Devos, D., and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet.* 17, 429–431. doi: 10.1016/S0168-9525(01)02348-4

Ding, W., Baumdicker, F., and Neher, R. A. (2017). panx: pan-genome analysis and exploration. *Nucleic Acids Res.* 46, e5–e5. doi: 10.1093/nar/gkx977

Dupoiron, S., Zischek, C., Ligat, L., Carbonne, J., Boulanger, A., de Bernonville, T. D., et al. (2015). The n-glycan cluster from xanthomonas campestris pv. campestris a toolbox for sequential plant n-glycan processing. *J. Biol. Chem.* 290, 6022–6036. doi: 10.1074/jbc.M114.624593

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

Enright, A. J., Van Dongen, S., and Ouzounis, 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

Field, D., Sansone, S.-A., Collis, A., Booth, T., Dukes, P., Gregurick, S. K., et al. (2009). 'omics data sharing. *Science* 326, 234–236. doi: 10.1126/science.1180598

Green, M., and Karp, P. (2005). Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic Acids Res.* 33, 4035–4039. doi: 10.1093/nar/gki711

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hillmer, R. A. (2015). Systems biology for biologists. *PLoS Pathog.* 11, e1004786. doi: 10.1371/journal.ppat.1004786

Häweker, H., Rips, S., Koiwa, H., Salomon, S., Saijo, Y., Chinchilla, D., et al. (2010). Pattern recognition receptors require n-glycosylation to mediate plant immunity. *J. Biol. Chem.* 285, 4629–4636. doi: 10.1074/jbc.M109.063073

Jalan, N. U. (2012). *Comparative genomic and transcriptomic analyses of xanthomonas citri subsp. citri and related species provides insights into virulence and host-specificity.* Ph.d. thesis, University of Florida. doi: 10.1186/1471-2164-14-551

Klimke, W., O'Donovan, C., White, O., Brister, J. R., Clark, K., Fedorov, B., et al. (2011). Solving the problem: genome annotation standards before the data deluge. *Stand Genomic Sci.* 5, 168. doi: 10.4056/sigs.2084864

Kobourov, S. G. (2012). Spring embedders and force-directed graph drawing algorithms. 2012. Available: http://arxiv.org/abs/1201.3011

Laia, M. L., Moreira, L. M., Dezajacomo, J., Brigati, J. B., Ferreira, C. B., Ferro, M. I., et al. (2009). New genes of xanthomonas citri subsp. citri involved in pathogenesis and adaptation revealed by a transposon-based mutant library. *BMC Microbiol.* 9, 12. doi: 10.1186/1471-2180-9-12

Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., et al. (2010). Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* 11, 461. doi: 10.1186/1471-2105-11-461

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Lin, C.-H., Lee, C.-N., Lin, J.-W., Tsai, W.-J., Wang, S.-W., Weng, S.-F., et al. (2010). Characterization of xanthomonas campestris pv. campestris heat shock protein a (hspa), which possesses an intrinsic ability to reactivate inactivated proteins. *Appl. Microbiol. Biotechnol.* 88, 699–709. doi: 10.1007/s00253-010-2776-z

Lunak, Z. R., and Noel, K. D. (2015). A quinol oxidase, encoded by cyoabcd, is utilized to adapt to lower o2 concentrations in rhizobium etli cfn42. *Microbiology* 161, 203. doi: 10.1099/mic.0.083386-0

Moreira, L. M., Soares, M. R., Facincani, A. P., Ferreira, C. B., Ferreira, R. M., Ferro, M. I., et al. (2017). Proteomics-based identification of differentially abundant proteins reveals adaptation mechanisms of xanthomonas citri subsp. citri during citrus sinensis infection. *BMC Microbiol.* 17, 155. doi: 10.1186/s12866-017-1063-x

Nascimento, R., Gouran, H., Chakraborty, S., Gillespie, H. W., Almeida-Souza, H. O., Tu, A., et al. (2016). The type ii secreted lipase/esterase lesa is a key virulence factor required for xylella fastidiosa pathogenesis in grapevines. *Sci. Rep.* 6, 18598. doi: 10.1038/srep18598

Naushad, H. S., and Gupta, R. S. (2013). Phylogenomics and molecular signatures for species from the plant pathogen-containing order xanthomonadales. *PLoS One* 8, e55216. doi: 10.1371/journal.pone.0055216

Obolski, U., Gori, A., Lourenco, J., Thompson, C., Thompson, R., French, N., et al. (2018). Identifying streptococcus pneumoniae genes associated with invasive disease using pangenome-based whole genome sequence typing. *bioRxiv* 314666, 9. doi: 10.1101/314666

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Pieretti, I., Royer, M., Barbe, V., Carrere, S., Koebnik, R., Cociancich, S., et al. (2009). The complete genome sequence of xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited xanthomonadaceae. *BMC Genomics* 10, 616. doi: 10.1186/1471-2164-10-616

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

Ramos, P. L., Moreira-Filho, C. A., Van Trappen, S., Swings, J., Vos, P. D., Barbosa, H. R., et al. (2011). An mlsa-based online scheme for the rapid identification of stenotrophomonas isolates. *Mem. Inst. Oswaldo Cruz* 106, 394–399. doi: 10.1590/S0074-02762011000400003

Santiago, C., Pereira, V., and Digiampietri, L. (2018). Homology detection using multilayer maximum clustering coefficient. *J. Comput. Biol.* 25, 1328–1338. doi: 10.1089/cmb.2017.0266

Sharma, V., and Patil, P. B. (2011). Resolving the phylogenetic and taxonomic relationship of xanthomonas and stenotrophomonas strains using complete rpob gene sequence. *PLoS Curr.* 3. doi: 10.1371/currents.RRN1239

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Steinegger, M., and Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026. doi: 10.1038/nbt.3988

Tierney, M., and Lamour, K. (2005). An introduction to reverse genetic tools for investigating gene function. *Plant Health Instructor* 10. doi: 10.1094/PHI/A-2005-1025-01

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154. doi: 10.1016/j.mib.2014.11.016

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., et al. (2016). Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* 45, D535–D542. doi: 10.1093/nar/gkw1017

Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., et al. (2016). Msaviewer: interactive javascript visualization of multiple sequence alignments. *Bioinformatics* 32, 3501–3503. doi: 10.1093/bioinformatics/btw474

Zamioudis, C., and Pieterse, C. M. (2012). Modulation of host immunity by beneficial microbes. *Mol. Plant Microbe Interact.* 25, 139–150. doi: 10.1094/MPMI-06-11-0179

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., et al. (2014). Pangp: a tool for quickly analysing bacterial pan-genome profile. *Bioinformatics* 30, 1297–1299. doi: 10.1093/bioinformatics/btu017

Zhao, Y., Sun, C., Zhao, D., Zhang, Y., You, Y., Jia, X., et al. (2018). Pgap-x: extension on pan-genome analysis pipeline. *BMC Genomics* 19, 36. doi: 10.1186/s12864-017-4337-7

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., and Yu, J. (2011). Pgap: pan-genomes analysis pipeline. *Bioinformatics* 28, 416–418. doi: 10.1093/bioinformatics/btr655

Zhou, X., Yan, Q., and Wang, N. (2017). Deciphering the regulon of a GntR family regulator *via* transcriptome and ChIP-exo analyses and its contribution to virulence in Xanthomonas citri. *Mol. Plant Pathol.* 18 (2), 249–262. doi: 10.1111/mpp.12397

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

TECHNOLOGY AND CODE
published: 22 October 2019
doi: 10.3389/fgene.2019.01027

XitoSBML: A Modeling Tool for Creating Spatial Systems Biology Markup Language Models From Microscopic Images

*Kaito Ii[1], Kota Mashimo[1], Mitsunori Ozeki[1], Takahiro G. Yamada[1], Noriko Hiroi[1,2] and Akira Funahashi[1]**

[1] Systems Biology Laboratory, Department of Biosciences and Informatics, Keio University, Yokohama, Japan, [2] Laboratory of Physical Chemistry for Life Science, Faculty of Pharmaceutical Sciences, Sanyo-Onoda City University, Sanyo-Onoda City, Japan

XitoSBML is a software tool designed to create an SBML (Systems Biology Markup Language) Level 3 Version 1 document from microscopic cellular images. It is implemented as an ImageJ plug-in and is designed to create spatial models that reflect the three-dimensional cellular geometry. With XitoSBML, users can perform spatial model simulations based on realistic cellular geometry by using SBML-supported software tools, including simulators such as Virtual Cell and Spatial Simulator. XitoSBML is open-source and is available at https://github.com/spatialsimulator/XitoSBML/. XitoSBML is confirmed to run on most 32/64-bit operating systems: Windows, MacOS, and Linux.

Keywords: spatial model simulation, spatial modeling, systems biology markup language, ImageJ, image processing

OPEN ACCESS

Edited by:
Helder Nakaya,
University of São Paulo, Brazil

Reviewed by:
Lucian Paul Smith,
University of Washington, United States
Nicholas Roehner,
Raytheon, United States

***Correspondence:**
Akira Funahashi
funa@bio.keio.ac.jp

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

Received: *25 March 2019*
Accepted: *24 September 2019*
Published: *22 October 2019*

Citation:
Ii K, Mashimo K, Ozeki M, Yamada TG, Hiroi N and Funahashi A (2019) XitoSBML: A Modeling Tool for Creating Spatial Systems Biology Markup Language Models From Microscopic Images. Front. Genet. 10:1027. doi: 10.3389/fgene.2019.01027

INTRODUCTION

With the recent development of imaging technologies, we can quantitatively analyze spatial localization and concentration gradients of biochemicals within living cells (Chen et al., 2014; Keller and Ahrens, 2015). As a result, the importance of biochemical spatial localization and concentration gradients has become apparent. The effect of dynamics related to biochemical spatial distribution and cellular shape can be analyzed by using spatial model simulations (Rangamani et al., 2013).

However, in most spatial model simulations, cellular regions are defined as two- or three-dimensional spatial models based on simple mathematical equations. Because cell shape in such models differs from actual cells, these simulations will not produce appropriate results. Therefore, it is crucial to perform three-dimensional spatial model simulations by using spatial models with the actual cellular shape.

Moreover, the advance of microscopic imaging technologies has made it possible to acquire a considerable amount of microscopic cellular images from biological experiments. Therefore, providing a software tool that can automatically generate a spatial model from microscopic cellular images will play an essential role in Systems Biology.

Software tools such as Virtual Cell (Loew and Schaff, 2001), Smoldyn (Andrews et al., 2010), and Morpheus (Starruss et al., 2014) are capable of performing simulations with actual cellular shape. Virtual Cell is a computational environment for modeling and numerical simulation that provides a graphical user interface (GUI) to create biochemical network models and perform ordinary differential equations, partial differential equations, and stochastic numerical simulations. Virtual Cell is popular for its partial differential equation model simulation. Smoldyn is a stochastic

model simulator that can perform spatial model simulation with a particle-based model. The molecules in the model are defined as particles that diffuse with Brownian motion. The software is mostly used for biochemical reaction simulation at the single-cell scale (e.g., nanometer-scale spatial resolution). Morpheus is a modeling environment for simulation with ordinary or partial differential equations and can be used to model a reaction-diffusion system for multiscale and multicellular systems. These software tools provide outstanding functionalities in terms of spatial modeling and simulation but are limited by their unique file format. Morpheus can import SBML (Systems Biology Markup Language) (Hucka et al., 2003) files but cannot import the spatial information from these files. Virtual Cell can import and export SBML files, including spatial information. Smoldyn supports the SBML file format using Virtual Cell as a proxy. Virtual Cell, Smoldyn, and Morpheus are also limited in that they do not offer an interface to overcome the difficulty of creating a spatial model from images.

Virtual Cell provides functionality to create a spatial model from microscopic images and export it as a spatial SBML document, but it only supports the import of grayscale or multichannel TIFF images; this is problematic for the following reasons: 1) When importing a grayscale image, users have to apply a segmentation task manually from the distribution of intensity provided by Virtual Cell. Because, in general, segmentation is a difficult task in image processing (Rajasekaran et al., 2016), manual segmentation for each organelle with the distribution of intensity would requires enormous modifications to each pixel in the image; and 2) Even though Virtual Cell supports the import of a multichannel TIFF image so that each channel can be segmented and assigned to each organelle, the program requires users to manually assign a membrane between two organelles in their model. The number of possible membrane positions would increase with $O(n^2)$ [$\binom{n}{2}$, where n is the number of organelles]. When the number of organelles is small, it would not be a critical problem for users but advances in microscopic technologies have enabled 9 to 24 multichannel images to be obtained for a single cell (Niehörster et al., 2016; Wei et al., 2017).

To solve these problems, we present XitoSBML, which is capable of creating spatial models from microscopic images in SBML format. XitoSBML uses images to construct spatial models with more flexibility in defining compartment shapes compared to models created with mathematical equations. Because XitoSBML is implemented as a plug-in for ImageJ (Schindelin et al., 2012; Schneider et al., 2012), users can call sophisticated segmentation algorithms for each channel through the user interface of ImageJ and directly apply the segmentation result to XitoSBML. This means that users can process images and create the spatial model within the same application. XitoSBML supports the import of plural segmented (binary) images for each organelle. Moreover, XitoSBML automatically assigns membranes between domains from given inclusion properties of organelles. XitoSBML also automatically adjusts the segmentation result so that users can create a spatial model without manually performing morphological operations and interpolation on the segmented images. SBML is compatible with more than 290 different software tools. Although only a limited number of software tools currently support spatial SBML simulation (Loew and Schaff 2001; Matsui et al., 2015), the demand for such modeling is increasing. We therefore expect that the number of spatial simulators will increase in the next few years. XitoSBML is a user-friendly and extensive modeling software, providing the environment to create a spatial model on the fly. Users may efficiently perform spatial model simulations and export the model to any compatible simulator.

MATERIALS AND METHODS

Here, we briefly describe XitoSBML and outline the process the program takes to create a model with JSBML 1.2 (Dräger et al., 2011). XitoSBML operates as a plug-in for ImageJ. Once the microscopic images are well organized and segmented (Nketia et al., 2017), XitoSBML can take them as input to create an SBML level 3 version 1 model with spatial processes (Schaff et al., 2015).

Software Architecture

XitoSBML is open-source software distributed under Apache License, 2.0; it is written in Java and is platform-independent. XitoSBML uses an ImageJ plug-in application programming interface (API) to import images from ImageJ and to create the GUI (**Figure 1**). The imported images are passed to several image processing algorithms (morphological operations, interpolation, and labeling) implemented in XitoSBML. The JSBML API then converts the processed images to a spatial SBML model, which is converted to an SBML Level 3 Version 1 object that can be modified through the XitoSBML GUI. The converted spatial SBML model contains the spatial geometry of the original images as well as information on molecular concentrations, locations, biochemical reactions, and parameters. This information will be used by SBML-supported simulators to perform spatial simulation on the model.

Preprocess Of Images

XitoSBML takes in two-dimensional or z-stack three-dimensional images as input and creates a spatial model. Before doing so, the images must be in a specific format as outlined below. XitoSBML assumes the image is segmented and represented as a specific domain within the cell. Usually, the segmented image is a binary image that only contains black or white pixels. For example, input images with a segmented image of a nucleus and a cytosol will produce a spatial model with domains of extracellular matrix, cytosol, and nucleus. Therefore, to obtain a reasonable model, segmentation of the microscopic images is essential. ImageJ provides a variety of tools for this purpose. One of the benefits of XitoSBML is that one can process the image and create the model simultaneously on ImageJ: i.e., the user just has to process the images on ImageJ and import them into XitoSBML on ImageJ.

Software Functionalities

Creating Spatial Model From Images

XitoSBML provides an easy-to-use GUI to create the spatial models. Before doing so, the microscopic images must be processed to binary representing one component of the cell;

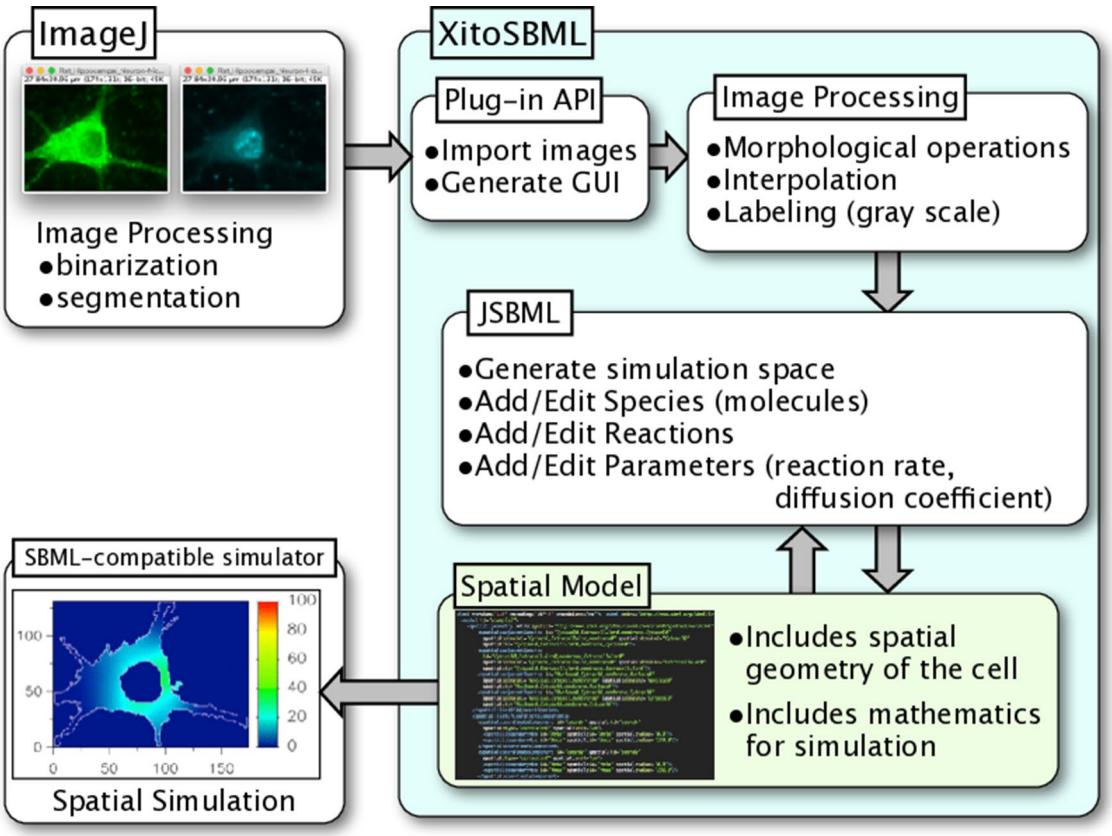


FIGURE 1 | Architecture and workflow of XitoSBML. Inputs to XitoSBML are segmented images of a cell or spatial SBML model. The segmented images are easily generated by using existing functions in ImageJ. XitoSBML imports these images through ImageJ plug-in API and applies several image processing algorithms implemented in XitoSBML. The converted images are then exported to spatial SBML by using JSBML API. The generated spatial SBML model is used to perform a spatial simulation on SBML-supported simulators.

this can be performed on ImageJ. Given the input images, the software will generate a spatial model as follows.

1. The binary images (**Figure 2A**) are filled by morphological operation and interpolated if necessary for the sake of simulation (**Figure 2B**).
2. The software then combines the images into a single grayscale image, assigning a distinct pixel value to each component given by the input (**Figure 2C**).
3. After the software generates a simulation space from the given images, users may add molecular species, parameters, and reactions to the model.
4. The resulting image is visualized by surface rendering using a 3D Viewer (Schmid et al., 2010) (**Figure 2D**).
5. In addition, the inclusion property between domains is shown (**Figure 2E**). Using this relationship, one can check whether the domains in the model are biologically valid by showing which domains are adjacent to each other. Thus, the program can determine whether a model is biologically impossible: e.g., nucleus adjacent to the extracellular matrix.
6. Finally, the model is exported as an SBML document, along with the grayscale image (**Figure 2F**).

Figures 2G, H show the post-processing of imported images. When merging two segmented images, a gap might occur between two segmented regions (e.g., nucleus and cytosol) when the segmentation did not work correctly. XitoSBML will automatically fill the gap between these two regions by a morphological operation (**Figure 2G**). Most of the three-dimensional microscopic images (z-stack images) have low resolution on the z-axis. This induces anisotropic voxels in the spatial model, which in turn would cause inaccurate spatial simulation. To solve this problem, XitoSBML interpolates z-slice images from the given input images by the nearest neighbor method (**Figure 2H**). Common pitfalls of segmentation are covered by applying morphological operation and interpolation as a post-process.

Figure 3 shows how the domains are written in SBML. From the original image (e.g., **Figure 2A**), each domain is assigned a specific pixel value creating a single grayscale image (e.g., **Figure 2C**). In the example in **Figure 3** (left side), Nuc (nucleus) has a value of 170, Cyt (cytosol) has a value of 85, and EC (extracellular matrix) has a value of 0. From the grayscale image, the adjacency of domains is found, and membranes (with no thickness) are created between the domains. After the

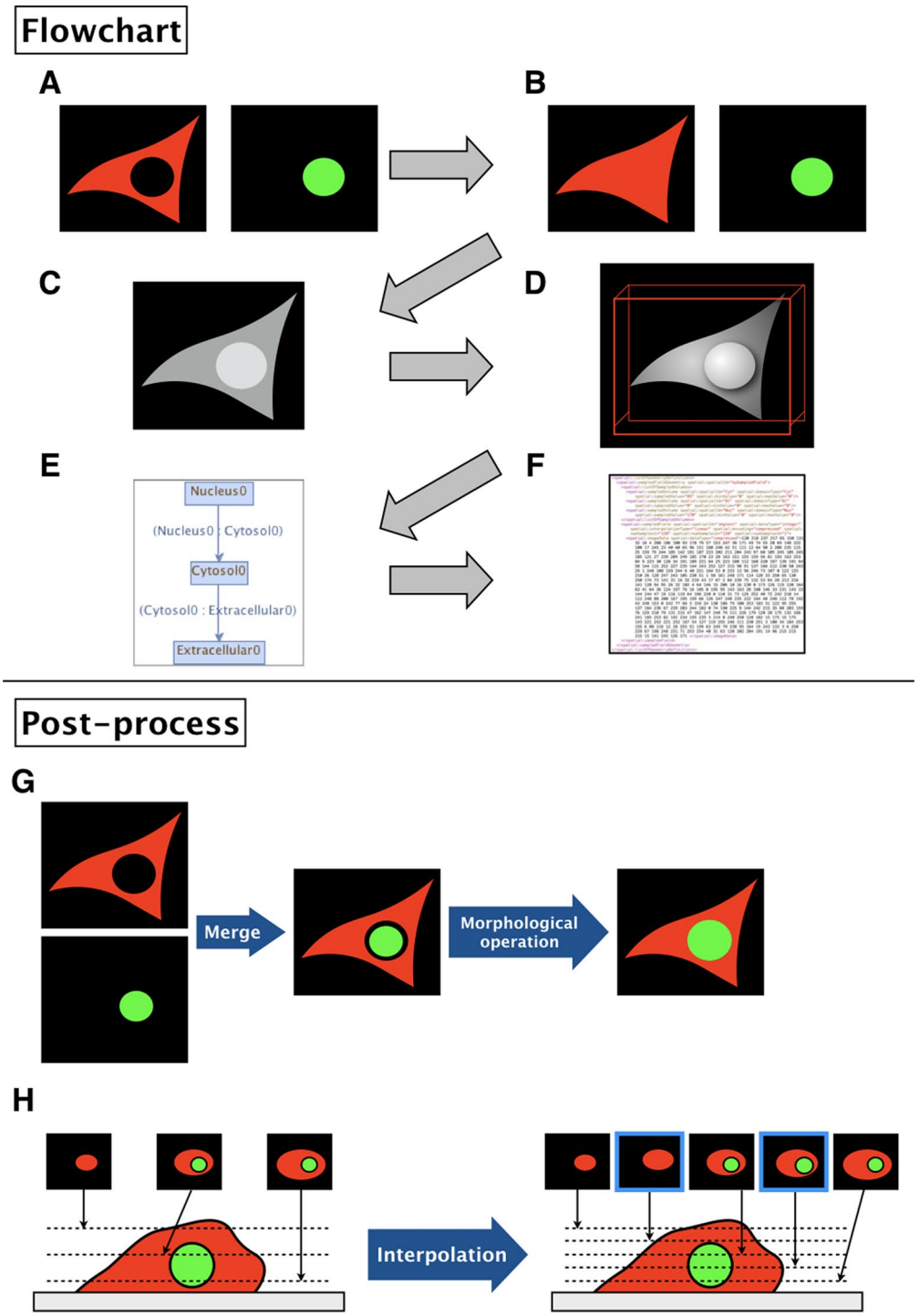


FIGURE 2 | Flowchart of XitoSBML and post-processing of imported images in XitoSBML. **(A)** Segmented images of cytosol (red) and nucleus (green). The two images are in color for visualization purposes; however, when in actual use, they have to be binarized. These two images are set as inputs of XitoSBML. **(B)** Each input is filled by morphological operation and interpolated as necessary. The interpolation is performed with the nearest neighbor method. **(C)** The inputs are combined creating a new grayscale image. The domains, in this case, cytosol and nucleus, are assigned with the specific pixel value. If each domain overlaps with each other or creates a gap in between, the grayscale images are corrected. **(D)** The result of the grayscale image projected three-dimensionally with 3D viewer. Each color represents a different domain of the input. **(E)** Inclusion property within the model. The box refers to the domain, and the arrow refers to the adjacency of domains, which apparently corresponds with **(C)**. **(F)** The resulting model for the SBML document. **(G)** When a gap exists between two segmented regions (e.g., nucleus and cytosol), XitoSBML will automatically fill the gap by a morphological operation. **(H)** If the imported simulation space contains anisotropic voxels, XitoSBML will use the nearest neighbor method to interpolate z-slice images from the given input images.

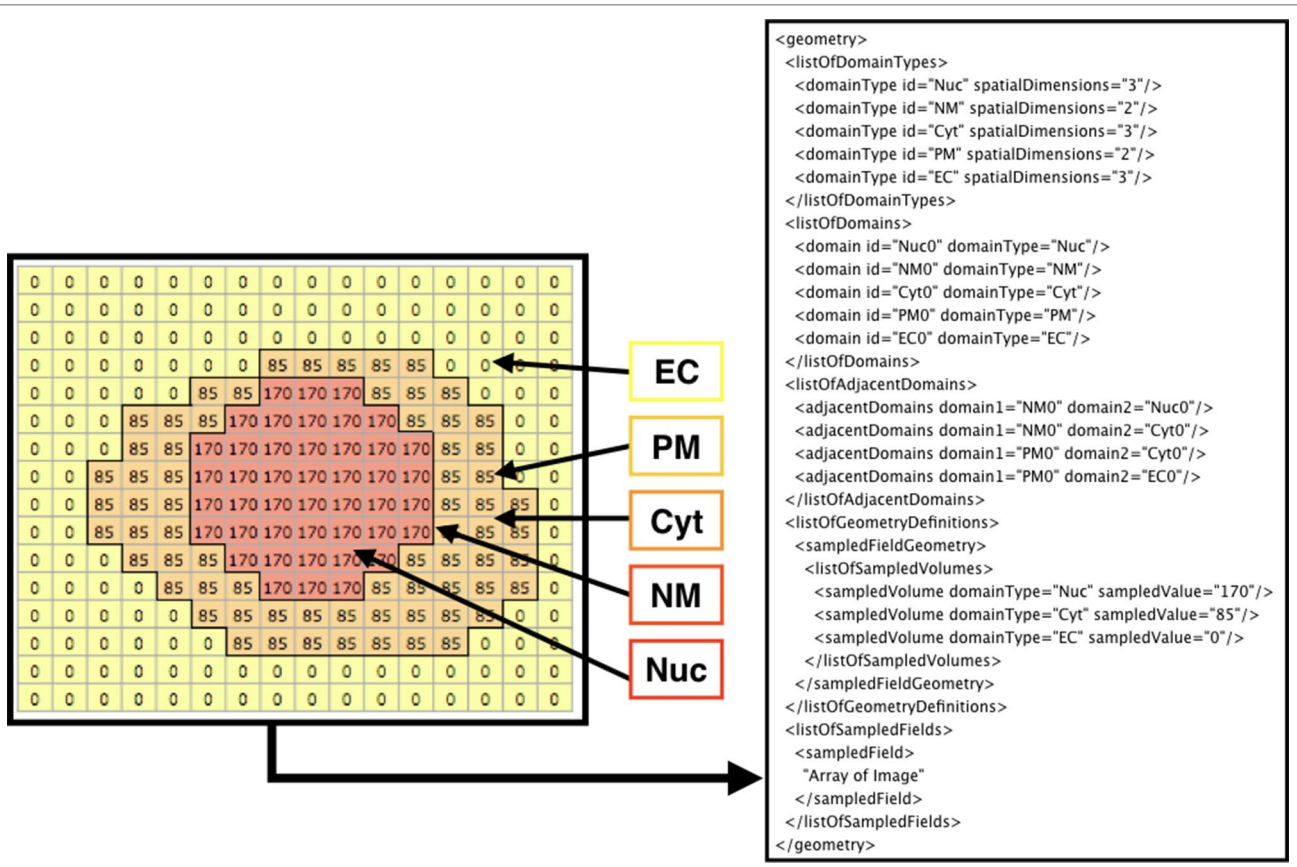


FIGURE 3 | Diagram of how the image is written in the SBML document. The array on the left side is a sample image representing the grayscale image in **Figure 2C**. Each pixel value corresponds to a domainType, and every contiguous region on a domainType is defined as the domain of that domainType. Membranes are created between the domains and defined in the SBML document. In this instance, EC (extracellular matrix), PM (plasma membrane), Cyt (cytosol), NM (nuclear membrane), and Nuc (nucleus) are defined as domains. While creating the membrane, the adjacency of domains is resolved and defined as adjacentDomain. The corresponding value of the domain, excluding membranes, is defined as the sampledValue. Lastly, the array of the whole image is stored in the sampledField.

domains are created from given images, users can manually add molecular species and parameters (e.g., advection coefficient, boundary condition, or diffusion coefficient) into the necessary domains by XitoSBML (model editor). Then, all the information is written in an SBML document. While exporting the spatial model as an SBML Level 3 Version 1 document, XitoSBML executes both syntax and semantic validation on the SBML core package by using an API provided by JSBML and executes syntax validation on the SBML spatial package by using an online libSBML validator. Moreover, XitoSBML has a custom implementation of a validator that can semantically validate the spatial information inside the model.

The user needs to do only three easy steps to create a spatial model from images: 1) process the microscopic images to binary images, 2) add molecular species and parameters into the necessary domains, and 3) save the created model as a file.

Editing an Existing Model

XitoSBML also can handle existing spatial SBML models, thereby allowing users to modify their spatial SBML model by opening it from XitoSBML. Using the "run Model Editor" menu item from the "XitoSBML" plug-in menu, the molecular species and parameter in a model can be modified. With the correct version and extensions, any model can be modified.

RESULTS

In **Figure 4**, we present an example of the use of XitoSBML software to demonstrate the basic work flow. As an input, we will use three-dimensional images of SH-SY5Y cells, which are derived from human neuroblastoma. Before construction of the model, the images were segmented using ImageJ, with each segmented image representing a geometry of a domain in a cell.

XitoSBML offers an easy way to create a spatial model from images. The obtained spatial model is usable for spatial model simulation with the appropriate simulator. Below, we present the result of a spatial model simulation with a spatial model exported from XitoSBML to validate the usefulness of the obtained spatial model. As an example simulator, we chose Spatial Simulator (Matsui et al., 2015), which is an in-house software implemented as a partial differential equation simulator specialized for SBML documents. Before performing

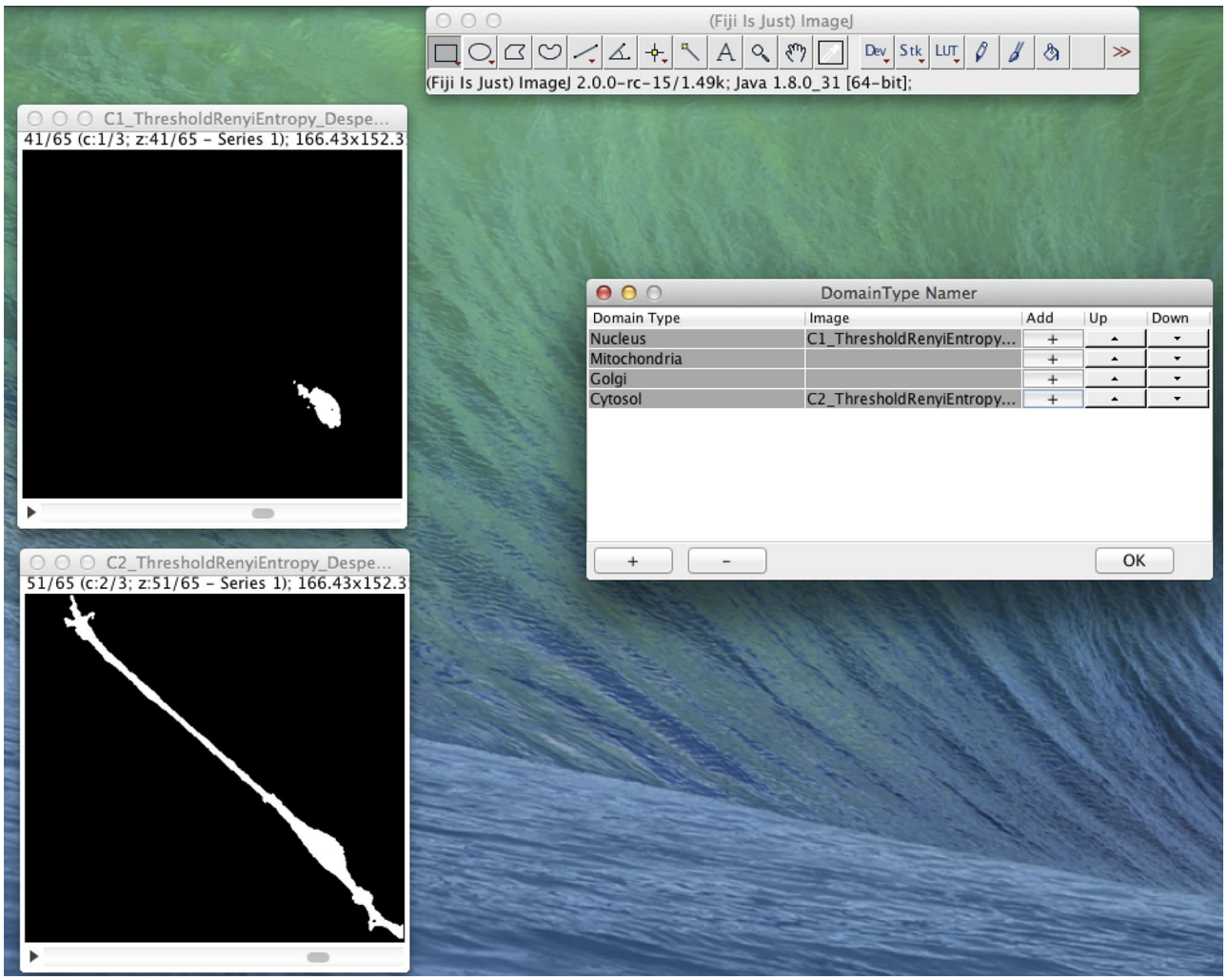


FIGURE 4 | Screenshot of XitoSBML generating an SBML document from microscopic images. The segmented three-dimensional SH-SY5Y cell images are shown on the left. Top left is the nucleus, and bottom left is the cytosol. On the right is the interface of XitoSBML that allows the user to add the images to the corresponding domainType.

the simulation, the XitoSBML output model requires further modification: information on biochemical reactions must be added because Spatial Simulator lacks the ability to add this information. XitoSBML provides a GUI for this purpose; users can add molecular species, reactions, parameters, and reaction rates to the model. In this example, the model is the simple transportation of molecules from Cyt to EC and simple diffusion combined. The result of the spatial model simulation is shown in **Figure S1**.

Even though the model is three-dimensional, to visualize the entire result, we show the results of a time series of a particular Z slice of the model. The colors inside the cell represent a concentration of molecule. By using XitoSBML and Spatial Simulator, users can easily create a spatial model, add mathematics to their model, and execute a spatial simulation from microscopic images.

DISCUSSION

Ever since the SBML Spatial Processes package was proposed, spatial models could be created in a standardized format. XitoSBML is one of the first software tools to create a pure SBML spatial model. Thus, we have provided a platform within a laboratory to perform spatial modeling in which acquisition of microscopic images and the addition of molecular species and parameters is conducted manually through the GUI of XitoSBML.

XitoSBML is a significant step toward more user-friendly tools for spatial biochemical modeling that provides the environment to create spatial models that reflect three-dimensional cellular geometry. It provides a GUI to easily create SBML Level 3 Version 1 documents and operates on ImageJ to simultaneously process images and create SBML documents. The exported model is compatible with SBML-supported software tools and can be

used to perform spatial modeling. Thus, XitoSBML works as the gateway between bioimaging and spatial model simulation. As such, it provides a fast and easy way for biologists, who do not have detailed knowledge of modeling but can produce microscopic z-stack images, to perform spatial model simulations.

In the future, XitoSBML will be extended to automatically add the distribution of the initial concentration for each molecular species: in this new functionality, the fluorescent microscopic image of the localization of the molecule would be received and added as the distribution of initial concentration for that molecule in the SBML model.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found in https://github.com/spatialsimulator/XitoSBML/.

AUTHOR CONTRIBUTIONS

AF conceived and led the project. KI implemented the software and wrote the manuscript with TY. NH provided biological expertise. AF gave technical advice on the implementation. KM and MO provided advice on the image processing algorithms implemented in this software. All authors were involved in drafting or revising the content of the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) (Grant Number 24300112) and the Imaging Science Project of the Center for Novel Science Initiatives (CNSI), National Institutes of Natural Sciences (NINS) (Grant Number IS271002).

ACKNOWLEDGMENTS

The authors thank Ryuichi Tanimoto, Keio University, and Yuta Tokuoka, Keio University for providing the three-dimensional SH-SY5Y cell images and for productive discussions on recent segmentation algorithms.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01027/full#supplementary-material

REFERENCES

Andrews, S. S., Addy, N. J., Brent, R., and Arkin, A. P. (2010). Detailed simulations of cell biology with smoldyn 2.1. *PLoS Comput. Biol.* 6, e1000705. doi: 10.1371/journal.pcbi.1000705

Chen, B.-C., Legant, W. R., Wang, K., Shao, L., Milkie, D. E., Davidson, M. W., et al. (2014). Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science* 346, 1257998–1257998. doi: 10.1126/science.1257998

Dräger, A., Rodriguez, N., Dumousseau, M., Dörr, A., Wrzodek, C., Le Novère, N., et al. (2011). JSBML: a flexible Java library for working with SBML. *Bioinformatics* 27, 2167–2168. doi: 10.1093/bioinformatics/btr361

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The systems biology markup language (SBML): a medium for representationand exchange of biochemical network models. *Bioinformatics* 19, 524–531. doi: 10.1093/bioinformatics/btg015

Keller, P. J., and Ahrens, M. B. (2015). Visualizing whole-brain activity and development at the single-cell level using light-sheet microscopy. *Neuron* 85, 462–483. doi: 10.1016/j.neuron.2014.12.039

Loew, L. M., and Schaff, J. C. (2001). The virtual cell: a software environment for computational cell biology. *Trends Biotechnol.* 19, 401–406. doi: 10.1016/s0167-7799(01)01740-1

Matsui, T., Mashimo, K., Ii, K., Hiroi, N., and Funahashi, A. (2015). Spatial simulator. Available at: https://github.com/spatialsimulator/Spatial_Simulator/

Niehörster, T., Löschberger, A., Gregor, I., Krämer, B., Rahn, H.-J., Patting, M., et al. (2016). Multi-target spectrally resolved fluorescence lifetime imaging microscopy. *Nat. Methods* 13, 257. doi: 10.1038/nmeth.3740

Nketia, T. A., Sailem, H., Rohde, G., Machiraju, R., and Rittscher, J. (2017). Analysis of live cell images: methods, tools and opportunities. *Methods* 115, 65–79. doi: 10.1016/j.ymeth.2017.02.007

Rajasekaran, B., Uriu, K., Valentin, G., Tinevez, J.-Y., and Oates, A. C. (2016). Object segmentation and ground truth in 3d embryonic imaging. *PloS ONE* 11, e0150853. doi: 10.1371/journal.pone.0150853

Rangamani, P., Lipshtat, A., Azeloglu, E. U., Calizo, R. C., Hu, M., Ghassemi, S., et al. (2013). Decoding information in cell shape. *Cell* 154, 1356–1369. doi: 10.1016/j.cell.2013.08.026

Schaff, J. C., Lakshminarayana, A., Smith, L., Bergmann, F., and Sullivan, D. P. (2015). *SBML level 3 package specification spatial processes*. Available at: https://sourceforge.net/p/sbml/code/HEAD/tree/trunk/specifications/sbml-level-3/version-1/spatial/specification/spatial-v1-sbml-l3v1-rel0.90.pdf

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–682. doi: 10.1038/nmeth.2019

Schmid, B., Schindelin, J., Cardona, A., Longair, M., and Heisenberg, M. (2010). A high-level 3d visualization api for java and imagej. BMC. *Bioinformatics* 11, 274. doi: 10.1186/1471-2105-11-274

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675. doi: 10.1038/nmeth.2089

Starruss, J., de Back, W., Brusch, L., and Deutsch, A. (2014). Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology. *Bioinformatics* 30, 1331–1332. doi: 10.1093/bioinformatics/btt772

Wei, L., Chen, Z., Shi, L., Long, R., Anzalone, A. V., Zhang, L., et al. (2017). Super-multiplex vibrational imaging. *Nature* 544, 465. doi: 10.1038/nature22051

Conflict of Interest: 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

BRIEF RESEARCH REPORT
published: 24 October 2019
doi: 10.3389/fgene.2019.00971

Assessing the Impact of Sample Heterogeneity on Transcriptome Analysis of Human Diseases Using MDP Webtool

*André N. A. Gonçalves[1], Melissa Lever[1], Pedro S. T. Russo[1], Bruno Gomes-Correia[2], Alysson H. Urbanski[1], Gabriele Pollara[3], Mahdad Noursadeghi[3], Vinicius Maracaja-Coutinho[2] and Helder I. Nakaya[1,4]**

[1] Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil, [2] Advanced Center for Chronic Diseases–ACCDiS, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile, [3] Division of Infection and Immunity, University College London, London, United Kingdom, [4] Scientific Platform Pasteur–USP, São Paulo, Brazil

OPEN ACCESS

Edited by:
Argyris Papantonis, University Medical Center Göttingen, Germany

Reviewed by:
Debashis Sahoo, University of California, San Diego, United States
Lin Zhang, China University of Mining and Technology, China

***Correspondence:**
Helder I. Nakaya
hnakaya@usp.br

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

Received: *09 April 2019*
Accepted: *11 September 2019*
Published: *24 October 2019*

Citation:
Gonçalves ANA, Lever M, Russo PST, Gomes-Correia B, Urbanski AH, Pollara G, Noursadeghi M, Maracaja-Coutinho V and Nakaya HI (2019) Assessing the Impact of Sample Heterogeneity on Transcriptome Analysis of Human Diseases Using MDP Webtool. Front. Genet. 10:971. doi: 10.3389/fgene.2019.00971

Transcriptome analyses have increased our understanding of the molecular mechanisms underlying human diseases. Most approaches aim to identify significant genes by comparing their expression values between healthy subjects and a group of patients with a certain disease. Given that studies normally contain few samples, the heterogeneity among individuals caused by environmental factors or undetected illnesses can impact gene expression analyses. We present a systematic analysis of sample heterogeneity in a variety of gene expression studies relating to inflammatory and infectious diseases and show that novel immunological insights may arise once heterogeneity is addressed. The perturbation score of samples is quantified using nonperturbed subjects (i.e., healthy subjects) as a reference group. Such a score allows us to detect outlying samples and subgroups of diseased patients and even assess the molecular perturbation of single cells infected with viruses. We also show how removal of outlying samples can improve the "signal" of the disease and impact detection of differentially expressed genes. The method is made available *via* the mdp Bioconductor R package and as a user-friendly webtool, webMDP, available at http://mdp.sysbio.tools.

Keywords: heterogeneity, transcriptome analysis, gene expression profiling, infectious diseases, inflammatory diseases

INTRODUCTION

Gene expression profiling methods such as microarrays and RNA-seq have been extensively used to examine the molecular changes associated with a biological "perturbation." This perturbation can be drug treatments, vaccinations, infections, cancers, and autoimmune or inflammatory diseases (Nakaya et al., 2012; Prada-Medina et al., 2017; Jochems et al., 2018). For human diseases, the initial analysis usually tries to find genes whose expression is significantly altered in the perturbed condition (i.e., patients with the disease) compared to the nonperturbed subjects (i.e., the healthy subjects). However, the definition of health and disease is broad, and the inherent variation among individuals can make any group of human samples highly heterogeneous. Variation can be due to genetic and environmental factors, as well as undetected health problems (Whitney et al., 2003; Albert and

Kruglyak, 2015). Similarly, patients with the same disease can present huge variation in terms of symptoms or score (Hersh and Prahalad, 2015; Garg and Smith, 2015). Thus, the removal of outlier samples can impact downstream analyses, especially in studies investigating mild diseases or the administration of inactivated vaccines.

Transcriptome datasets typically contain expression values of tens of thousands of genes from a relatively small number of samples. This presents a dimensionality problem when trying to identify significant changes in gene expression (Wang et al., 2008). Most methods will classify a gene as differentially expressed if there is a large difference in the mean expression between classes and a low variance within classes (De Hertogh et al., 2010). Therefore, genes that have heterogeneous expression within a class due to technical or biological outliers will have their detection as differentially expressed hindered. Individual heterogeneity can arise from past infections, environmental factors, microbiota, and genetics (Gibson, 2008), as well as undetected problems such as chronic disease, worms, food poisoning, or asymptomatic infection. In order to reduce biological heterogeneity, scientists try to enroll subjects with similar characteristics, controlling them for gender, clinical information, age, and so on. However, many hidden factors will invariably remain in the final set of samples and contribute to individual differences.

The molecular distance to health (Pankla et al., 2009) is a method that analyzes sample heterogeneity by scoring samples based on how distant their expression is to healthy and has been applied to quantify the perturbation of samples from diseased subjects (Berry et al., 2010; Banchereau et al., 2012; Bell et al., 2016). However, there has been no systematic assessment of how human heterogeneity affects downstream analyses. Also, none of the previous studies have used specific knowledge-based gene sets to evaluate subject perturbation or provided a tool for users to assess the heterogeneity in their own datasets.

Here we describe a systematic analysis on heterogeneity of several RNA-seq and microarray datasets from a diverse set of human diseases. Our approach, called the molecular degree of perturbation (MDP), is available as a Bioconductor R package (https://bioconductor.org/packages/release/bioc/html/mdp.html) and can identify potentially problematic subject data from transcriptomic dataset, as well as to quantify the perturbation score of healthy and diseased samples. Meanwhile, our user-friendly web-based application (https://mdp.sysbio.tools/) allows scientists to run MDP without any knowledge of bioinformatics or programming languages. We demonstrated that the application of our method on inflammatory and infectious disease datasets can affect the detection of differentially expressed genes (DEGs). Finally, these tools were used to analyze RNA-seq data of single cells infected with dengue virus (DENV), revealing the individual cell heterogeneity of infected cells.

METHODS

MDP Algorithm

The MDP score measures how much a sample is distant from a reference group of samples. Let G be the genes in a given expression dataset with N samples, out of which h are the healthy control samples. Also, let C_i^h be a centrality measurement (either the mean or the median; the default is median), and S_i^h, a measure of the variability (the standard deviation or the MAD) for each gene i in the control samples. Finally, let z_i be a modified z-score transformation using C_i^h and S_i^h as parameters. The absolute values of z_i are taken, and values less than 2 are set to 0. The values that remain represent significant deviations from the healthy control samples. The MDP score for each sample j (both in the control and perturbed groups) is then the mean of the modified absolute z_i values considering all genes or just the perturbed ones. The "perturbed genes" represent the top (default is 25%) genes with the highest absolute z_i values across all samples in a perturbed group. Additionally, the MDP package can automatically identify outlier samples based on the number of standard deviations (default = 2) from the mean of MDP scores of all samples within each class.

Data Acquisition and Processing

Normalized gene expression data from RNA-seq and microarray studies were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). If normalized data were not available, we processed the raw CEL files using the affy Bioconductor R package (Gautier et al., 2004) and performed data quality control using the arrayQualityMetrics Bioconductor R package (Kauffmann et al., 2009). Normalization was performed using the "RMA" function from the affy package. Samples that failed at least two quality control tests before or after normalization were removed from downstream analyses. For the single-cell RNA-seq data, we utilized the gene counts table from Supplementary File 7 published by Zanini et al. (2018). Prior to the calculation of MDP on single-cell data, we kept only the top 30% genes with the highest mean expression on all single cells and then removed the genes with zero values in 40% or more single cells.

Differential Gene Expression Analysis

Student t test was used to identify DEGs between patients with a disease and the healthy subjects. Different $\log_2$ fold change and adjusted P value (Benjamini and Hochberg) cutoffs were used and are shown in **Table S1**.

Pathway and Network Analyses

We used the NetworkAnalyst tool (Xia et al., 2015) to create the protein–protein interaction network with the DEGs. For the JIA analysis, we used the protein–protein interaction database STRING (score >900) and the minimum network. For the single-cell RNA-seq analysis, we used the protein–protein interaction database STRING (score >900) and the zero-order network. Overrepresentation analyses using the Gene Ontology gene sets were performed using the genes in the networks. Cytoscape software (Shannon et al., 2003) was used to display the networks.

MDP Webtool Implementation

The code of the tool was implemented in HTML, CSS, JavaScript, PHP, and R. To upload files, check for errors and check the

structure of the data; we used the languages JavaScript and PHP. An R script containing the https://cloud.r-project.org repo packages: data.table, withr, ggplot2, plotly, and pandoc was used to process the data and generate the results in HTML.

For defining style and appearance of pages, we used CSS with Bootstrap, which is a front-end framework with several components included. For dynamic manipulation of the page, we used JavaScript with Jquery. The latter is a framework for JavaScript itself, where its main purpose is to facilitate, streamline, and reduce the complexity in development.

In the infrastructure, we used the concept of containers and microservice with the platform Docker. In parallel, we used the tool Docker Compose to orchestrate and to deploy these containers. In total, we have three containers: proxy, nginx, and php-fpm. In the proxy container, the functions of reverse proxy and load balancing were performed, which were left in charge of the traefik service (https://traefik.io/). It also implements SSL certificate management through the Let's Encrypt project (https://letsencrypt.org/). The nginx container is our webserver, and the php-fpm is the backend that processes requests to php files.

RESULTS

Molecular Degree of Perturbation Algorithm and Webtool

We developed a user-friendly tool that inspects sample heterogeneity by assigning a score to each sample based on the cumulative perturbation of its gene expression levels relative to control samples. The algorithm performs a *Z*-score normalization of gene expression values for noncontrol samples, using the control samples to compute the median (M) and median absolute deviation (MAD). Absolute normalized expression values less than 2 are designated as unperturbed and are set to 0. Sample MDP scores are the average of normalized expression values for a given gene set (**Figure 1A**).

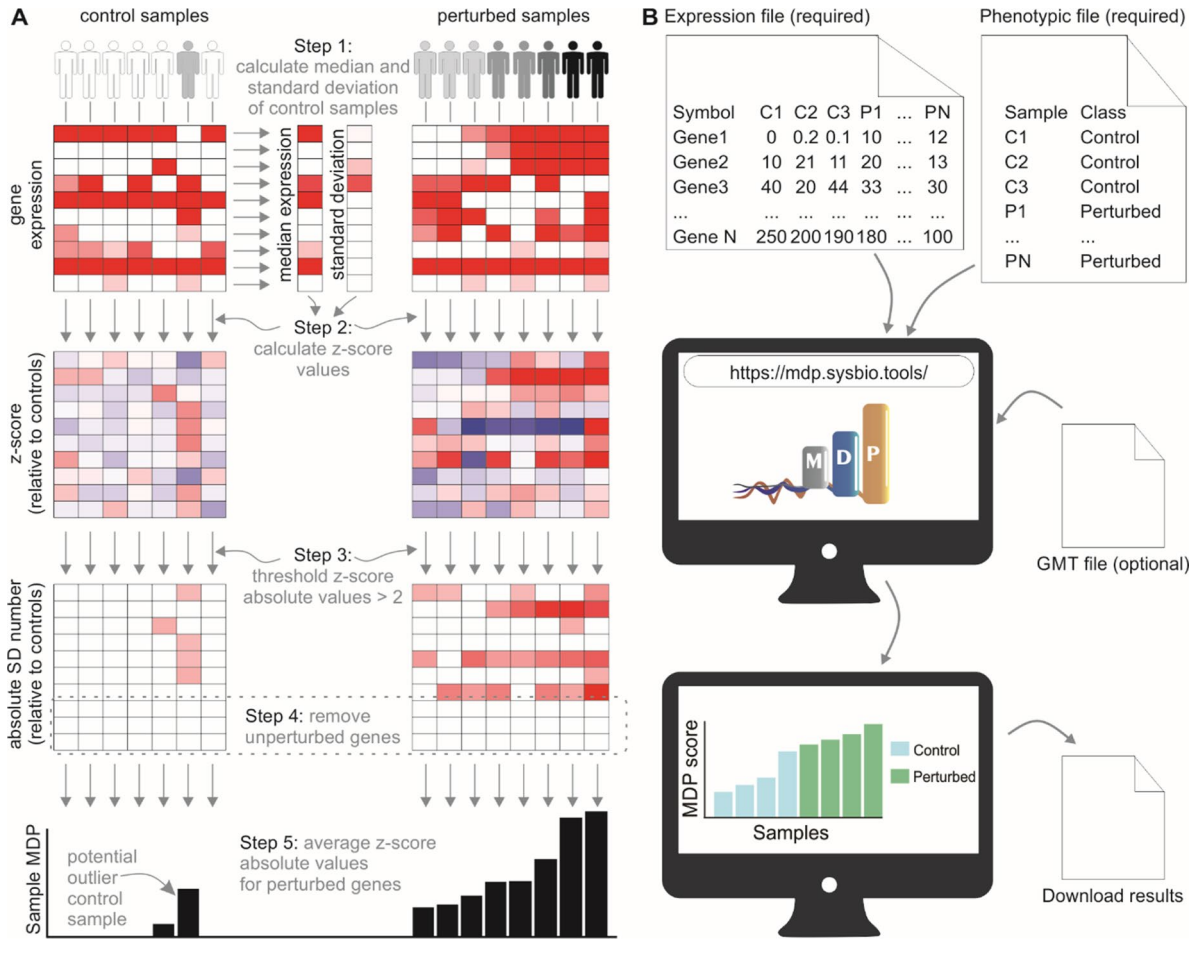


FIGURE 1 | The molecular degree of perturbation approach to calculating sample heterogeneity. **(A)** The MDP algorithm scores samples based on their perturbation from user-defined control samples (often healthy subjects). A *Z*-score normalization is performed using the control samples as a reference. The absolute values of the normalized scores are then taken, and values below 2 are set to 0. The sample scores are the average of these gene scores for each sample. **(B)** Running the MDP webtool. Expression and phenotypic files are required to run MDP; the results are a simple barplot and boxplot showing molecular perturbation for each submitted sample. An optional feature allows users to run MDP using a specific gene set, provided as a.gmt file.

The web interface for MDP (http://mdp.sysbio.tools) has been developed to allow non–bioinformatics users to quickly assess the MDP in their samples without the need for any previous computational knowledge or additional software (**Figure 1B**). The minimal requirements to execute the webtool are the input gene expression file and the phenotype data file. As long as the data are already normalized (CPM, TMM, FPKM, RMA, etc.), gene expression data from both RNA-seq and microarray experiments are supported.

The MDP tool has an additional feature that allows users to assess the MDP using a specific gene set or pathway. This may be useful in cases where there is a prior knowledge about the pathways involved with the disease. For running this optional analysis, users must provide a pathway annotation file in.gmt format and then select a specific gene set or pathway to calculate the perturbation score.

The Sample Perturbation Score for Different Human Diseases

We applied the MDP to 20 transcriptome studies (11 microarray and 9 RNA-seq) obtained from the GEO (Edgar et al., 2002) and SRA (Leinonen et al., 2011) databases in order to investigate how sample heterogeneity can impact the downstream differential expression analysis. Studies were related to tuberculosis (TB), cancer, juvenile idiopathic arthritis (JIA), sepsis, and other autoimmune and infectious diseases.

We initially showed that the perturbation scores of samples broadly vary within and between different diseases or treatments (**Figure S1**). Infection with the bacteria *Staphylococcus aureus*, for instance, seems to be a stronger perturbation than infection with influenza virus (**Figure S1A**) (Ramilo et al., 2007). Similarly, different types of cancer may show lower or higher perturbation scores regardless of their known prognostic values (**Figure S1B**) (Best et al., 2015). Our approach also differentiates between several subtypes of inflammatory diseases such as JIA, Crohn disease, and ulcerative colitis (**Figure S1C**) (Mo et al., 2018).

MDP Identifies Potential Outlier Samples

By assessing the sample perturbation scores, we were able to identify potential outlier samples for each of the 20 microarray and RNA-seq studies. One representative boxplot (**Figure 2A**) shows that one of the healthy subjects may be in fact "perturbed" when

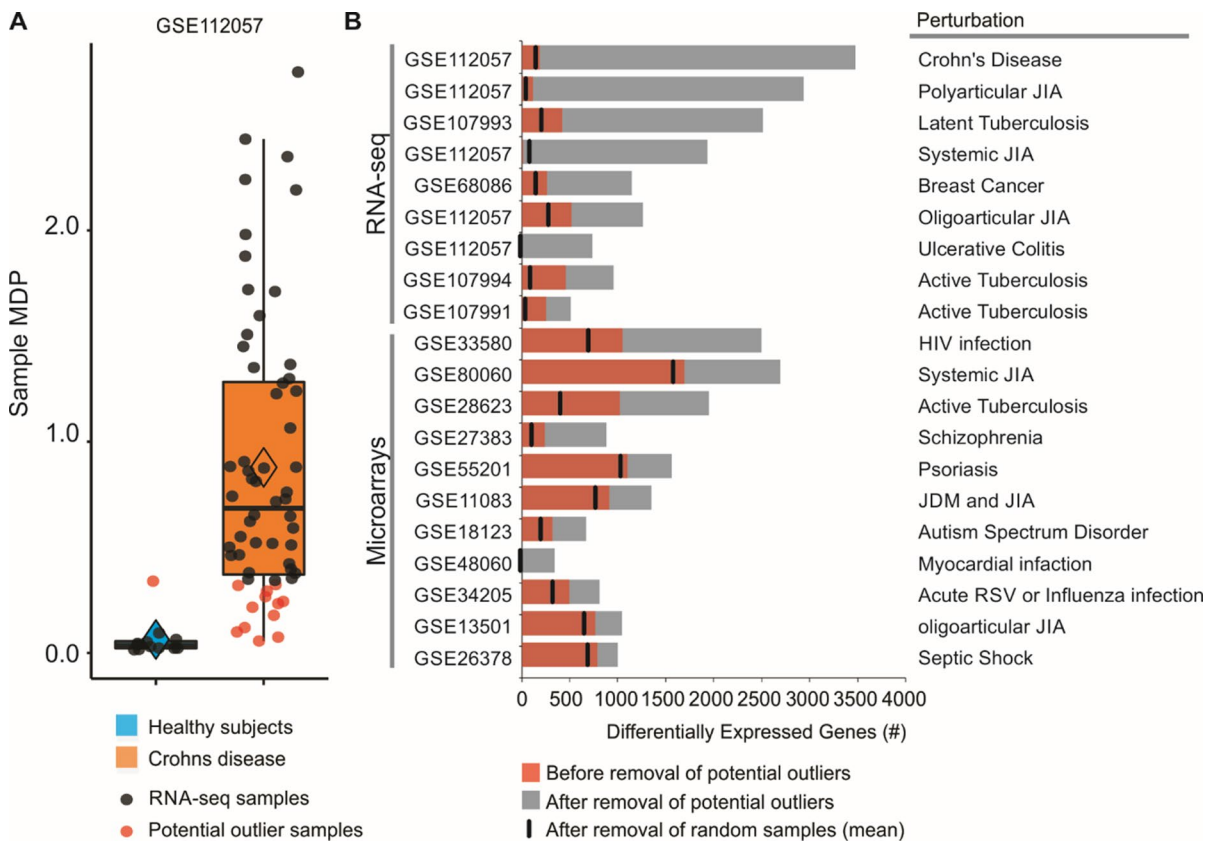


FIGURE 2 | Removal of potential outlier samples impacts differential expression analyses. **(A)** Sample MDP scores were calculated for 60 patients with Crohn disease using as a reference group 12 healthy subjects. Data were obtained from whole blood and are available under GEO accession GSE112057. Healthy subjects (blue) were used as reference group. Potential outlier samples are shown as red dots. **(B)** Differential expression analyses between patients with a disease and healthy controls. Numbers of DEGs before and after removal of potential outlier samples are shown as red and black bars, respectively. Random removal of samples followed by differential expression analysis was performed 1,000 times for each comparison, and the number of DEGs was averaged (black vertical line).

compared to the rest of the healthy group. Similarly, 12 of Crohn disease patients do not seem greatly perturbed at the molecular level (**Figure 2A**). Treating these samples as outliers and thus removing them from differential expression analyses increased the number of DEGs. For the GSE112057 comparison between healthy subjects and Crohn disease patients, we identified 188 DEGs before the removal of outliers (**Figure 2B**). After removal, the number of DEGs for this comparison was 3,477 (18.50-fold increase). If only the single control outlier sample is removed (**Figure 2B**), the number of DEGs increases to 1,931 (10.1-fold increase). We also randomly removed the same number of samples considered as outliers and counted the number of DEGs for each comparison. This process was repeated 1,000 times showing that the increase in DEG number is not due to random chance (**Figure 2B**). We performed this analysis for the 19 other comparisons as well. In all of them, the number of DEGs increased after removing the potential outliers (**Figure 2B**).

Removal of Potential Outlier Samples Increases Biological Consistency Across Similar Studies

Five JIA datasets (three RNA-seq and two microarrays) were used to assess the consistency between DEGs before and after removal of potential outlier samples identified by MDP. After removal, we found 21 genes that were differentially expressed in at least four JIA datasets, and none using all original samples (**Figure 3A**). Overrepresentation analysis of the genes consistently up-regulated in three or more datasets revealed that the top 1 gene set, neutrophil degranulation (GO:0043312), was recently associated with JIA (Brown et al., 2018) (**Figure 3B**). We then created a protein–protein interaction network with these consistently up-regulated genes (**Figure 3C**). This approach revealed highly connected genes, which may be central to JIA, such as STAT3, UBE2D1, MAPK14, and TLR4 (**Figure 3C**).

Using a Specific Gene set to Determine the MDP

T cells play a critical role in the outcome of *Mycobacterium tuberculosis* infection (Jasenosky et al., 2015). One important cytokine released by these cells is interferon gamma (IFNg). However, Berry et al. (2010) have shown that the blood transcriptome of patients with active TB was dominated by neutrophil-driven type I IFN-related genes. We thus decided to evaluate if gene modules related to specific blood immune cell populations can capture the MDP of patients with active TB. In the analysis, we used transcriptional modules that have been extensively validated to be highly specific for different immune cell types (Pollara et al., 2017). We also used modules derived from the unique transcriptome of human monocyte-derived macrophages (Mϕ) stimulated *in vitro* with different cytokines (Bell et al., 2016). For the study GSE19435 (Berry et al., 2010), the sample MDP scores calculated with gene modules of macrophages treated with IFNg for 4 h, neutrophils and T cells were higher in patients with active TB compared to those from healthy controls (**Figure S2A**). We also performed the same analysis for all 15 gene modules and all 7 TB datasets (**Figure S2B**) and found that the genes associated with macrophages treated with IFNg for 4 or 24 h are greatly perturbed in active TB. This analysis demonstrated that prior knowledge about a disease can be used to quantify sample perturbation and that the gene set used will impact the MDP scores.

MDP Analysis for Single-Cell RNA-Seq Dataset

Finally, we applied the MDP approach to analyze the molecular perturbation caused by a viral infection at single-cell level. Zanini et al. (2018) developed an approach named viscRNA-seq (virus-inclusive single-cell RNA-seq) to probe the host single-cell transcriptome together with intracellular viral RNA. We first evaluated if the MDP score was correlated with the DENV counts (herein defined as viral load or VL). Using uninfected single cells as the reference control, we calculated the MDP score for all cells infected with DENV and then compared these scores with VL (**Figure 4A**). No clear correlation was seen between MDP score and VL. Based on the VL (cutoff VL = 10^3) and on the MDP score (cutoff MDP = 1), we split the single cells into four subsets: $MDP^{high}VL^{low}$, $MDP^{high}VL^{high}$, $MDP^{low}VL^{low}$, and $MDPl^{ow}VL^{high}$. We then performed differential expression analyses between these subsets to assess the transcriptomic alterations caused by DENV infection. **Figure 4B** shows that the highest number of DEGs was found when we compared $MDP^{high}VL^{high}$ with $MDP^{low}VL^{low}$ subsets (1,158 DEGs), rather than either of these criteria alone. Comparing cells with high MDP score ($MDP^{high}VL^{low}$ + $MDP^{high}VL^{high}$) with those with low MDP score ($MDP^{low}VL^{low}$ + $MDPl^{ow}VL^{high}$) resulted in 872 DEGs. The lowest number of DEGs (196 DEGs) was found when we compared cells with high VL ($MDP^{high}VL^{high}$ + $MDP^{low}VL^{high}$) with those with low VL ($MDP^{high}VL^{low}$ + $MDPl^{ow}VL^{low}$) (**Figure 4B**). These results suggest that VL alone cannot be a strong marker of cell perturbation.

Network and pathway analyses were then performed on the 1,158 DEGs identified in the $MDP^{high}VL^{high}$ with $MDP^{low}VL^{low}$ comparison (**Figure 4C**). The top associated pathways were "regulation of cell cycle," "viral infectious cycle," and "endoplasmic reticulum unfolded protein response" (**Figure 4C**). In addition to VL, MDP provided another layer of information for quantifying heterogeneity at single-cell level and generated novel insights associated to viral infections.

DISCUSSION

We have shown that the MDP tool provides an intuitive way to inspect gene expression data and identify samples that are potential biological outliers. Although it can be argued that it is important to embrace the heterogeneity of samples and use all of them to perform analyses, we have shown that, for DEG analyses, sample removal can result in a dramatic improvement in the number of DEGs found, particularly removal of clear outlier

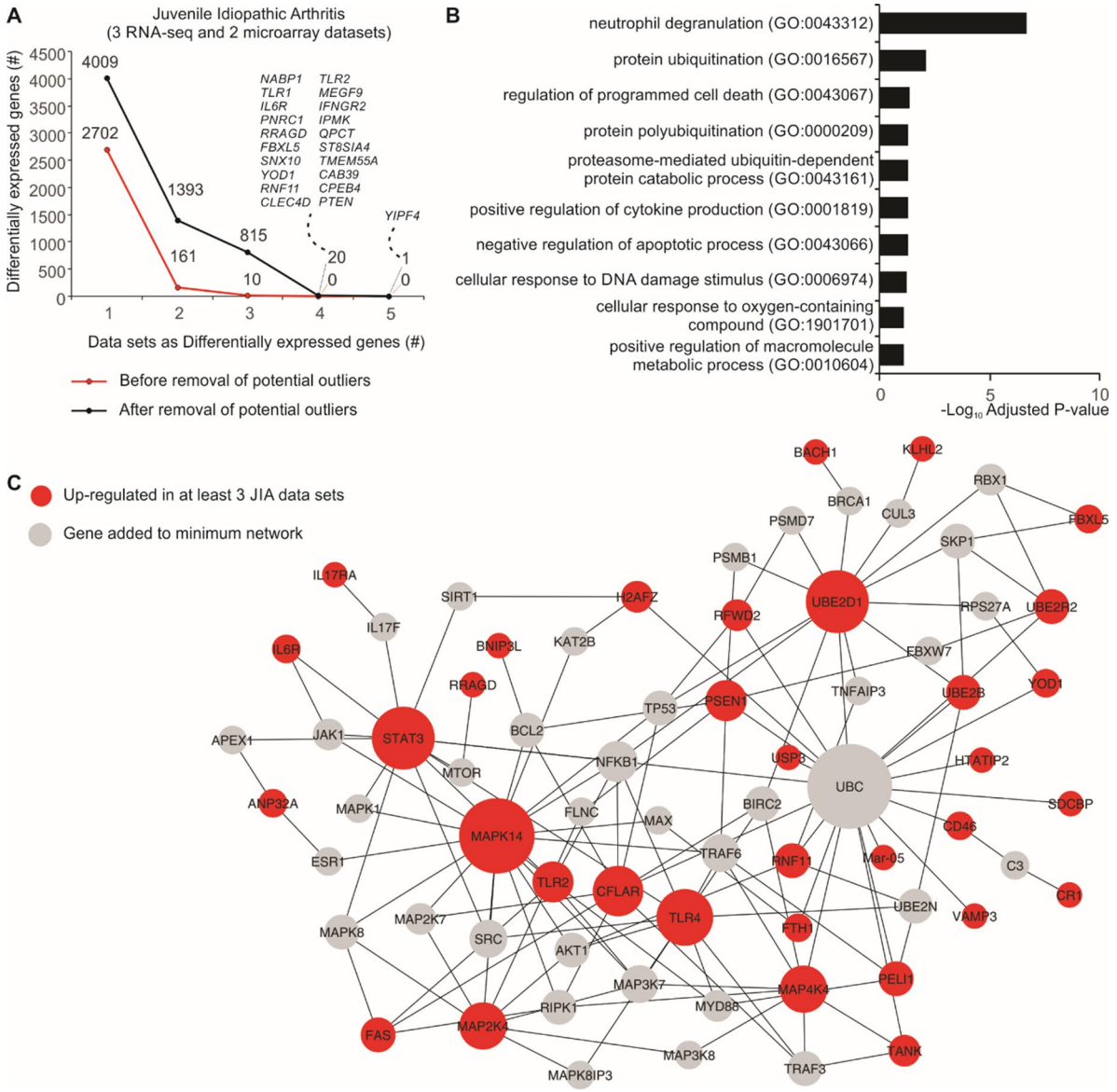


FIGURE 3 | Consistency of JIA signatures increases after removal of potential outlier samples. **(A)** Number of DEGs before and after removal of potential outlier samples in five JIA datasets. The lines show the number of genes (*y* axis) considered as DEGs in one or more JIA datasets (*x* axis). **(B)** Enrichment pathway analysis of genes consistently up- or down-regulated in three or more JIA datasets after removal of potential outlier samples. Bar graph shows the $\log_{10}$ adjusted *P* value (*x* axis) of top Gene Ontology gene sets (*y* axis). **(C)** Protein–protein interaction network showing the connectivity of up-regulated DEGs in at least three JIA datasets. Genes added to minimum network are shown as gray nodes. Edges were defined by InnateDB (Breuer et al., 2013).

samples in an otherwise uniform control group. Removing perturbed outliers could also potentially prove useful for finding disease classifiers by increasing the consistency of DEGs between similar studies. For single-cell analyses, it is not clear, however, how dropouts and cells with low MDP scores may impact the interpretation of the results since zero-inflated datasets may affect the calculation of MDP.

We observe that there is a great variation in the transcriptional profile of patients with different diseases. Part of this variability is due to the genetic contributions of each individual, as well as their prior infections, nutritional condition, stress, microbiota, and so on (Nakaya et al., 2012). There is still the possibility of hidden comorbidities in the diseased individuals, which were not part of the exclusion criteria of the clinical trials. The degree of molecular perturbation can provide a good indication of the health status of the individual and also identify the genes most perturbed by the disease in question.

Finally, the MDP approach can also be used to identify disease-associated perturbation in a priori–defined clinical or immunological factors (Bell et al., 2016; Pollara et al., 2017). In this way, the analysis can be used to split patients with the same disease into new subgroups with distinct gene expression profiles.

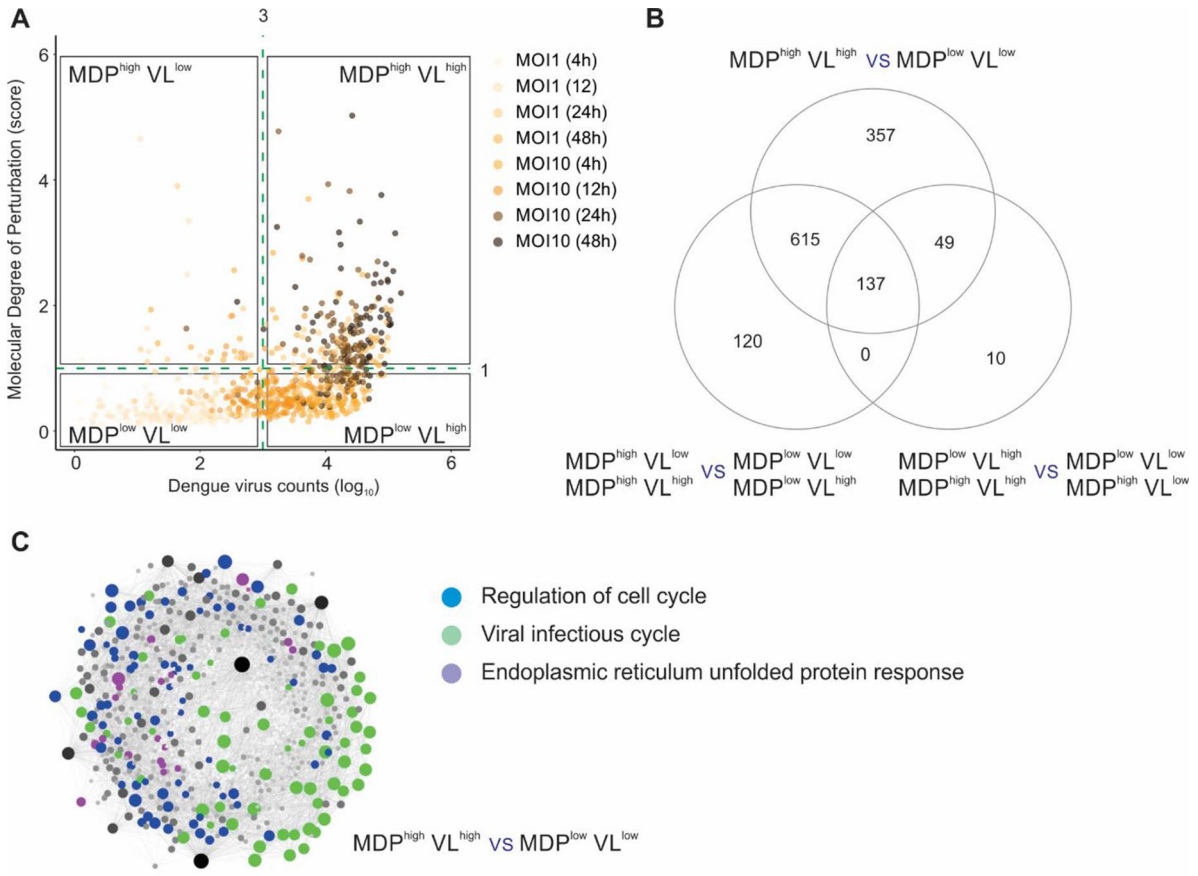


FIGURE 4 | Molecular degree of perturbation at single-cell level. **(A)** Sample MDP score (*y* axis) and DENV count (VL, *x* axis) for all single cells infected with DENV. The quadrants divide cells according to their MDP score (cutoff = 1) and/or the VL (cutoff = 3). Each circle represents a single cell, and the gradient color represents the different MOI (*m*ultiplicity *o*f *i*nfection) and time postinfection. **(B)** Differential expression analyses between cell subsets shown in A. The numbers of DEGs are indicated in the Venn diagram. **(C)** Protein–protein interaction network using the DEGs from $MDP^{high}VL^{high}$ versus $MDP^{low}VL^{low}$ comparison. Genes related to regulation of cell cycle (blue circles), viral infectious cycle (green circles), and endoplasmic reticulum unfolded protein response (purple circles) are shown.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

AG, ML, and HN performed the analyses, wrote the initial draft, and developed the tools. PR, AU, GP, and MN performed analyses. BG-C and VM-C implemented and help developed the webtool version. HN supervised the work. All authors wrote the final version of the manuscript.

FUNDING

This work was supported by grants from FAPESP (2012/19278-6, 2013/08216-2, 2018/14933-2), CNPq (313662/2017-7), FONDECYT-CONICYT (11161020), and PAI-CONICYT (PAI79170021). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001. MN and GP were supported by the Wellcome Trust and National Institute for Health Research Biomedical Research Centre at University College London Hospitals.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00971/full#supplementary-material

FIGURE S1 | Sample MDP scores of different human diseases. **(A)** Sample MDP scores of patients acutely infected with either virus or bacteria. Data were obtained from blood leukocytes and are available under GEO accession GSE6269. Healthy subjects (blue) were used as reference group. **(B)** Sample MDP scores of different types of cancer. Data were obtained from platelets and are available under GEO accession GSE68086. Healthy subjects (blue) were used as reference group. **(C)** Sample MDP scores of patients with inflammatory diseases. Data were obtained from whole blood and are available under GEO accession GSE112057. Healthy subjects (blue) were used as reference group.

FIGURE S2 | MDP calculated with specific gene modules. **(A)** Sample MDP score of patients with active TB (brown bars) and healthy controls (blue bars) using three different specific gene modules. Data were obtained from whole blood and are available under GEO accession GSE19435. **(B)** Sample MDP score calculated using all gene modules and for all TB datasets. The circles represent the difference between the median sample MDP score of patients with active TB and the healthy controls with no active TB within each study. The size and color of the circles are proportional to this difference. MΦ: macrophages.

TABLE S1 | Differential expression analysis with or without removal of potential sample outliers. The transcriptomic studies are shown as rows. StudyId = number of the study; GEOId = GEO accession ID with the type of disease; TotalControlSamples = number of samples in control group; TotalTreatedSamples = number of samples in disease group; TotalControlOutliers = number of samples in control group that were considered outlier by MDP; TotalTreatedOutliers = number of samples in disease group that were considered outlier by MDP; TotalOutliers = number of samples in total that were considered outlier by MDP; DEGsBefore = number of differentially expressed genes without removing any potential sample outlier (using samples in TotalControlSamples and TotalTreatedSamples); DEGsAfter = number of differentially expressed genes after removing potential sample outliers (using samples in TotalControlOutliers and TotalTreatedOutliers); DEGMin = minimum number of differentially expressed genes found after removing random samples (number of samples removed on each itineration is equivalent to the corresponding number in TotalOutliers) from TotalControlSamples and TotalTreatedSamples; DEGMax = maximum number of differentially expressed genes found after removing random samples (number of samples removed on each itineration is equivalent to the corresponding number in TotalOutliers) from TotalControlSamples and TotalTreatedSamples; DEGMean = average number of differentially expressed genes found after removing random samples (number of samples removed on each itineration is equivalent to the corresponding number in TotalOutliers) from TotalControlSamples and TotalTreatedSamples; AdjPcut = Adjusted P-value cutoff used on the differential expression analysis.

REFERENCES

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16 (4), 197–212. doi: 10.1038/nrg3891

Banchereau, J., Pascual, V., and O'Garra, A. (2012). From IL-2 to IL-37: the expanding spectrum of anti-inflammatory cytokines. *Nat. Immunol.* 13 (10), 925–931. doi: 10.1038/ni.2406

Bell, L. C., Pollara, G., Pascoe, M., Tomlinson, G. S., Lehloenya, R. J., Roe, J., et al. (2016). In vivo molecular dissection of the effects of HIV-1 in active tuberculosis. *PLoS Pathog.* 12 (3), e1005469. doi: 10.1371/journal.ppat.1005469

Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A., Oni, T., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466 (7309), 973–U998. doi: 10.1038/nature09247

Best, M. G., Verschueren, H., Post, E., Koster, J., Ylstra, B., Ameziane, N., et al. (2015). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* 28 (5), 666–676. doi: 10.1016/j.ccell.2015.09.018

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., et al. (2013). InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* 41 (Database issue), D1228–D1233. doi: 10.1093/nar/gks1147

Brown, R. A., Henderlight, M., Do, T., Yasin, S., Grom, A. A., DeLay M., et al. (2018). Neutrophils from children with systemic juvenile idiopathic arthritis exhibit persistent proinflammatory activation despite long-standing clinically inactive disease. *Front. Immunol.* 9, 2995. doi: 10.3389/fimmu.2018.02995

De Hertogh, B., De Meulder, B., Berger, F., Pierre, M., Bareke, E., Gaigneaux, A., et al. (2010). A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC Bioinf.* 11. doi: 10.1186/1471-2105-11-17

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30 (1), 207–210. doi: 10.1093/nar/30.1.207

Garg, N., and Smith, T. W. (2015). An update on immunopathogenesis, diagnosis, and treatment of multiple sclerosis. *Brain Behav.* 5 (9). doi: 10.1002/brb3.362

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20 (3), 307–315. doi: 10.1093/bioinformatics/btg405

Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nat. Rev. Genet.* 9 (8), 575–581. doi: 10.1038/nrg2383

Hersh, A. O., and Prahalad, S. (2015). Immunogenetics of juvenile idiopathic arthritis: a comprehensive review. *J. Autoimmun.* 64, 113–124. doi: 10.1016/j.jaut.2015.08.002

Jasenosky, L. D., Scriba, T. J., Hanekom, W. A., and Goldfeld, A. E. (2015). T cells and adaptive immunity to *Mycobacterium tuberculosis* in humans. *Immunol. Rev.* 264 (1), 74–87. doi: 10.1111/imr.12274

Jochems, S. P., Marcon, F., Carniel, B. F., Holloway, M., Mitsi, E., Smith, E., et al. (2018). Inflammation induced by influenza virus impairs human innate immune control of *Pneumococcus*. *Nat. Immunol.* 19 (12), 1299–1308. doi: 10.1038/s41590-018-0231-y

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25 (3), 415–416. doi: 10.1093/bioinformatics/btn647

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database C (2011). The sequence read archive. *Nucleic Acids Res.* 39 (Database issue), D19–D21. doi: 10.1093/nar/gkq1019

Mo, A., Marigorta, U. M., Arafat, D., Chan, L. H. K., Ponder L., Jang, S. R., et al. Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. *Genome Med.* 10 (1), 48. doi: 10.1186/s13073-018-0558-x

Nakaya, H. I., Gardner, J., Poo, Y. S., Major, L., Pulendran, B., and Suhrbier, A. (2012). Gene profiling of Chikungunya virus arthritis in a mouse model reveals significant overlap with rheumatoid arthritis. *Arthritis Rheum.* 64 (11), 3553–3563. doi: 10.1002/art.34631

Nakaya, H. I., Li, S., and Pulendran, B. (2012). Systems vaccinology: learning to compute the behavior of vaccine induced immunity. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 4 (2), 193–205. doi: 10.1002/wsbm.163

Pankla, R., Buddhisa, S., Berry, M., Blankenship, D. M., Bancroft, G. J., Banchereau, J., et al. (2009). Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis. *Genome Biol.* 10 (11), R127. doi: 10.1186/gb-2009-10-11-r127

Pollara, G., Murray, M. J., Heather, J. M., Byng-Maddick, R., Guppy, N., Ellis, M., et al. (2017). Validation of immune cell modules in multicellular transcriptomic data. *PLoS One* 12 (1), e0169271. doi: 10.1371/journal.pone.0169271

Prada-Medina, C. A., Fukutani, K. F., Pavan Kumar, N., Gil-Santana, L., Babu, S., Lichtenstein, F., et al. (2017). Systems immunology of diabetes-tuberculosis comorbidity reveals signatures of disease complications. *Sci. Rep.* 7 (1), 1999. doi: 10.1038/s41598-017-01767-4

Ramilo, O., Allman, W., Chung, W., Mejias, A., Ardura, M., Glaser, C., et al. (2007). Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 109 (5), 2066–2077. doi: 10.1182/blood-2006-02-002477

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi: 10.1101/gr.1239303

Wang, Y., Miller, D. J., and Clarke, R. (2008). Approaches to working in high-dimensional data spaces: gene expression microarrays. *Brit. J. Cancer* 98 (6), 1023–1028. doi: 10.1038/sj.bjc.6604207

Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J. C., Relman, D. A., et al. (2003). Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci. U. S. A.* 100 (4), 1896–1901. doi: 10.1073/pnas.252784499

Xia, J., Gill, E. E., and Hancock, R. E. (2015). NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* 10 (6), 823–844. doi: 10.1038/nprot.2015.052

Zanini, F., Pu, S. Y., Bekerman, E., Einav, S., and Quake, S. R. (2018). Single-cell transcriptional dynamics of flavivirus infection. *Elife* 7. doi: 10.7554/eLife.32942

Conflict of Interest: 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: 13 November 2019
doi: 10.3389/fgene.2019.01120

Leveraging User-Friendly Network Approaches to Extract Knowledge From High-Throughput *Omics* Datasets

Pablo Ivan Pereira Ramos[1], Luis Willian Pacheco Arge[2], Nicholas Costa Barroso Lima[3], Kiyoshi F. Fukutani[4] and Artur Trancoso L. de Queiroz[1]*

[1] Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil, [2] Laboratório de Genética Molecular e Biotecnologia Vegetal, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, [3] Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Ceará, Fortaleza, Brazil, [4] Multinational Organization Network Sponsoring Translational and Epidemiological Research (MONSTER) Initiative, Fundação José Silveira, Salvador, Brazil

OPEN ACCESS

Edited by:
Juilee Thakar, University of Rochester, United States

Reviewed by:
Monika Heiner, Brandenburg University of Technology Cottbus-Senftenberg, Germany
Paolo Tieri, Italian National Research Council (CNR), Italy

***Correspondence:**
Pablo Ivan Pereira Ramos
pablo.ramos@fiocruz.br

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

Received: *29 March 2019*
Accepted: *16 October 2019*
Published: *13 November 2019*

Citation:
Ramos PIP, Arge LWP, Lima NCB, Fukutani KF and de Queiroz ATL (2019) Leveraging User-Friendly Network Approaches to Extract Knowledge From High-Throughput Omics Datasets. Front. Genet. 10:1120. doi: 10.3389/fgene.2019.01120

Recent technological advances for the acquisition of multi-*omics* data have allowed an unprecedented understanding of the complex intricacies of biological systems. In parallel, a myriad of computational analysis techniques and bioinformatics tools have been developed, with many efforts directed towards the creation and interpretation of networks from this data. In this review, we begin by examining key network concepts and terminology. Then, computational tools that allow for their construction and analysis from high-throughput *omics* datasets are presented. We focus on the study of functional relationships such as co-expression, protein–protein interactions, and regulatory interactions that are particularly amenable to modeling using the framework of networks. We envisage that many potential users of these analytical strategies may not be completely literate in programming languages and code adaptation, and for this reason, emphasis is given to tools' user-friendliness, including plugins for the widely adopted Cytoscape software, an open-source, cross-platform tool for network analysis, visualization, and data integration.

Keywords: correlation networks, graph, high-throughput sequencing, network analysis, omics, protein–protein interaction, regulatory networks, systems biology

INTRODUCTION

The analysis of high-throughput datasets using the framework of networks has gained widespread adoption in the biological sciences. With approaches in this field shifting from a mostly reductionist perspective towards a more holistic view of natural phenomena (Barabási and Oltvai, 2004; Berlin et al., 2017), the analytical tools used to extract knowledge from data have also adapted. The vocabulary of networks is particularly suitable for studying problems that explicitly focus on the *relationships* among elements, where the latter can be any entity under study, including but not limited to genes, transcripts, proteins, or metabolites. With sheer amounts of data that can be obtained from instruments such as high-throughput sequencers, analytical strategies that permit broader insights of the functional roles of each element are warranted, and this can be achieved by the use of network approaches.

In this Review, we focus on the various uses of network methods to the analysis of large-scale *omics* datasets, which are those generated using medium- and high-throughput technologies in genomics, transcriptomics, proteomics, and metabolomics experiments. First, key concepts and terminology of this area are presented, followed by the introduction of biological network variants, namely correlation networks (*Correlation networks allow disclosing of relevant associations in omics datasets*), gene regulatory networks (GRNs) (*Gene regulatory networks permit an improved understanding of the cell's transcriptional circuitry*), and protein–protein interaction (PPI) networks (*Protein–protein interaction networks provide an integrated view of the proteome's organization and interactions*). Methods to perform key analysis in a network are presented in *A primer on network analysis and visualization*. With every approach, computational tools that we considered both appropriate and user-friendly are presented. User-friendly tools were defined as those that provide a point-and-click graphical user interface, which does not mean that they have limited functionality or that they are only used by those without extensive programming literacy. Rather, they can be used to complement analyses performed in different environments, such as R or Python scripts, and usually offer improved layouts and visualization modes compared to less friendly alternatives. Our Review differs from that of others who have engaged in similar challenges (for instance, the works of Aittokallio and Schwikowski, 2006; Stevens et al., 2014; Huang et al., 2017), since we primarily target the non-programmer who wants to apply network methods to a dataset of interest. Luckily, network analysis is an area that has greatly benefited from the existence of excellent analysis software such as Cytoscape (Shannon et al., 2003) (https://cytoscape.org/), Gephi (Bastian et al., 2009) (https://gephi.org), and NAViGaTOR (Brown et al., 2009), to name a few. Gephi and Cytoscape, in particular, can be extended by the many plugins created by third-party developers and available in official repositories (Saito et al., 2012), and these were at the heart of the current review. While the aforementioned types of networks are widely employed, there are many other applications that are not in the scope of this work. As an example, the modeling of (bio)chemical networks using graph–theoretic approaches have advanced our understanding of bacterial and eukaryotic metabolism (Klein et al., 2012; Dutta et al., 2014; Jha et al., 2015), and were the object of previous reviews (see, e.g., Lacroix et al., 2008; Cottret and Jourdan, 2010). Biology and Biomedicine are, indeed, areas which have been greatly benefited by the use of network techniques resulting from cross-pollination among disciplines.

Beyond the Empirical, Towards Formalism: What Are Networks?

Network is a general term used in many different contexts: social networks, traffic networks, ecological networks, computer networks, among others, all share a common theme related to the interaction among a set of disparate elements, viz. people, vehicles, species, and computers. The topology of networks and the interactions within can be formally studied from a graph–theoretic viewpoint, which allows for a mathematical representation and formalism, while also facilitating visualization of the network. Since several distinct graph representations exist, for generality we will focus on the description of simpler types of graphs. In general, a graph $\Gamma = (V,E)$ is composed of a finite set V of nodes (or vertices), and E of (directed or undirected) edges (or links). In the case of *omics* datasets, each node $v \in V$ could represent a (bio)chemical entity such as a gene, transcript, protein, or metabolite, and an edge $e = \{v_1,v_2\} \in E$ exists between two nodes when there is evidence for their interaction, which in turn depends on the specific aim of the modeled network, which guides the definition of interaction. For instance, in the simplest type of correlation network, one could specify a hard threshold over all pairwise values of Pearson's correlation coefficients in order to determine whether any two nodes are connected. On the other hand, in a PPI network, edges between protein nodes exist when evidence for their physical interaction is available, which could be obtained by a wealth of techniques that include co-immunoprecipitation, affinity purification, proteomics, and computational approaches (Ngounou Wetie et al., 2014).

The edges in a graph can be undirected (**Figure 1A**) or directed (**Figures 1B**, **C**). In directed graphs, there is a specific sense pointing at the direction of a given interaction, such as a transcription factor (TF) that regulates a given gene in a regulatory network (a causal relationship), while undirected graphs describe two-way associations such as the co-expression of genes in a correlation network, in which a significant correlation *per se* does not provide sufficient evidence to infer whether any of the compared genes regulates or is being regulated by the other, or even by an upstream regulator acting on both simultaneously. That is, correlation does not imply causation, and hence the undirected graph is a more appropriate representation of this relationship.

Graphs can also have numerical weights associated with each interaction, the interpretation of which depends on the specific application under study (**Figure 1C**). In a correlation network, for instance, weights could represent the magnitude of the correlation statistic. Also possible is to set weights based on the confidence of the interaction as measured by a relevant parameter. As an example, the STRING database (http://string-db.org), which harbors information on physical and functional PPIs, quantifies interaction weights between proteins as a combined score dependent on the nature (experimental or computational prediction) and quality of the supporting evidence (Szklarczyk et al., 2017). **Table 1** summarizes the biological interpretation of nodes, edges, and edge weights for the three types of networks considered in this study. While these interpretations are typical for these kinds of biological networks, studies may employ different analytical strategies that lead to variations on how to account edge directionality or weights, for instance. As an example, regulatory networks are usually inferred using a bipartite graph representation, where nodes are of two different types (either a TF or a target gene). In this case, edge directionality characterizes an underlying regulatory event (activation or inhibition) of a TF towards a target gene, hence these networks are usually modeled as a directed graph (Narasimhan et al., 2009; Song et al., 2017).

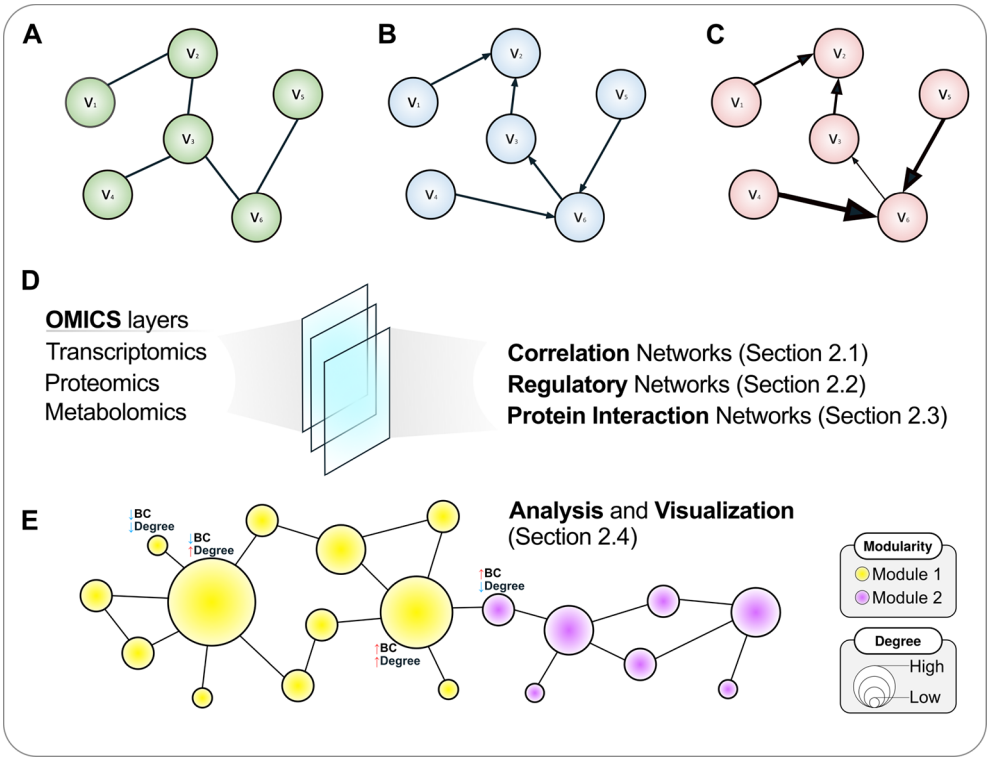


FIGURE 1 | A roadmap to network concepts covered in this review. Three simple six-node graphs are shown in the upper panel. These graphs can be undirected **(A)**, directed **(B)** or weighted directed **(C)**. In the latter, the thickness of edges reflects the weights of the interactions. Various *omics* datasets can be analyzed using the language of networks, which are discussed in the following sections **(D)**. **(E)** Once a network is attained, further analyses are warranted, which include disclosing modules or communities and calculating topological metrics such as node degree and betweenness centrality (BC), covered in *A primer on network analysis and visualization*. The size of a node is proportional to its degree, while the color reflects the community structure in this illustrative example where two modules are disclosed. For selected nodes, interpretations of node BC and degree are presented.

TABLE 1 | Biological interpretation of nodes, edges, and edge weights for the *omics*-derived networks under study.

Type of network	Graph representation	Edge directionality	Biological interpretation of		
			nodes	edges	edge weights
Correlation network	Simple graph	Undirected	Genes, proteins, or metabolites	Correlation (co-expression) between a pair of biological entities, which is calculated from a measure of abundance, such as gene expression or metabolite concentration	The strength of correlation (co-expression) between the pair of nodes
Gene regulatory network	Simple or bipartite graphs	Usually directed	Genes in the simple graph; transcription factors and target-genes in the bipartite graph	A regulatory relationship	The degree of the regulatory relationship
Protein–protein interaction network	Simple graph	Usually undirected	Proteins	The direct contact (physical binding) between proteins, but can represent indirect (functional) interactions between the peptides	Usually unweighted, but can be valued to represent the support (confidence) for a given interaction

HOW TO DISCLOSE NETWORKS FROM HIGH-THROUGHPUT *OMICS* DATASETS

In the following sections, we review and discuss methods to construct various types of networks using a wealth of *omics* datasets as input (**Figure 1D**). While many different computational methodologies to achieve the construction of a network exist, we focus on those that we considered more apt for users without a computational background, especially those that are based on plugins for the popular software Cytoscape (Shannon et al., 2003), which allows visualization, rendering, and analysis of networks in the same computational environment, with the

advantage of being open-source, platform-independent, and continuously updated. Once the tools to build these biological networks are covered, we shift our focus towards analysis and visualization aspects of graphs, which are presented in *A Primer on Network Analysis and Visualization* (**Figure 1E**).

Correlation Networks Allow Disclosing of Relevant Associations in *Omics* Datasets

Recent advances in high-throughput technologies have increased our capacity to assess the elements in different *omics* layers, allowing their simultaneous treatment in single grouped mechanisms that together explain biological events (Carpenter and Sabatini, 2004; Vella et al., 2017). In this sense, the processes that allow for life maintenance in cells can be regarded as an intricate web of complex relationships between molecules such as proteins, lipids, metabolites, and nucleic acids (RNA and DNA) (Barabási et al., 2011). Correlations are arguably the dominant way to infer relationships not only between the elements in these distinct layers of information but also within each layer, as it allows simultaneously examining the associations that drive an observed biological effect, and there are several ways of calculating correlation coefficients. Statistically, the correlation is a measure of the two-way linear association between a pair of variables (Mukaka, 2012). The correlation coefficient permits estimating the degree or strength of this association. The most common and classic correlation statistic is the Pearson's correlation coefficient (or r), which measures linear associations between two variables under the assumption that the data be normally distributed and that observations are independent (Walter and Altman, 1992). Non-parametric methods based on ranks avoid the assumption of normality and are preferred when the data is ordinal, skewed, or presents extreme values (outliers). One such method is the Spearman correlation coefficient, which is a calculation of Pearson's correlation coefficient on the ranks of the observations, rather than on the raw data, and yields an r_S statistic (also called ρ, rho). The Kendall rank correlation coefficient (also called τ, tau) uses the number of concordant and discordant rank pairs to evaluate association. The biweight midcorrelation is less prone to outlier influence because it is a median-based estimation and, like the two previous, yields a robust measurement of association, with the drawback that few tools are available that calculate this metric (Langfelder and Horvath, 2012). Correlation coefficients (r, r_s, ρ, or τ) are a dimensionless quantity ranging from -1 to 1, where values close to zero indicate no (linear) association whilst values equal to or near 1 (or -1) indicate strong, positive (or negative) correlations, although absolute values as low as 0.3 can already be considered a weak correlation depending on the context (Mukaka, 2012).

Since the relationships between genes, proteins, metabolites and biological entities in general are complex and often nonlinear, while having distributions that can be non-normal, alternative measurements of association are often required (Hardin et al., 2007), and include information-theoretical measures such as mutual information (MI). MI quantifies the dependence between a pair of random variables and, based on the concept of entropy, estimates how much knowledge is gained about a variable (say, expression values of a gene X) by observing a second variable (say, expression values of a gene Y), hence its name. The MI is zero when the variables are statistically independent, while a positive value denotes a degree of dependence (Steuer et al., 2002). In a scenario of statistical independence, the distribution of values of variable X is not altered at all when those of variable Y changes. It is worth noting that traditional association measures that disclose only linear relationships are insufficient to reveal statistical independence, exactly because there can be non-linear relationships in the data that these methods do not adequately capture. We refer the reader to the review of de Siqueira Santos et al. (2014) on statistical dependency identification, who further provide illustrative biological examples and simulations using various association statistics.

Correlations can be visually assessed by plotting the data as a scatter plot fitted by a line, where the further the data lie from the straight line, the weaker the correlation (**Figure 2A**). While this approach is feasible when few variables are compared, it has limited practicality when dealing with large-scale *omics* datasets, such as high-throughput expression profiling and proteomics. In these cases, methods that create correlation networks are preferred (Zhang and Horvath, 2005; Langfelder and Horvath, 2008; Vella et al., 2017). Once a correlation (or other association statistic) matrix is attained (**Figure 2B**), a network can be inferred (**Figure 2C**). A co-expression network is a particular case of correlation network constructed using genome-wide expression data, although the term is sometimes used to refer to networks created by correlating the abundance of protein or metabolites in proteomics and metabolomics studies. In this network, the nodes

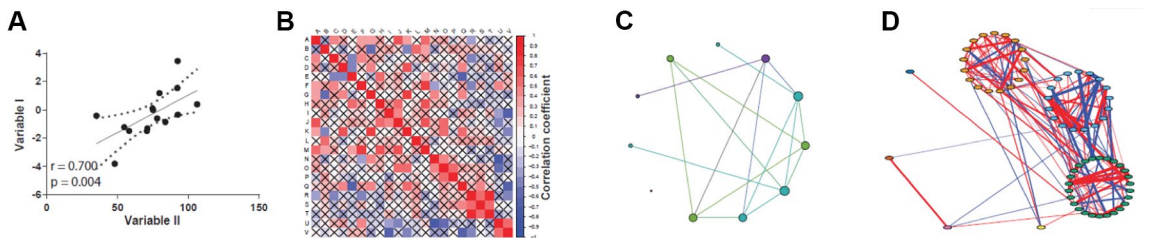


FIGURE 2 | Different views on assessing correlations. **(A)** Classic scatter plot with correlation curve (straight black line). **(B)** Correlation matrix plot, designed with the *corrplot* package (Wei and Simko, 2017). **(C)** Circular layout correlation network, designed with Gephi (Bastian et al., 2009). **(D)** Complex correlation network with modularity coloring, designed with *qgraph* package (Espkamp et al., 2012).

are elements such as genes, proteins, or metabolites, and an undirected edge connects a pair of nodes if the correlation statistic between them exceeds a given threshold (**Figure 2C**). This "hard-threshold" approach represents the simplest form of inducing a network from *omics* data, and is limited by the arbitrary nature of the threshold used, which will dismiss slightly undervalued correlations that could be potentially relevant. An alternative, more sophisticated approach to disclose co-expression networks is by using soft-thresholding approaches, of which the weighted gene co-expression network analysis (WGCNA) algorithm is among the most widely employed methods (Langfelder and Horvath, 2008). The main advantage of the WGCNA approach is that no arbitrary thresholding on the correlation values is enforced, which effectively preserves the continuous nature of the correlation distribution. In addition, it is not impacted by the arbitrariness of hard-thresholding methods. In WGCNA, once all pairwise correlations are calculated, an adjacency matrix, which holds information on edge strengths, is obtained by applying a power transformation of the form $f(x) = x^{\beta}$, where x are correlation values and β is the soft-thresholding parameter, a positive value set by the user such that the resulting network presents an approximately scale-free property while maintaining high connectivity (see **Box 1** for a primer of important network definitions). As a result, high correlations are emphasized at the expense of low correlations, but without the need of setting an explicit threshold on the correlation values themselves.

User-Friendly Tools for Constructing Correlation Networks

Gene/protein correlation network analysis can be performed using in-house scripts and packages for general-purpose programming languages such as R, Python, Perl, or Java. However, alternatives exist for the bioinformatics user that wants to apply such methods to their data in the absence of a solid computational background (**Table 2**). One of them is based on the Cytoscape environment, which also allows for installing third-party plugins. A specific app developed for correlation network analysis, the *ExpressionCorrelation* app (available at http://apps.cytoscape.org/apps/expressioncorrelation), presents a Pearson's correlation-based solution. Thus, a table of gene/protein/metabolites measurements is the input and Cytoscape can generate the gene and sample correlation network. This plugin has been applied to the construction of many networks, exemplified by an *Anopheles* gene co-expression network (Shrinet et al., 2014), a correlation network from *Aspergillus* metabolites highlighting those significantly associated to anticancer and antitrypanosomal bioactivity (Tawfike et al., 2019), and co-expression networks from cancer datasets (Wang et al., 2016b; Zhang et al., 2016). Pearson's correlation statistic, however, presents several limitations as pointed out in the previous section. The Cyni toolbox app circumvents this difficulty by allowing calculation of rank-based correlations such as Spearman's and Kendall's, in addition to Pearson's coefficient (Guitart-Pla et al., 2015). **Figure 3** shows a bacterial co-expression network constructed using Cyni.

Another user-friendly solution is *geWorkbench* (Floratos et al., 2010). This tool is an open source Java desktop application that allows correlation using an ARACNe (mutual information-based) implementation (Margolin et al., 2006a), and is particularly suitable for finding regulatory networks from transcriptomic data. In addition, the workbench allows for parameter estimation and is fairly flexible for user customization. Its advantages over the Cytoscape *ExpressionCorrelation* app include the possibility of p-value threshold modification and correction, as well as bootstrap resampling. Thus, the program permits evaluating the statistical significance of the network and keep the more robust associations. However, the user-friendly advantage is not without its costs: the plugin is limited to the calculation of regular correlations (Pearson's and Spearman's) and mutual information. Also, the use of more robust correlation statistics, such as the biweight midcorrelation, still requires proficiency in programming languages/R packages, since so far there are no alternatives that incorporate this measure.

The construction of weighted networks using the soft-thresholding approach employed by WGCNA requires the execution of a multi-step pipeline implemented as an R package (Langfelder and Horvath, 2008), thus requiring programming skills to correctly adapt and parametrize the functions and the dataset itself. To circumvent this need, a webserver adaptation of the WGCNA method was recently published as *webCEMiTool*, allowing an user-friendly approach to disclose a weighted co-expression network, detect modules therein, and produce publication-quality visualizations (https://cemitool.sysbio.tools/) (Cardozo et al., 2019). In this context, modules are considered as groups of genes with similar expression profiles, which tend to have related biological functions or be under the influence of the same transcriptional regulator, but a more ample discussion of modularity is presented in *A primer on network analysis and visualization*. *webCEMiTool* also has a built-in method to automatically select the optimal value of β (the soft-thresholding parameter), which is described elsewhere (Russo et al., 2018) and, like the original WGCNA algorithm, it could also be used to disclose correlation networks from proteomics or metabolomics datasets. Pathway enrichment analysis can be run directly from the *webCEMiTool* application, as it interfaces with the Enrichr platform (Kuleshov et al., 2016) which comprises over a hundred gene set libraries, thus facilitating the interpretation and extraction of knowledge from the inferred network.

Gene Regulatory Networks Permit an Improved Understanding of the Cell's Transcriptional Circuitry

Gene (transcriptional) regulatory networks, or GRNs, are models that aim at the elucidation of genetic information processing, aiding on the understanding of organism development. A GRN is based on the following elements: TFs, target genes, and their regulatory elements in the upstream region. TFs are identified using computational tools based on sequence homology and through motif conservation across TF families. Each TF can act on the transcription of multiple genes. In the upstream region of each target gene, there exist elements/motifs that are recognized by the TF, and the gene is subsequently transcribed. When located upstream of a gene, these motifs are called *cis*-elements.

BOX 1 | Key concepts applied to biological networks

Biological networks are composed of nodes that can represent different bioentities and have different biological importance for a given network. Regardless of the network size, shared commonalities exist between different biological networks, which allow their comparison. The concepts below describe some characteristics of biological networks and different metrics for topological evaluation of nodes, allowing for prioritization of important elements in the network.

Scale-free. A network is considered scale-free when its degree distribution follows a power law. Thus, it is characterized by the presence of many small-degree nodes together with a few highly connected nodes (or hubs), forming an inhomogeneous network. Many biological networks exhibit the scale-free property, including protein interaction and gene co-expression networks.

Small-world. When networks exhibit a low number of node intermediates separating any two nodes in the network (*ie.*, low average distance), it is considered a small-world network.

Modularity. Biological networks tend to form modules, or clusters of highly connected nodes (**Figure box A**). Modularity takes values between -1 and 1 and reflects the link density within a module as compared to links between

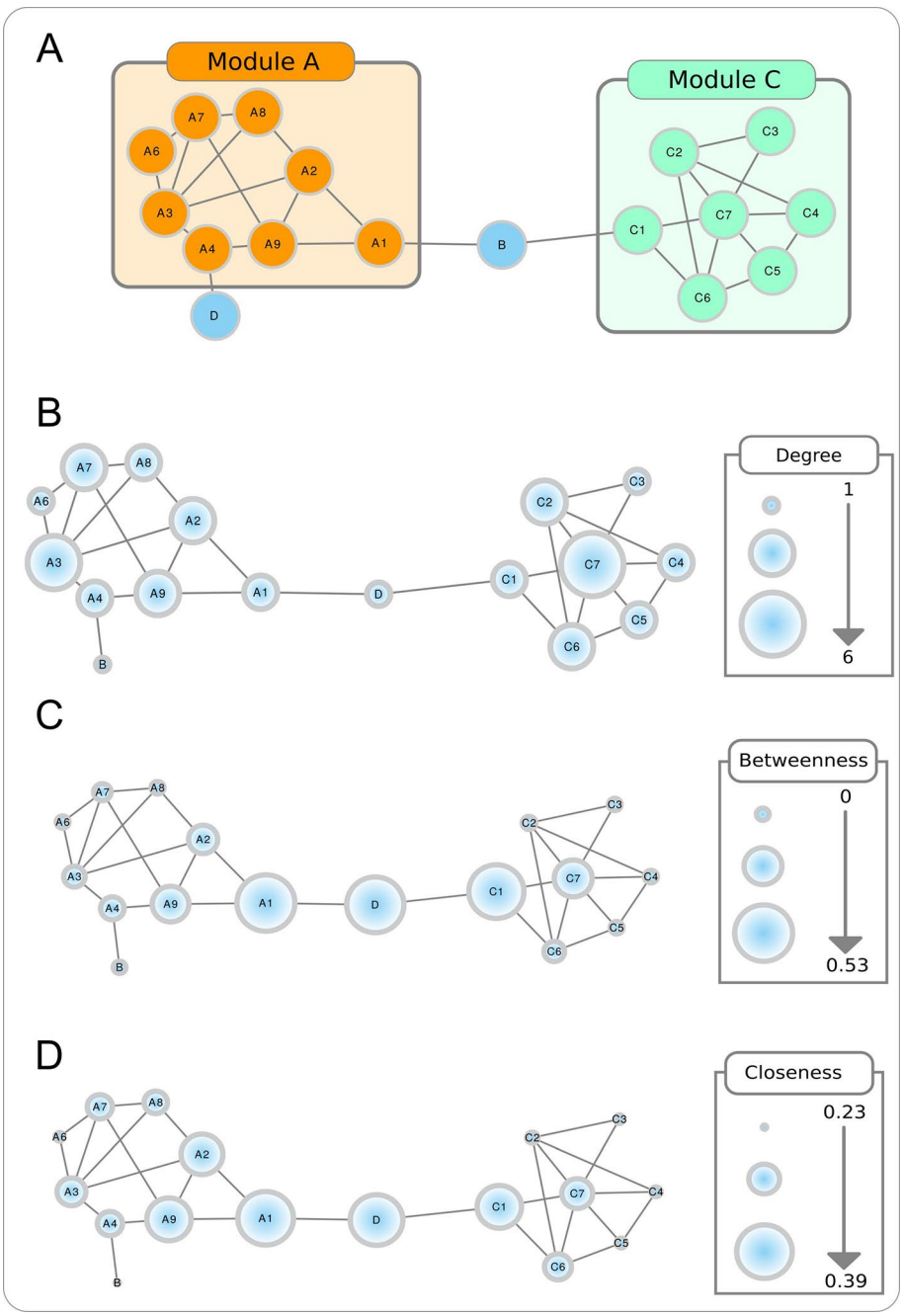


FIGURE BOX | Topological properties of a toy network. The modular aspect of the network is apparent in **A**, with two modules (or partitions) shown. The size of the nodes in **B**–**D** are proportional to, respectively, the node degree, betweenness centrality, and closeness centrality.

BOX 1 | Continued

modules. In biological networks, nodes with similar functions have a bias to form functional modules.

Hubs. The most highly linked nodes in a network are called hub nodes, which play an important role in defining network scale-freeness. The term is also used to refer to nodes that display high centrality as measured using a relevant metric (see below).

Shortest (or geodesic) path. A shortest path is the minimum series of edges that should be traversed to connect two nodes in a network. In a weighted graph, it is the path lending to the minimum sum of edge weights between a node pair.

Node centrality metrics

Each component of a network presents topological characteristics that can be translated into biological knowledge and help establish the identification of relevant nodes:

Node degree. Refers to the number of nodes directly connected to a specific node, and is obtained by counting the number of interactions that a specific node has with other nodes in the network (**Figure box B**). When the network is directed, this is separated into out-degree (the number of outgoing links from a node) and in-degree (the number of ingoing links in a node). The higher is the degree of a node, the higher will be the probability that it is a hub. Nodes with high degree centrality have more influence on the structure and functionality of a network than nodes with a low degree.

Betweenness centrality. Measures the importance of a node to the connection of different parts of a network (**Figure box C**). The betweenness centrality for a node is the proportion, among all shortest paths, of those that use the given node as intermediate. Nodes with these characteristics are usually referred as bottlenecks and can also be considered hubs.

Closeness centrality. Measures how close a node is to all the other nodes in the network (**Figure box D**). It is calculated by the reciprocal sum of all shortest paths to all other nodes of the network. The higher the closeness centrality for a node, the closer is the relationship with the remaining nodes in the network.

TABLE 2 | User-friendly computational tools for inferring correlation networks.

Tool	Description	Platform	Reference/URL
Cyni toolbox (Cytoscape)	Performs several correlation analyses and includes other networks inference algorithms.	Multi	http://apps.cytoscape.org/apps/cynitoolbox; (Guitart-Pla et al., 2015)
Expression Correlation app (Cytoscape)	Performs Pearson correlation analysis and network inference.	Multi	http://apps.cytoscape.org/apps/expressioncorrelation
ARACNe/Mutual Information (geWorkbench)	Creates a network based on Mutual Information.	Multi	http://wiki.c2b2.columbia.edu/workbench/index.php/Home; (Floratos et al., 2010)
webCEMiTool	Performs comprehensive modular analyses in a fully automated manner, generating co-expression networks based on the WGCNA method.	Webserver	https://cemitool.sysbio.tools/; (Cardozo et al., 2019)

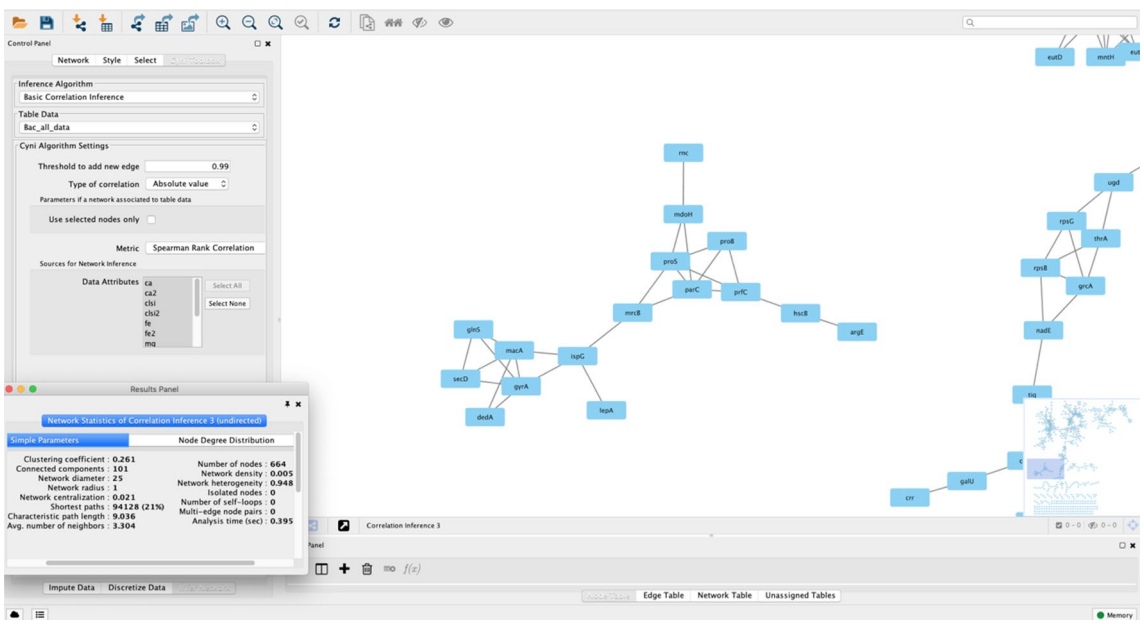


FIGURE 3 | A correlation network constructed using Cytoscape 3.2. The network was built using a bacterial expression dataset, and nodes represent annotated genes, with edges connecting nodes if they pass a correlation threshold calculated using Spearman's rank correlation in the Cyni Toolbox. In the picture a pop-up menu with the calculated network metrics (using the NetworkAnalyzer plugin in Cytoscape) is shown. Besides the network zoom, the program also shows the whole network in the lower-right screen, as a miniature.

Identification of *cis*-elements can be performed by biological experiments, such as by chromatin immunoprecipitation (ChIP)-seq methodology (Lee et al., 2006), or computationally by alignment of known motifs or by the identification of novel motifs. The latter are called *de novo* approaches and employ mathematical structures such as hidden Markov models (HMM) (Bailey et al., 2009). Typically, after the identification or discovery of new *cis*-elements, an enrichment analysis is performed using Fisher's exact test for identification of enriched motifs in the set of upstream regions from target genes.

On the other hand, the prediction of TFs-target genes interactions can be performed using a reverse engineering-based strategy. The top-down approach is particularly suitable in this context and uses information from gene expression datasets to detect expression patterns and then induce a GRN (Hartemink, 2005; Hache et al., 2009). The first models used to infer GRNs were based on the Pearson correlation coefficient but failed to capture non-linear pattern dependencies (as previously addressed). Other approaches were subsequently developed and applied to disclose GRNs in a more robust way, and included regression (Huynh-Thu et al., 2010), mutual information (Margolin et al., 2006a), partial correlations (Wille et al., 2004), and variations of these (Luo et al., 2008; Meyer et al., 2008). Despite each method having its peculiarities, GRNs inferred by diverse techniques usually do not present large differences (de Matos Simoes et al., 2013), and bootstrap analysis could be used to infer more robust GRNs. Another difficulty is the existence of regulation patterns that occur in rare conditions and cannot be easily detected, requiring specific wet-lab experiments for this purpose.

The study of gene regulation can take two main paths: i) GRN inference and ii) dynamic modeling, which can be performed either in isolation or in conjunction. We focused on methods that accomplish the first goal, while the latter can be attained using a diverse array of techniques that include Boolean formalism (logical models), Bayesian dynamic networks, and Ordinary Differential Equations (studied elsewhere, e.g., Kaderali and Radde 2008; Naldi et al. 2009; and Chai et al. 2014). The representation of inferred GRNs can be in the form of bipartite graphs which, in contrast to the simple graphs presented in the Introduction and in the construction of co-expression networks, have nodes of two types: TFs or target genes, and edges between them indicate a regulatory interaction (**Table 1**, **Figure 4A**). This type of representation is usually employed to GRNs originated from co-expression relationships because usually no *a priori* information is available about the type of regulation that the TF exerts on the target genes. Logical models, on the other hand, incorporate prior information on gene activation and repression, and the modeling of these relationships permit the capturing of the global dynamic behavior of the regulatory network in a simple fashion. An example of such a network from the human GRN, available in Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRRUST) database, is shown in **Figure 4B**.

User-Friendly Tools for Constructing Gene Regulatory Networks

As seen above, construction of GRNs is based on interaction inference between TFs and target genes, and on the identification of *cis*-elements in the upstream region of target genes. Next, we present user-friendly tools to perform both steps. GRNs inferred based on gene expression patterns are considered of intermediate value because they require improvement and validation with biological experiments. Traditionally, the inference of GRNs has been performed with tools based on command-line or in the R programming language such as ARACNe (Margolin et al., 2006a), but current alternatives include more user-friendly approaches which are listed in **Table 3**. These include

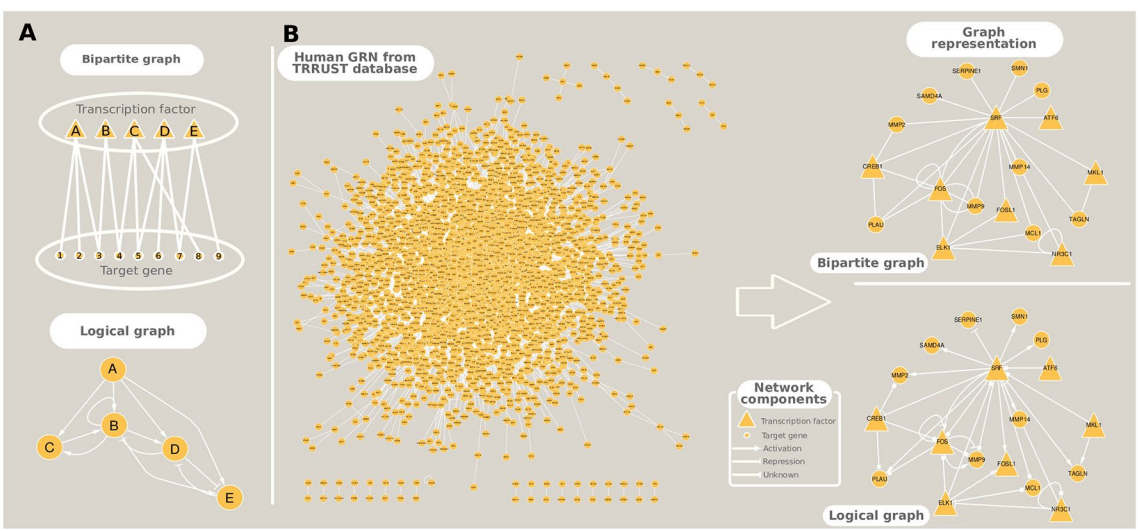


FIGURE 4 | Different ways to represent gene regulatory networks. **(A)** Toy networks exemplifying bipartite and logical (Boolean) graphs. **(B)** A real example of the human gene regulatory network extracted from TRRUST database, and its graphical representation as a bipartite and a logical networks.

TABLE 3 | User-friendly computational tools for inferring gene regulatory networks.

Tool	Description	Platform	Type of data		Reference/URL
			Expression	Promoter	
ARACNe	Creates a network based on Mutual Information	Multi	✓		http://apps.cytoscape.org/apps/aracne; (Floratos et al., 2010)
CyGenexpi	A toolset for identifying regulons and validating gene regulatory networks using time-course expression data	Multi	✓		https://apps.cytoscape.org/apps/cygenexpi; (Modrák and Vohradský, 2018)
CyNetworkBMA	Infers gene regulatory networks from expression measurements using Bayesian Model Averaging	Multi	✓		https://apps.cytoscape.org/apps/cynetworkbma; (Fronczuk et al., 2015)
GRNCOP2	Model-free combinatorial optimization algorithm to infer time-delayed gene regulatory networks from genome-wide time series datasets	Multi	✓		https://apps.cytoscape.org/apps/grncop2; (Gallo et al., 2011)
iRegulon	Allows identification of regulons using motif and track discovery in an existing network	Multi		✓	https://apps.cytoscape.org/apps/iregulon; (Janky et al., 2014)
NetworkAnalyst	Allows establishing TF-target genes and miRNAs-target genes associations.	Webserver	✓		http://www.networkanalyst.ca; (Zhou et al., 2019)
TRRUST	TFs and target genes interactions, and TFs cis-regulatory elements	Webserver	✓	✓	https://www.grnpedia.org/trrust/Network_search_form.php; (Han et al., 2018)
RegNetwork	Genic regulations by TFs and microRNAs	Webserver	✓		http://www.regnetworkweb.org/search.jsp; (Liu et al., 2015)
ORegAnno	Regulatory regions, transcription factor binding sites, etc.	Webserver		✓	http://www.oreganno.org/; (Lesurf et al., 2016)
rSNPBase	Harbors curated information on regulatory SNPs	Webserver		✓	http://rsnp.psych.ac.cn/; (Guo and Wang, 2018)
MEME	Sequence analysis tools for motifs discovery	Webserver		✓	http://meme-suite.org/ (Bailey et al., 2009)

an ARACNe implementation in *geWorkbench*, which was listed previously in the correlation network section, and also available are the Cytoscape plugins CyGenexpi (Modrák and Vohradský, 2018), CyNetworkBMA (Fronczuk et al., 2015), GRNCOP2 (Gallo et al., 2011), and iRegulon (Janky et al., 2014) (**Table 3**).

The ARACNe package is based on mutual information index to establish interactions between a pair of genes, such as a TF and a target gene; moreover, this tool employs bootstrapping to generate a consensus and robust network (Margolin et al., 2006b). CyGenexpi is based on an ordinary differential equation model applied on time series data that together with static binding (e.g., ChIP-seq) or information obtained from the literature allows inferring of gene regulatory modules in bacteria (Modrák and Vohradský, 2018). CyNetworkBMA employs a Bayesian model averaging algorithm to infer GRNs with a user-friendly interface and executes network processing on top of R code, which accelerates the inference process by allowing parallel processing (Fronczuk et al., 2015). Additionally, CyNetworkBMA can compute some statistics for the network evaluation, including receiver operating characteristic and precision-recall curves. The package GRNCOP2 has an algorithm based on machine learning with a model-free combinatorial optimization to infer time-delayed GRNs from genome-wide time series datasets (Gallo et al., 2011). The GRNs inference from the iRegulon package is based on analysis of *cis*-regulatory sequences from target genes and performs a genome-wide ranking-and-recovery strategy to detect enriched motifs related to TFs and their optimal sets of direct targets (Janky et al., 2014).

Like other types of biological data, GRNs can be stored on public databases which can be queried by other scientists. In this context, databases that permit storing and downloading of GRNs include TRRUST (Han et al., 2018), RegNetwork (Liu et al., 2015), ORegAnno (Lesurf et al., 2016), and rSNPBase (Guo and Wang, 2018) (**Table 3**). TRRUST database contains information obtained by computational mining and curated TFs-target genes interactions, and about TFs *cis*-regulatory elements in human and mouse. RegNetwork contains information of genic regulations by TFs and microRNAs, also in human and mouse. Similarly, NetworkAnalyst is a webserver that offers an integrated environment to establish TF-target gene and miRNA-target gene interactions (with data sourced from TarBase and miRTarBase). It works by mapping significant genes (such as those found differentially expressed in an RNA-seq experiment) to the corresponding molecular interaction database, and the resulting network can be exported to a Cytoscape-friendly input format. ORegAnno contains information about regulatory regions, TF binding sites, RNA binding sites, regulatory variants, haplotypes, and other regulatory elements for 18 species. Finally, rSNPBase contain information about SNPs on regulatory networks facilitating genetic studies, especially QTL studies.

In the context of *cis*-regulatory elements, this step of GRN inference can be performed either by ChIP-chip experimental approaches or using computational tools from the MEME suite (Bailey et al., 2009), which is a user-friendly web tool (**Table 3**).

Protein–Protein Interaction Networks Provide an Integrated View of the Proteome's Organization and Interactions

Proteins are intrinsically involved in every aspect of cellular bioprocesses. Simplistically, they do so by interacting with other

TABLE 4 | Online resources for acquiring protein interaction information.

Abbreviation	Name	URL	Availability	Data Source
DIP	Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	Academic license	Primary
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org/	License purchase	Secondary
IntAct	IntAct Molecular Interaction Database	http://www.ebi.ac.uk/intact	Free	Primary
BioGRID	Biological General Repository for Interaction Datasets	http://www.thebiogrid.org/	Free	Primary
MINT	Molecular Interaction Database	http://mint.bio.uniroma2.it/	Free	Primary
I2D	Interologous Interaction Database	http://ophid.utoronto.ca/	Academic license	Secondary
CCSB	Center for Cancer Systems Biology Interactome Database	http://interactome.dfci.harvard.edu/	Free	Primary
APID	Agile Protein Interactomes DataServer	http://apid.dep.usal.es/	Free	Secondary
HuRI	The Human Reference Protein Interactome Mapping Project	http://interactome.baderlab.org/	Academic license	Primary
IID	Integrated Interactions Database	http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/	Academic license	Primary

proteins and other biocomponents and the resulting interactions may be strong or transient depending on the biological mechanisms at hand. Thus, the analysis of PPIs is a valuable way to study protein complexes, protein function annotation, and states of health and disease (Barabási et al., 2011; Snider et al., 2015).

To begin understanding the emergent characteristics of PPI one has to retrieve interaction data, which can be obtained from high-throughput techniques, interaction databases, or interaction prediction algorithms. The yeast two-hybrid (Y2H) experimental approach verifies the binary interactions between proteins by fusing them to separate *Gal4* TF DNA binding and activating domains (BD and AD, respectively). The principle of the technique relies on the interaction of a protein fused to BD, called *bait*, to the protein fused to AD, called *prey*. If *bait* and *prey* proteins interact, so do BD and AD, restoring the TF activity which is reported in the assay. The Y2H is scalable and can be used to test protein interaction of many proteins in parallel with some automatization (Fields and Song, 1989).

Along with Y2H, the affinity precipitation coupled to mass spectrometry (AP-MS) yields high-throughput interaction data. Affinity purification methods use the specificity of antibody–epitope interaction to co-purify tightly interacting proteins (Bauer and Kuster, 2003). Coupling the purification phase to an identification step using MS provides means to massively generate interaction data. More PPI data can be retrieved from primary databases that store interaction information from experimental data or computational methods for interaction prediction that may involve protein sequence comparison, interologs comparison, protein surface docking, or evolutionary information using co-mutation profiles (Liu et al., 2008; Wiles et al., 2010; Schoenrock et al., 2017).

The nodes in a PPI network are proteins, and an edge is formed between a protein pair when there is evidence of interaction between them (**Table 1**). Interaction evidence may be accompanied by a score or by the qualification of that evidence, which can be set as an edge attribute to weight the support for that interaction. Usually, scores are calculated to assess the confidence in the interaction, i.e., whether the interaction is confirmed by experimental and/or computational methods. The edges in a PPI network are usually undirected, but depending on the specific objective of the reconstruction it could also be set as a directed network (Vinayagam et al., 2011, Vinayagam et al., 2016).

User-Friendly Tools for Constructing Protein–Protein Interaction Networks

Many online resources of PPI data are available from different experimental or computational methods and for diverse organisms in varying conditions. The webpage Pathguide[1] presents a comprehensive list of metabolic pathways and molecular interaction resources available online and indicating if the resources are free to access, whether they follow a systems biology standard for information description and if they are still available. On the PPI section of Pathguide there are 320 listed databases, from which 246 are still online and accessible. On **Table 4** we have listed some general protein–protein database resources. The databases listed are either free or available through academic licensing, with the exception of STRING, which is free to use online, but in order to download the whole database a license must be purchased. The databases are classified as primary, when they gather experimental or literature-based knowledge, or secondary when they gather predicted protein interactions or reflect only a portion of the information available from primary databases (usually performing secondary analyses therein). The DIP database (Xenarios et al., 2000; Salwinski et al., 2004) has experimental interaction information that is curated automatically and manually giving the data high accuracy. STRING, which was briefly presented in the Introduction, is a database that provides experimental and/or predicted protein interaction data for over 5,000 organisms. The IntAct database (Hermjakob et al., 2004; Kerrien et al., 2012) is open-source and maintained by the European Bioinformatics Institute, gathering experimental protein–protein and protein–compound interaction data. With both protein and genetic interaction data from experimental studies, BIOGRID is a freely available primary database (Stark et al., 2006; Chatr-Aryamontri et al., 2017). It is an excellent source of curated experimental data for many model organisms and especially valuable for budding and fission yeasts. The MINT database (Chatr-Aryamontri et al., 2008) provides interaction data derived from the literature and is freely accessible. The I2D database (Brown and Jurisica, 2005, Brown and Jurisica, 2007) is available online and provides data

[1]http://www.pathguide.org; the webpage is maintained by Dr. Gary Bader at the University of Toronto.

for human PPIs which it imported from primary databases. It can also derive PPI data for other model organisms if they can be mapped to human data. The Center for Cancer Systems Biology (CCSB) provides a primary interaction database named CCSB Interactome Database (http://interactome.dfci.harvard.edu/). The CCSB Interactome Database has experimental binary interaction data for model organisms which can be downloaded and searched freely. APID is a secondary database (Alonso-López et al., 2019) which gathers information from many primary databases, including the Protein Data Bank where protein structures are defined with interacting proteins. As an online web-tool, APID provides the possibility to select interaction properties and interactive mapping of the functional environment of proteins. HuRI, a derivation of the CCSB Interactome Database, is a database with binary PPIs for the human proteome and has three proteome scale protein–protein network reconstructions for the human genome available. Finally, the IID (Kotlyar et al., 2016) database provides tissue-specific interaction data for model organisms and human, harboring both experimental and predicted interactions.

To analyze interaction data, as for the other two previously discussed network approaches, programmable and graphical user interface options are available. For more advanced users with a programming background, tools such as iGraph and NetworkX allow for automation and processing of large-scale datasets (Csardi and Nepusz, 2006; Hagberg et al., 2013), but user-friendly alternatives also exist, which are compiled in **Table 5**. The first step towards constructing a protein interaction network (PIN) is to get interaction data for proteins of interest. This can be done either by experimentation, as briefly described earlier, and/or by retrieving interaction data from the primary and secondary interaction databases described earlier. Interaction data can be directly downloaded or indirectly retrieved using programs or plugins, as is the case for Cytoscape. On the *Interaction database* category in **Table 5** we list Cytoscape apps that can be used to interrogate and retrieve interaction data from various databases. Bisogenet searches for molecular interaction data from an in-house database, SysBiomics, which integrates data from other interaction databases such as DIP, BIOGRID, BIND, MINT, and IntAct. The searches can be filtered to narrow the interaction space, and protein annotations are retrieved from National Center for Biotechnology Information, Uniprot, KEGG, and GO. The Bisogenet app also includes PIN analysis tools. CyPath2 searches for interaction data from the Pathway Commons integrated BioPAX pathway database. PSICQUIC is a built-in feature of Cytoscape that harbors over 10 million binary interactions from 22 active data providers. The list of active providers of interaction data for PSICQUIC can

TABLE 5 | User-friendly computational tools for inferring and analyzing protein interaction networks.

Tool	Description	Category	Reference/URL
Bisogenet	Retrieves interactions associated with input IDs. Sophisticated UI gives links to GO, KEGG, etc.	*Interaction database*	Martin et al., 2010
CyNetSVM	Developed for identification of cancer biomarkers using machine learning approaches.	*PPI-network*	Shi et al., 2017
CyPath2	Pathway Commons (BioPAX L3 database) web service graphical user interface client app.	*Interaction database*	http://apps.cytoscape.org/apps/cypath2
CytoGEDEVO	Pairwise global alignment of PPI or other networks.	*PPI-network*	Malek et al., 2016
CytoMOBAS	Identifies and analyses disease associated and highly connected subnetworks.	*Disease-disease association* *PPI-network*	https://apps.cytoscape.org/apps/cytomobas
DeDal	Applies data dimensionality reduction methods for designing insightful network visualizations.	*PPI-network*	Czerwinska et al., 2015
INTERSPIA	Free online resource for protein interaction comparison between species	Not a Cytoscape app	Kwon et al., 2018
NetworkAnalyst	Free online resource for network construction and analysis	Not a Cytoscape app	Zhou et al., 2019
PathLinker	Reconstructs the interactions in a signaling pathway of interest from the receptors and TFs in a pathway, and can be broadly used to compute and analyze a network of protein interactions.	*PPI-network*	Gil et al., 2017
PEmeasure	Compute links weights and assess the reliability of the links in a network including PPI.	*PPI-network*	Zaki et al., 2013
PEPPER	Find meaningful pathways / complexes connecting a protein set members within a PPI-network using multi-objective optimization.	*Functional module detection*	Winterhalter et al., 2014
PINA	Free online resource capable of PIN construction, filtering, analysis, visualization and management.	Not a Cytoscape app	Wu et al., 2009 Cowley et al., 2012;
PINBPA	Protein-interaction-network-based Pathway Analysis.	*Random walk with restart algorithm*	Wang et al., 2015
PSICQUIC Universal Client	PSICQUIC Web Service Client for importing interactions from public databases.	*Interaction database*	Aranda et al., 2011
stringApp	Import and augment Cytoscape networks from STRING.	*Gene-disease association;* *PPI-network*	Doncheva et al., 2019

PINA, protein interaction network analysis; INTERSPIA, inter-species protein interaction analysis; PINBPA, protein interaction network-based pathway analysis.

be seen at the PSICQUIC Registry page[2]. StringApp imports PPI data from STRING with a user provided protein list (or gene, compound, or disease list). Once imported, a matching network of interactions is disclosed, and functional enrichment analysis can be subsequently performed. The previously cited NetworkAnalyst is an online tool for multi-*omics* analysis, also allowing PPI visualization and analysis. It can take a network in standard format, render visualizations and perform network analysis, also receiving a gene list as input to construct an interaction network. Another online option is the Protein Interaction Network Analysis platform (PINA), which generates PINs from a single protein, a list of proteins, a list of protein pairs or two lists of proteins. Networks generated by PINA can be modified with custom data or with different information from other public interaction databases. Lastly, DeDal is a Cytoscape app that embeds data information into the layout of the network, which can facilitate the user in data interpretation (**Table 5**).

For PPI network analysis, besides the previously described online resources, Cytoscape apps can be used. Apps with the *PPI-Network* tag (**Table 5**) can be applied to study the resulting network. CyNetSVM, specifically geared towards identification of cancer biomarkers, takes as input PINs and applies artificial intelligence techniques with gene expression data to aid in the prediction of clinical outcome. CytoGEDEVO is a Cytoscape app that is capable of aligning networks, especially PINs, which can be used to study the evolution and conservation of proteins interactions. A different approach on comparison of PPIs is used by the online application INTERSPIA, which is freely available. INSTERSPIA can identify interacting proteins in a user-specified list and disclose similar interaction patterns across multiple species. PE-measure, another Cytoscape app, can be used to confirm protein interactions in a network based on its structure, also helping users to identify spurious interactions. Further analysis in PPI networks can be achieved using other tools in Cytoscape. PEPPER, for instance, identifies protein complexes or pathways that are highly condensed using a gene set list as input, helping to integrate information such as protein connections with proteins on the gene set list that are involved in a particular phenotype change, e.g., disease, by finding functional modules. PINBPA is another app that aids in module discovery and is especially suited to integrate GWAS data into protein–protein networks, which can help identify enriched sub-networks and prioritize relevant genes. In the following section we return to the identification of modules in networks in general using algorithms that rely only on the network topology. Finally, PathLinker, a Cytoscape app, can infer signaling networks from PPI networks by computing short paths in a PIN between receptor proteins, as source nodes, and target proteins, as TFs.

A Primer on Network Analysis and Visualization

Once a network of interest is attained, downstream analyses are warranted to extract relevant information and gain knowledge from the reconstruction. These analyses can be broadly divided into *knowledge extraction* and *visualization* steps. There are many methods to evaluate a network and leverage knowledge to help guide interpretation, and this usually begins by exploring local and global interactions within the network. Metrics such as modularity, degree distribution, and other centrality measures are commonly applied to assist in the identification of important or influential nodes in a network (Freeman, 1978; Jeong et al., 2001; Barabási, 2016) (see **Box 1**). Cytoscape has the built-in plugin *NetworkAnalyzer* (Assenov et al., 2008) that computes many centrality metrics, and these can be extended by the Centiscape plugin, which implements ten centrality indexes (Scardoni et al., 2009). Gephi also provides built-in methods to calculate betweenness, eigenvector, and closeness centrality measures, while bridging centrality can be calculated via a third-party plug-in (Bastian et al., 2009). Different centrality methods will usually arrive at distinct rankings of important nodes, which is not unexpected since in order to establish importance each method takes into account different aspects of the data. Betweenness centrality, for instance, emphasizes the importance of a node by considering its contribution in allowing information to pass from one part of the network to the other (thus, a global measure of centrality), while degree centrality simply counts the number of connections between a node and its direct neighbors (thus, a local measure of centrality). For some applications, a combination of centrality metrics may be more appropriate, as has been suggested for metabolic network analysis (Rio et al., 2009). In **Box 1** we present a comparison between selected centrality measures using a toy network, but an exhaustive evaluation is out of the scope of the current work, and efforts have been made to categorize and describe the various centrality indexes, such as the CentiServer online resource (http://www.centiserver.org) (Jalili et al., 2015), which harbors 232 measures of centrality in its last 2017 update, allowing users to input a network and calculate 55 centralities indexes in an interactive web-based application. The use of centrality measures in biological networks dates back to 2001, when Jeong et al. (2001) postulated the 'centrality-lethality rule' using a yeast PIN, and found that the most highly connected proteins in the fungi's cellular network were those more important for its survival, establishing a connection between centrality (a graph-theoretical concept) and essentiality (a biological concept).

Biological networks usually display internal structures that can be identified as subnetworks in modularity analysis (Blondel et al., 2008), which present as densely connected regions, and the disclosed modules can be visually inspected by applying, for instance, the *qgraph* approach (Epskamp et al., 2012) (**Figure 2D**). Modularity (or *Q*) is used as a metric for defining the partitioning of a network and increases its value with increasing network community structure (Newman, 2006). The maximum modularity for a network is $Q = 1$, but in practice values for networks with strong community structure are typically in the range of 0.3-0.7 (Newman and Girvan, 2004). Many module detection techniques have been developed in the recent years and broadly divide into clustering, decomposition, and biclustering methods, which have been subject of recent reviews (Saelens et al., 2018 Rahiminejad et al., 2019). Another use of this approach is to infer biological functions using the guilty-by-association

[2] Available at http://www.ebi.ac.uk/Tools/webservices/psicquic/registry/registry?action=STATUS.

principle, where the role of an uncharacterized gene (or protein) can be predicted by considering the broad functions of the genes with which it clusters in a modularity analysis. As an example, groups of co-expressed genes have a greater chance of being functionally coupled, either by participating in a common biological pathway or by a shared regulatory mechanism, such as an upstream regulator. In this way, novel hypotheses about gene function are generated which can be subsequently explored using as basis a co-expression network. This strategy has successfully led to the identification of novel schizophrenia risk genes, where a co-expression gene set enriched for protein-coding genes associated with the disease was disclosed (Pergola et al., 2017). As was the case for centrality metrics, both Gephi and Cytoscape offer modules to perform clustering analysis, and a Cytoscape example is shown in **Figure 5**. Gephi implements natively the Louvain algorithm, that finds modules by exploring the idea of increasing the network modularity in two phases: first, local modularity gains when neighboring nodes are included in the same cluster in an iterative fashion, which leads to local modularity maxima; second, by considering the disclosed modules from the first phase as communities and aggregating these communities iteratively (forming meta-communities) until attaining a new modularity maximum which cannot be increased further (Blondel et al., 2008). The efficiency of this algorithm allows its application to very large networks on the order of millions of nodes, one of the reasons why it has gained widespread adoption, with almost 9,000 citations (Blondel et al., 2008), including its application to disclose modules related to hepatic dysfunction (Soltis et al., 2017) and cancer (Ajorloo et al., 2017). Other clustering methods available in Gephi through third-party plugins are the Leiden (Traag et al., 2019) and the Girvan-Newman algorithms (Girvan and Newman, 2002). Girvan-Newman works by sequentially removing edges from the network until reaching a maximum modularity, and the nodes that remain connected in the resulting network represent the communities. It has been applied to a wealth of problems (accumulating over 11,000 citations), including to the successful recovery of communities of taxonomically-related organisms using protein sequence data as input (Andrade et al., 2011), but has the drawback of scaling cubically with the number of nodes in its worst case scenario, which limits its use to networks having not more than a few thousand nodes (Girvan and Newman, 2002; Rahiminejad et al., 2019). The Leiden method appeared more recently and claims to improve the quality of the disclosed modules compared to Louvain's method, as well as address some of its shortcomings (Traag et al., 2019). Other clustering methods are available through Cytoscape packages such as *clusterMaker* (Morris et al., 2011) and *CytoCluster* (Li et al., 2017b), with the latter implementing six clustering methods including OH-PIN. In contrast to the previous algorithms that only detect modules containing non-overlapping elements, OH-PIN discloses overlapping clusters typical of many biological networks, such as enzymes that catalyze reactions across multiple pathways.

Once a network is constructed and analyzed from a topological standpoint using the previous approaches, several layout algorithms can be employed to generate visualizations

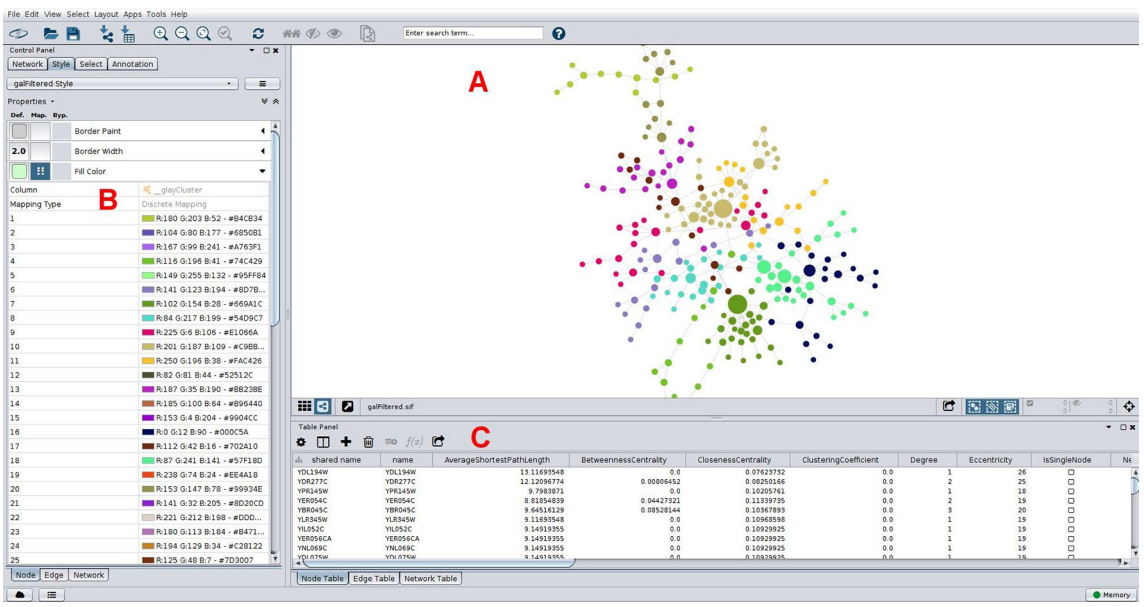


FIGURE 5 | Typical network analyses performed using Cytoscape. A network of yeast protein interaction data is presented **(A)**, with node size scaled with betweenness centrality, which help in straightforward identification of important nodes in this network. Nodes are colored according to its membership to a community as determined using the Girvan-Newman fast greedy algorithm implementation in the *clusterMaker* plugin (Morris et al., 2011). Colors for each community were chosen automatically using a color-generating function and a discrete mapping, with modules numbered sequentially in the left column shown in **(B)**, and colors (in RGB and hex formats) on the right. Properties of nodes are shown below in **(C)**, including some centrality measures. These can be downloaded in-whole as a table for downstream analyses. The network is arranged according to a force-directed layout algorithm.

of the network. While different visualization strategies do not alter the connectivity patterns between nodes, they aid during the identification of influential nodes and communities, while also allowing the organization of the network according to specific properties it may present, such as an underlying node hierarchy. Many layout algorithms are constrained by network size and can perform poorly (consuming extensive memory and CPU) when applied to the ordering of very large networks. Both Gephi (Bastian et al., 2009) and Cytoscape (Shannon et al., 2003) have a plethora of built-in visualization algorithms. In order to arrive at a suitable and pleasant network visualization a number of trial-and-error is involved, not only by qualitatively selecting layout algorithms (which can be coupled in sequence), but also by experimenting with different parameterizations schemes. Force-based algorithms are widely used to arrange networks and follow the general rule that linked nodes attract each other and non-linked nodes are mutually repelled, with inspiration from mechanical forces such as tension and compression acting through a spring, temperature gradients, or even electromagnetic forces. These methods rely only on the topology of the graph in order to arrange the nodes. Consequently, networks laid out according to force-directed strategies usually present similar edge lengths which have a low number of crossings, resulting in an aesthetically pleasing visualization. In Cytoscape, force-directed-based algorithms include the compound spring embedder and prefuse force-directed spring layout, while Gephi implements ForceAtlas2, Fruchterman-Reingold, Yifan-Hu, and OpenOrd. OpenOrd is particularly suitable for large graphs, scaling well for networks over 1 million nodes, and can be followed by the Yifan-Hu layout in order to produce appealing visualizations in such large networks (Pavlopoulos et al., 2017). Both Gephi and Cytoscape can expand their repertoire of layout methods using third-party plugins, such as the proprietary yFiles plugin for Cytoscape which offers nine options for network layout, many of which are multi-purpose such as the force-directed organic (which works well for large graphs) and orthogonal layouts (best applicable to medium-sized networks, routing edges orthogonally), as well as the hierarchic (useful for portraying precedence relationships) and circular layouts (producing star and ring topologies that are useful for visualization of regulatory relationships).

NETWORKS, NETWORKS EVERYWHERE: HEALTH AND DISEASE FROM A GLOBAL STANDPOINT

Networks are now widely employed to help make sense of high-throughput *omics* data. **Figure 6** shows that usage of the networks methods that were covered in this Review is on the rise in the scientific literature. Particularly in the last 5 years, there has been a steep increase in their adoption, especially for co-expression networks, which can be partly due to the falling of sequencing costs, but also to the recent availability of some of the more user-friendly tools that were put available and reviewed herein.

Integrative approaches are particularly suitable for the study of diseases, as they are hardly the effect of single perturbations. These networks allow the identification of associations between

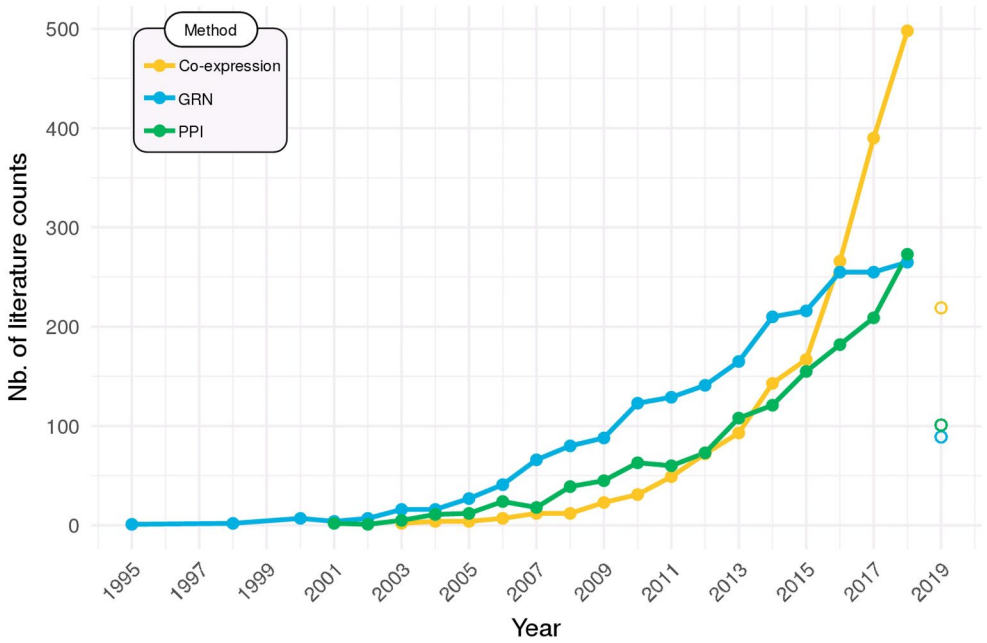


FIGURE 6 | Network methods on the rise. Searches in PubMed (http://ncbi.nlm.nih.gov/pubmed) were performed to identify the all-time use of co-expression networks (query: "co-expression network" OR "coexpression network"), gene regulatory networks (GRN; query: "gene regulatory network"), and protein–protein interaction networks (PPI; query: "protein–protein interaction network"). Data for 2019 is partial (up to March) and are displayed as open points.

the measured components as well as identifying communities (or modules) that could mediate a link between normal and diseased states, including regulatory interactions. Applications of correlation networks include hub genes identification in several diseases such as cancer (Oh et al., 2015), chronic fatigue syndrome (Presson et al., 2008), diabetes (Keller et al., 2008), and in the multivariate disease autism (Voineagu et al., 2011). The use of networks in the context of the neglected tropical disease leishmaniasis was also recently reviewed (Veras et al., 2018). Also performed were the stratification of breast cancer subtypes using human plasma metabolomics (Fan et al., 2016), the study of extracellular proteins in serum to disclose information on human disease states (Emilsson et al., 2018), and the evaluation of coordinated expression patterns in different brain regions in Alzheimer's disease (Wang et al., 2016a). These many studies revealed important pathways and networks of interconnected bioelements that associate with health and disease phenotypes. Co-expression and correlation networks were also used to understanding the immune response of humans to vaccination, disclosing vaccine-induced transcriptional signatures that correlated to protection (Nakaya et al., 2015; Li et al., 2017c), and have also been derived from multi-*omics* data to the understanding and tackling of disease complications from diabetes-tuberculosis comorbidity, where a correlation network constructed from whole-blood gene expression and plasma cytokine measurements was obtained (Prada-Medina et al., 2017).

Finally, disease-disease association uses the information of disease-modules in order to identify common nodes (proteins, genes, metabolites) between diseases which can help pinpoint disease comorbidity or predisposition between conditions. This approach can potentially accelerate drug design since drugs that target interactions that are common between conditions could have a better treatment impact (Barabási et al., 2011). These methods were widely employed to construct disease-disease and gene-disease networks (Serão et al., 2011; Li et al., 2017a; Wiredja et al., 2017; Dong et al., 2018; Liu et al., 2019; Zhang et al., 2018).

While co-expression and PPI networks are tightly related, they are both under the control of regulatory elements, thus the importance of GRNs. Environmental stimuli, pathogen exposure and other disease statuses can trigger a myriad of responses in a cell, including the cascade signals that are recognized by TFs, which in response modulate gene expression. Due to the specificity of GRNs for the conditions of interest, there are multiple GRNs that were generated from specific conditions, such as tissues, environments, pathologies, and the combination of these factors (Guan et al., 2012; Emmert-Streib et al., 2014). This availability of networks from specific conditions can be used to support other studies with similar conditions or used to improve GRNs for other species. In this context, GRNs can be used in health as maps and biomarkers to characterize genetic perturbations associated to rare hereditary variants such as SNPs in the regulatory region of a disease-related gene of interest (Guo and Wang, 2018).

CONCLUSIONS

A variety of tools are available to support the construction of biological networks from *omics* data. Although user-friendliness is usually not a top priority for developers, it can be readily attained with the help of excellent frameworks such as Cytoscape, for which a multitude of plugins are available that permits greatly expanding the capacities of the software beyond its original scope. Also, webserver versions of hitherto command-line only software are increasingly being published. We expect that user empowerment through the breaking of barriers imposed by programming language requirements will allow further adoption of network strategies and accelerate the extraction of knowledge and insights from biological data.

AUTHOR CONTRIBUTIONS

PR conceived the review scope and outline. AQ, KF, LA, NL, and PR wrote the review. PR edited the final version with support from the other authors. All authors read and approved the final version.

FUNDING

NL received financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil [Universal 28/2018; grant protocol 427183/2018-9]. LA received a postdoctoral fellowship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). AQ acknowledges funding from Fundação Oswaldo Cruz (INOVA - Process VPPIS-001-FIO-18-45). Publication fees were defrayed by Fundação Oswaldo Cruz. The funders had no role in study design, analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Aittokallio, T., and Schwikowski, B. (2006). Graph-based methods for analyzing networks in cell biology. *Brief. Bioinform.* 7, 243–255. doi: 10.1093/bib/bbl022

Ajorloo, F., Vaezi, M., Saadat, A., Safaee, S. R., Gharib, B., Ghanei, M., et al. (2017). A systems medicine approach for finding target proteins affecting treatment outcomes in patients with non-Hodgkin lymphoma. *PloS One* 12, e0183969. doi: 10.1371/journal.pone.0183969

Alonso-López, D., Campos-Laborie, F. J., Gutiérrez, M. A., Lambourne, L., Calderwood, M. A., Vidal, M., et al. (2019). APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database* 2019, baz005. doi: 10.1093/database/baz005

Andrade, R. F. S., Rocha-Neto, I. C., Santos, L. B. L., de Santana, C. N., Diniz, M. V. C., Lobão, T. P., et al. (2011). Detecting network communities: an application to phylogenetic analysis. *PloS Comput. Biol.* 7, e1001131. doi: 10.1371/journal.pcbi.1001131

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8, 528–529. doi: 10.1038/nmeth.1637

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics* 24, 282–284. doi: 10.1093/bioinformatics/btm554

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Barabási, A.-L. (2016). *Network Science*. Cambridge, United Kingdom: Cambridge University Press.

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

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

Bastian, M., Heymann, S., Jacomy, M., and Others. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm* 8, 361–362.

Bauer, A., and Kuster, B. (2003). Affinity purification-mass spectrometry: powerful tools for the characterization of protein complexes. *Eur. J. Biochem.* 270, 570–578. doi: 10.1046/j.1432-1033.2003.03428.x

Bavelas, A. (1950). Communication patterns in task-oriented groups. *J. Acoust Soc. America* 22, 725–730. doi: 10.1121/1.1906679

Berlin, R., Gruen, R., and Best, J. (2017). Systems medicine-complexity within, simplicity without. *J. Healthcare Inf. Res.* 1, 119–137. doi: 10.1007/s41666-017-0002-9

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008. doi: 10.1088/1742-5468/2008/10/P10008

Brown, K. R., and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* 21, 2076–2082. doi: 10.1093/bioinformatics/bti273

Brown, K. R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* 8, R95. doi: 10.1186/gb-2007-8-5-r95

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., et al. (2009). NAViGaTOR: network analysis, visualization and graphing toronto. *Bioinformatics* 25, 3327–3329. doi: 10.1093/bioinformatics/btp595

Cardozo, L. E., Russo, P. S. T., Gomes-Correia, B., Araujo-Pereira, M., Sepúlveda-Hermosilla, G., Maracaja-Coutinho, V., et al. (2019). webCEMiTool: Co-expression modular analysis made easy. *Front. Genet.* 10, 146. doi: 10.3389/fgene.2019.00146

Carpenter, A. E., and Sabatini, D. M. (2004). Systematic genome-wide screens of gene function. *Nat. Rev. Genet.* 5, 11–22. doi: 10.1038/nrg1248

Chai, L., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* 48, 55–65. doi: 10.1016/j.compbiomed.2014.02.011

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45, D369–D379. doi: 10.1093/nar/gkw1102

Chatr-Aryamontri, A., Zanzoni, A., Ceol, A., and Cesareni, G. (2008). Searching the protein interaction space through the MINT database. *Methods Mol. Biol.* 484, 305–317. doi: 10.1007/978-1-59745-398-1_20

Cottret, L., and Jourdan, F. (2010). Graph methods for the investigation of metabolic networks in parasitology. *Parasitology* 137, 1393–1407. doi: 10.1017/S0031182010000363

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., et al. (2012). PINA v2.0: mining interactome modules. *Nucleic Acids Res.* 40, D862–D865. doi: 10.1093/nar/gkr967

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Syst.* 1695, 1–9.

Czerwinska, U., Calzone, L., Barillot, E., and Zinovyev, A. (2015). DeDaL: cytoscape 3 app for producing and morphing data-driven and structure-driven network layouts. *BMC Syst. Biol.* 9, 46. doi: 10.1186/s12918-015-0189-4

de Matos Simoes, R., Dehmer, M., and Emmert-Streib, F. (2013). B-cell lymphoma gene regulatory networks: biological consistency among inference methods. *Front. Genet.* 4, 281. doi: 10.3389/fgene.2013.00281

de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., and Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief. Bioinform.* 15, 906–918. doi: 10.1093/bib/bbt051

Doncheva, N. T., Morris, J. H., Gorodkin, J., and Jensen, L. J. (2019). Cytoscape stringapp: network analysis and visualization of proteomics data. *J. Proteome Res.* 18, 623–632. doi: 10.1021/acs.jproteome.8b00702

Dong, H., Zhang, S., Wei, Y., Liu, C., Wang, N., Zhang, P., et al. (2018). Bioinformatic analysis of differential expression and core GENEs in breast cancer. *Int. J. Clin. Exp. Pathol.* 11(3), 1146–1156.

Dutta, N. K., Bandyopadhyay, N., Veeramani, B., Lamichhane, G., Karakousis, P. C., and Bader, J. S. (2014). Systems biology-based identification of mycobacterium tuberculosis persistence genes in mouse lungs. *mBio* 5, e01066–13. doi: 10.1128/mBio.01066-13

Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. *Science* 361, 769–773. doi: 10.1126/science.aaq1327

Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.* 5, 15. doi: 10.3389/fgene.2014.00015

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., and Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Software* 48, 1–18. doi: 10.18637/jss.v048.i04

Fan, Y., Zhou, X., Xia, T.-S., Chen, Z., Li, J., Liu, Q., et al. (2016). Human plasma metabolomics for identifying differential metabolites and predicting molecular subtypes of breast cancer. *Oncotarget* 7, 9925–9938. doi: 10.18632/oncotarget.7155

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245–246. doi: 10.1038/340245a0

Floratos, A., Smith, K., Ji, Z., Watkinson, J., and Califano, A. (2010). geWorkbench: an open source platform for integrative genomics. *Bioinformatics* 26, 1779–1780. doi: 10.1093/bioinformatics/btq282

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Soc Networks* 1, 215–239. doi: 10.1016/0378-8733(78)90021-7

Fronczuk, M., Raftery, A. E., and Yeung, K. Y. (2015). CyNetworkBMA: a Cytoscape app for inferring gene regulatory networks. *Source Code Biol. Med.* 10, 11. doi: 10.1186/s13029-015-0043-5

Gallo, C. A., Carballido, J. A., and Ponzoni, I. (2011). Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinf.* 12, 123. doi: 10.1186/1471-2105-12-123

Gil, D. P., Law, J. N., and Murali, T. M. (2017). The PathLinker app: connect the dots in protein interaction networks. *F1000Res.* 6, 58. doi: 10.12688/f1000research.9909.1

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

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

Guitart-Pla, O., Kustagi, M., Rügheimer, F., Califano, A., and Schwikowski, B. (2015). The Cyni framework for network inference in Cytoscape. *Bioinformatics* 31, 1499–1501. doi: 10.1093/bioinformatics/btu812

Guo, L., and Wang, J. (2018). rSNPBase 3.0: an updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks. *Nucleic Acids Res.* 46, D1111–D1116. doi: 10.1093/nar/gkx1101

Hache, H., Lehrach, H., and Herwig, R. (2009). Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J. Bioinform. Syst. Biol.* 2009(1), 617281. doi: 10.1155/2009/617281

Hagberg, A., Schult, D., Swart, P., Conway, D., Séguin-Charbonneau, L., Ellison, C., et al. (2013). "Networkx," in *high productivity software for complex networks*. Webová strá nka Available at: https://networkx. lanl. gov/wiki.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46, D380–D386. doi: 10.1093/nar/gkx1013

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007). A robust measure of correlation between two genes on a microarray. *BMC Bioinf.* 8, 220. doi: 10.1186/1471-2105-8-220

Hartemink, A. J. (2005). Reverse engineering gene regulatory networks. *Nat. Biotechnol.* 23, 554–555. doi: 10.1038/nbt0505-554

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–D455. doi: 10.1093/nar/gkh052

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8, 84. doi: 10.3389/fgene.2017.00084

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS One* 5, e12776. doi: 10.1371/journal.pone.0012776

Jalili, M., Salehzadeh-Yazdi, A., Asgari, Y., Arab, S. S., Yaghmaie, M., Ghavamzadeh, A., et al. (2015). CentiServer: a comprehensive resource, web-based application and r package for centrality analysis. *PloS One* 10, e0143111. doi: 10.1371/journal.pone.0143111

Janky, R., Verfaillie, A., Imrichová, H., Van de Sande, B., Standaert, L., Christiaens, V., et al. (2014). iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PloS Comput. Biol.* 10, e1003731. doi: 10.1371/journal.pcbi.1003731

Jeong, H., Mason, S. P., Barabási, -L.,. A., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138

Jha, A. K., Huang, S. C.-C., Sergushichev, A., Lampropoulou, V., Ivanova, Y., Loginicheva, E., et al. (2015). Network integration of parallel metabolic and transcriptional data reveals metabolic modules that regulate macrophage polarization. *Immunity* 42, 419–430. doi: 10.1016/j.immuni.2015.02.005

Kaderali, L., and Radde, N. (2008). Inferring gene regulatory networks from expression data. *Comput. Intell. Bioinf.* 94, 33–74. doi: 10.1007/978-3-540-76803-6_2

Keller, M. P., Choi, Y., Wang, P., Davis, D. B., Rabaglia, M. E., Oler, A. T., et al. (2008). A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* 18, 706–716. doi: 10.1101/gr.074914.107

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846. doi: 10.1093/nar/gkr1088

Klein, C. C., Cottret, L., Kielbassa, J., Charles, H., Gautier, C., Ribeiro de Vasconcelos, A. T., et al. (2012). Exploration of the core metabolism of symbiotic bacteria. *BMC Genomics* 13, 438. doi: 10.1186/1471-2164-13-438

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* 44, D536–D541. doi: 10.1093/nar/gkv1115

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Kwon, D., Lee, D., Kim, J., Lee, J., Sim, M., and Kim, J. (2018). INTERSPIA: a web application for exploring the dynamics of protein-protein interactions among multiple species. *Nucleic Acids Res.* 46, W89–W94. doi: 10.1093/nar/gky378

Lacroix, V., Cottret, L., Thebault, P., and Sago, -F.t, M. (2008). An Introduction to Metabolic Networks and Their Structural Analysis. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 5, 594–617. doi: 10.1109/TCBB.2008.79

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559

Langfelder, P., and Horvath, S. (2012). Fast R Functions for robust correlations and hierarchical clustering. *J. Stat. Software* 46(11), i11. doi: 10.18637/jss.v046.i11

Lee, T. I., Johnstone, S. E., and Young, R. A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* 1, 729–748. doi: 10.1038/nprot.2006.98

Lesurf, R., Cotto, K. C., Wang, G., Griffith, M., Kasaian, K., Jones, S. J. M., et al. (2016). ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.* 44, D126–D132. doi: 10.1093/nar/gkv1203

Li, D.-Y., Chen, W.-J., Luo, L., Wang, Y.-K., Shang, J., Zhang, Y., et al. (2017a). Prospective lncRNA-miRNA-mRNA regulatory network of long non-coding RNA LINC00968 in non-small cell lung cancer A549 cells: A miRNA microarray and bioinformatics investigation. *Int. J. Mol. Med.* 40(6), 1895–1906 doi: 10.3892/ijmm.2017.3187

Li, M., Li, D., Tang, Y., Wu, F., and Wang, J. (2017b). CytoCluster: a cytoscape plugin for cluster analysis and visualization of biological networks. *Int. J. Mol. Sci.* 18, 1880. doi: 10.3390/ijms18091880

Li, S., Sullivan, N. L., Rouphael, N., Yu, T., Banton, S., Maddur, M. S., et al. (2017c). Metabolic phenotypes of response to vaccination in humans. *Cell* 169, 862–877.e17. doi: 10.1016/j.cell.2017.04.026

Liu, D., Skomorovska, Y., Song, J., Bowler, E., Harris, R., Ravasz, M., et al. (2019). ELF3 is an antagonist of oncogenic-signaling-induced expression of EMT-TF ZEB1. *Cancer Biol. Ther.* 20(1) 90–100. doi: 10.1080/15384047.2018.1507256

Liu, S., Gao, Y., and Vakser, I. A. (2008). Dockground protein-protein docking decoy set. *Bioinformatics* 24, 2634–2635. doi: 10.1093/bioinformatics/btn497

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015, bav095. doi: 10.1093/database/bav095

Luo, W., Hankenson, K. D., and Woolf, P. J. (2008). Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinf.* 9, 467. doi: 10.1186/1471-2105-9-467

Malek, M., Ibragimov, R., Albrecht, M., and Baumbach, J. (2016). CytoGEDEVO-global alignment of biological networks with Cytoscape. *Bioinformatics* 32, 1259–1261. doi: 10.1093/bioinformatics/btv732

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006a). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf.* 7 Suppl 1, S7. doi: 10.1186/1471-2105-7-S1-S7

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006b). Reverse engineering cellular networks. *Nat. Protoc.* 1, 662–671. doi: 10.1038/nprot.2006.106

Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de-Cossio, J., and Bringas, R. (2010). BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinf.* 11, 91. doi: 10.1186/1471-2105-11-91

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinf.* 9, 461. doi: 10.1186/1471-2105-9-461

Modrák, M., and Vohradský, J. (2018). Genexpi: a toolset for identifying regulons and validating gene regulatory networks using time-course expression data. *BMC Bioinf.* 19, 137. doi: 10.1186/s12859-018-2138-x

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., et al. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinf.* 12, 436. doi: 10.1186/1471-2105-12-436

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24(3) 69–71.

Nakaya, H. I., Hagan, T., Duraisingham, S. S., Lee, E. K., Kwissa, M., Rouphael, N., et al. (2015). Systems analysis of immunity to influenza vaccination across multiple years and in diverse populations reveals shared molecular signatures. *Immunity* 43, 1186–1198. doi: 10.1016/j.immuni.2015.11.012

Naldi, A., Berenguiera, D., Fauréa, A., Lopeza, F., Thieffrya, D., and Chaouiya, C. (2009). Logical modelling of regulatory networks with GINsim 2.3. *BioSystems* 97, 134–139. doi: 10.1016/j.biosystems.2009.04.008

Narasimhan, S., Rengaswamy, R., and Vadigepalli, R. (2009). Structural properties of gene regulatory networks: definitions and connections. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6, 158–170. doi: 10.1109/TCBB.2007.70231

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 69, 026113. doi: 10.1103/PhysRevE.69.026113

Ngounou Wetie, A. G., Sokolowska, I., Woods, A. G., Roy, U., Deinhardt, K., and Darie, C. C. (2014). Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. Life Sci.* 71, 205–228. doi: 10.1007/s00018-013-1333-1

Oh, E.-Y., Christensen, S. M., Ghanta, S., Jeong, J. C., Bucur, O., Glass, B., et al. (2015). Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biol.* 16, 128. doi: 10.1186/s13059-015-0675-4

Pavlopoulos, G. A., Paez-Espino, D., Kyrpides, N. C., and Iliopoulos, I. (2017). Empirical comparison of visualization tools for larger-scale network analysis. *Adv. Bioinf.* 2017, 1278932. doi: 10.1155/2017/1278932

Pergola, G., Di Carlo, P., D'Ambrosio, E., Gelao, B., Fazio, L., Papalino, M., et al. (2017). DRD2 co-expression network and a related polygenic index predict imaging, behavioral and clinical phenotypes linked to schizophrenia. *Trans. Psychiatry* 7, e1006–e1006. doi: 10.1038/tp.2016.253

Prada-Medina, C. A., Fukutani, K. F., Pavan Kumar, N., Gil-Santana, L., Babu, S., Lichtenstein, F., et al. (2017). Systems immunology of diabetes-tuberculosis

comorbidity reveals signatures of disease complications. *Sci. Rep.* 7, 1999. doi: 10.1038/s41598-017-01767-4

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., et al. (2008). Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst. Biol.* 2, 95. doi: 10.1186/1752-0509-2-95

Rahiminejad, S., Maurya, M. R., and Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinf.* 20, 212. doi: 10.1186/s12859-019-2746-0

Rio, G., del Rio, G., Koschützki, D., and Coello, G. (2009).). How to identify essential genes from molecular networks? *BMC Syst. Biol.* 3, 102. doi: 10.1186/1752-0509-3-102

Russo, P. S. T., Ferreira, G. R., Cardozo, L. E., Bürger, M. C., Arias-Carrasco, R., Maruyama, S. R., et al. (2018). CEMiTool: a bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinf.* 19, 56. doi: 10.1186/s12859-018-2053-1

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika* 31, 581–603. doi: 10.1007/BF02289527

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9, 1090. doi: 10.1038/s41467-018-03424-4

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., et al. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9, 1069–1076. doi: 10.1038/nmeth.2212

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi: 10.1093/nar/gkh086

Scardoni, G., Petterlini, M., and Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe. *Bioinformatics* 25, 2857–2859. doi: 10.1093/bioinformatics/btp517

Schoenrock, A., Burnside, D., Moteshareie, H., Pitre, S., Hooshyar, M., Green, J. R., et al. (2017). Evolution of protein-protein interaction networks in yeast. *PloS One* 12, e0171920. doi: 10.1371/journal.pone.0171920

Serão, N. V. L., Delfino, K. R., Southey, B. R., Beever, J. E., and Rodriguez-Zas, S. L. (2011). Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. *BMC Med. Genomics* 4, 49. doi: 10.1186/1755-8794-4-49

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, X., Banerjee, S., Chen, L., Hilakivi-Clarke, L., Clarke, R., and Xuan, J. (2017). CyNetSVM: A cytoscape app for cancer biomarker identification using network constrained support vector machines. *PloS One* 12, e0170482. doi: 10.1371/journal.pone.0170482

Shrinet, J., Nandal, U. K., Adak, T., Bhatnagar, R. K., and Sunil, S. (2014). Inference of the oxidative stress network in Anopheles stephensi upon Plasmodium infection. *PloS One* 9, e114461. doi: 10.1371/journal.pone.0114461

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* 11, 848. doi: 10.15252/msb.20156351

Soltis, A. R., Kennedy, N. J., Xin, X., Zhou, F., Ficarro, S. B., Yap, Y. S., et al. (2017). Hepatic dysfunction caused by consumption of a high-fat diet. *Cell Rep.* 21, 3317–3328. doi: 10.1016/j.celrep.2017.11.059

Song, Q., Grene, R., Heath, L. S., and Li, S. (2017). Identification of regulatory modules in genome scale transcription regulatory networks. *BMC Syst. Biol.* 11, 140. doi: 10.1186/s12918-017-0493-2

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539. doi: 10.1093/nar/gkj109

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18, S231–S240. doi: 10.1093/bioinformatics/18.suppl_2.S231

Stevens, A., De Leonibus, C., Hanson, D., Dowsey, A. W., Whatmore, A., Meyer, S., et al. (2014). Network analysis: a new approach to study endocrine disorders. *J. Mol. Endocrinol.* 52, R79–R93. doi: 10.1530/JME-13-0112

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tawfike, A. F., Romli, M., Clements, C., Abbott, G., Young, L., Schumacher, M., et al. (2019). Isolation of anticancer and anti-trypanosome secondary metabolites from the endophytic fungus Aspergillus flocculus via bioactivity guided isolation and MS based metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 1106-1107, 71–83. doi: 10.1016/j.jchromb.2018.12.032

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 5233. doi: 10.1038/s41598-019-41695-z

Vella, D., Zoppis, I., Mauri, G., Mauri, P., and Di Silvestre, D. (2017). From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP J. Bioinform. Syst. Biol.* 2017, 6. doi: 10.1186/s13637-017-0059-z

Veras, P. S. T., Ramos, P. I. P., and de Menezes, J. P. B. (2018). In search of biomarkers for pathogenesis and control of leishmaniasis by global analyses of -infected macrophages. *Front. Cell. Infect. Microbiol.* 8, 326. doi: 10.3389/fcimb.2018.00326

Vinayagam, A., Gibson, T. E., Lee, H.-J., Yilmazel, B., Roesel, C., Hu, Y., et al. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. U. S. A.* 113, 4976–4981. doi: 10.1073/pnas.1603992113

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* 4, rs8. doi: 10.1126/scisignal.2001699

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., et al. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384. doi: 10.1038/nature10110

Walter, S. D., and Altman, D. G. (1992). practical statistics for medical research. *Biometrics* 48, 656. doi: 10.2307/2532320

Wang, L., Matsushita, T., Madireddy, L., Mousavi, P., and Baranzini, S. E. (2015). PINBPA: cytoscape app for network analysis of GWAS data. *Bioinformatics* 31, 262–264. doi: 10.1093/bioinformatics/btu644

Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K. J., et al. (2016a). Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med.* 8, 104. doi: 10.1186/s13073-016-0355-3

Wang, P., Wang, Y., Hang, B., Zou, X., and Mao, J.-H. (2016b). A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. *Oncotarget* 7, 55343–55351. doi: 10.18632/oncotarget.10533

Wei, T., and Simko (2017). V., *R package "corrplot": visualization of a correlation matrix (version 0.84)*, 2017, Retrieved from https://github. com/taiyun/corrplot.

Wiles, A. M., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B., et al. (2010). Building and analyzing protein interactome networks by cross-species comparisons. *BMC Syst. Biol.* 4, 36. doi: 10.1186/1752-0509-4-36

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol.* 5(11) R92. doi: 10.1186/gb-2004-5-11-r92

Winterhalter, C., Nicolle, R., Louis, A., To, C., Radvanyi, F., and Elati, M. (2014). Pepper: cytoscape app for protein complex expansion using protein-protein interaction networks. *Bioinformatics* 30, 3419–3420. doi: 10.1093/bioinformatics/btu517

Wiredja, D. D., Ayati, M., Mazhar, S., Sangodkar, J., Maxwell, S., Schlatzer, D., et al. (2017). Phosphoproteomics profiling of nonsmall cell lung cancer cells treated with a novel phosphatase activator. *Proteomics* 17, 1700214. doi: 10.1002/pmic.201700214

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat. Methods* 6, 75–77. doi: 10.1038/nmeth.1282

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291. doi: 10.1093/nar/28.1.289

Zaki, N., Efimov, D., and Berengueres, J. (2013). Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinf.* 14, 163. doi: 10.1186/1471-2105-14-163

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi: 10.2202/1544-6115.1128

Zhang, F., Xu, W., Liu, J., Liu, X., Huo, B., Li, B., et al. (2018). Optimizing miRNA-module diagnostic biomarkers of gastric carcinoma via integrated network analysis. *PloS One* 13, e0198445. doi: 10.1371/journal.pone.0198445

Zhang, W., Mao, J.-H., Zhu, W., Jain, A. K., Liu, K., Brown, J. B., et al. (2016). Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat. Commun.* 7, 12619. doi: 10.1038/ncomms12619

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* 47(W1) W234–W241. doi: 10.1093/nar/gkz240

Conflict of Interest: 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