MULTI-OMIC DATA INTEGRATION

EDITED BY: Paolo Tieri, Christine Nardini and Jennifer Elizabeth Dent

PUBLISHED IN: Frontiers in Cell and Developmental Biology and
Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88919-648-7
DOI 10.3389/978-2-88919-648-7

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**

MULTI-OMIC DATA INTEGRATION

Topic Editors:
Paolo Tieri, Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni del Calcolo, Italy
Christine Nardini, Shanghai Institutes for Biological Sciences, China
Jennifer Elizabeth Dent, Quintiles, UK

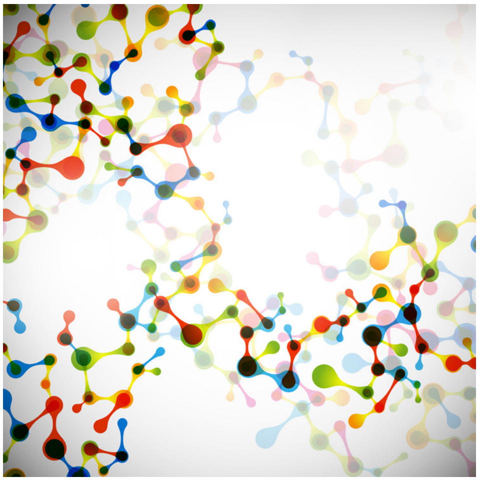


Image by Palau/Shutterstock.
http://www.shutterstock.com/pic-256604953/stock-vector-eps-beautiful-structure-of-the-dna-molecule.html?src=R8wMrezjeJJPjQcpjaDG-w-1-5

Stable, predictive biomarkers and interpretable disease signatures are seen as a significant step towards personalized medicine. In this perspective, integration of multi-omic data coming from genomics, transcriptomics, glycomics, proteomics, metabolomics is a powerful strategy to reconstruct and analyse complex multi-dimensional interactions, enabling deeper mechanistic and medical insight.

At the same time, there is a rising concern that much of such different omic data –although often publicly and freely available- lie in databases and repositories underutilised or not used at all. Issues coming from lack of standardisation and shared biological identities are also well-known.

From these considerations, a novel, pressing request arises from the life sciences to design methodologies and approaches that allow for these data to be interpreted as a whole, i.e. as intertwined molecular signatures containing genes, proteins, mRNAs and miRNAs, able to capture inter-layers connections and complexity.

Papers discuss data integration approaches and methods of several types and extents, their application in understanding the pathogenesis of specific diseases or in identifying candidate biomarkers to exploit the full benefit of multi-omic datasets and their intrinsic information content.

Topics of interest include, but are not limited to:

- Methods for the integration of layered data, including, but not limited to, genomics, transcriptomics, glycomics, proteomics, metabolomics;

- Application of multi-omic data integration approaches for diagnostic biomarker discovery in any field of the life sciences;
- Innovative approaches for the analysis and the visualization of multi-omic datasets;
- Methods and applications for systematic measurements from single/undivided samples (comprising genomic, transcriptomic, proteomic, metabolomic measurements, among others);
- Multi-scale approaches for integrated dynamic modelling and simulation;
- Implementation of applications, computational resources and repositories devoted to data integration including, but not limited to, data warehousing, database federation, semantic integration, service-oriented and/or wiki integration;
- Issues related to the definition and implementation of standards, shared identities and semantics, with particular focus on the integration problem.

Research papers, reviews and short communications on all topics related to the above issues were welcomed.

Citation: Tieri, P., Nardini, C., Dent, J. E., eds. (2015). Multi-omic Data Integration. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-648-7

Table of Contents

frontiers in Cell and Developmental Biology

EDITORIAL
published: 07 July 2015
doi: 10.3389/fcell.2015.00046

Editorial: Multi-omic data integration

Christine Nardini[1,2], *Jennifer Dent*[3] and *Paolo Tieri*[4]*

[1] Lazzari, Bologna, Italy, [2] Group of Clinical Genomic Networks, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, China, [3] Quintiles, Reading, UK, [4] Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni del Calcolo, Rome, Italy

Keywords: multi-omics, multi-omic data integration, integration, systems biology, network analysis

As researchers involved in molecular biology, we are witnessing tremendous paradigm changes in a time frame that becomes shorter and shorter. The epoch-making notion, originally put forward by the central dogma of biology (Crick, 1970), that there is a unidirectional process and a privileged level (genetic) of causality at which biological functions are determined, has already long and strongly been challenged. It is in fact well recognised that multi-level causality with feedback cycles among all former and newly identified biochemical levels (including small RNAs, epigenomic changes) is a fundamental attribute of biological systems (Noble, 2012).

Yet, the focus shift from single reactions to transcriptomics, promoted by microarray first and sequencers now, is already challenged by a novel, pressing offer from fast evolving technologies. Indeed, the possibility to have a *omic* view on virtually all molecular layers (genomes, metagenomes, transcriptomes, proteomes, epigenomes) pushes to integrate the study of systems at yet another level of complexity, a run harmed, and not negligibly, by the difficulties in formatting, storing, and reusing the deluge of data encompassing every level of biological organization.

In such a complex background, it is growingly acknowledged that tools and theoretical frameworks that could help in combining and giving account for both the multi-level causation scheme and the burden of data are still underdeveloped (Witzany and Baluska, 2012).

From these considerations, a novel, pressing request arises to design methodologies, approaches and frameworks that allow for these data to be interpreted as a whole, i.e., as intertwined molecular signatures containing genes, proteins, mRNAs, and miRNAs, but also epigenomic characterizations, as well as correlations with microbiomes' compositions, just to name the major, able to capture the inter-layers connections and the complexity of phenotypes. This request is seconded by demands and concerns about the storage and reusability of much of such different omic data. Indeed, although publicly and freely available, these data often lie in databases and repositories underutilized or not used at all. Issues coming from lack of standardization and shared biological identities are also well known to represent a hurdle for data reuse (Tieri and Nardini, 2013; Chowdhury and Sarkar, 2015).

The "Multi-Omic Data Integration" Research Topic is in our intention a dedicated forum to collect efforts that help in defining this emerging field, aimed to the integration of data, analyses and approaches from, and for multiple omics.

The articles here collected address these questions from a number of perspectives that we summarize as *experimental, network based,* and *methodological.* In the first category the authors extract and analyse different types of high-throughput data (epitomics, localisomics, transcriptomics, lipidomics) from different locations on model organisms [*Arabidopsis thaliana* (Wilson et al., 2015) and *rhesus macaques* (Lee et al., 2014)] to understand a complex biological question (roots' growth and response to anti-malarial drugs) that could not be addressed with single-omic approaches.

We transition from these approaches to more theoretical ones via the usage of graphs. Networks offer a complete, intuitive, versatile, and powerful approach to the representation of complex systems (genomics, epigenomics, transcriptomics, metabolomics, host-microbiome interface, diseases' phenomics) which is here exploited to represent the multifaceted aspects of complex autoimmune diseases (*rheumatoid arthritis,* Tieri et al., 2014) in order to evaluate complex side

Edited by:
*Raina Robeva,
Sweet Briar College, USA*

Reviewed by:
*Matteo Barberis,
University of Amsterdam, Netherlands*

***Correspondence:**
*Paolo Tieri,
p.tieri@iac.cnr.it*

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

Received: *27 May 2015*
Accepted: *25 June 2015*
Published: *07 July 2015*

Citation:
Nardini C, Dent J and Tieri P (2015) Editorial: Multi-omic data integration. Front. Cell Dev. Biol. 3:46. doi: 10.3389/fcell.2015.00046

effects of old and novel therapies; to identify disease molecules that can be both effective therapeutic targets relevant progression markers with application to *diabetic nephropathy* (Heinzel et al., 2014); to stratify patients with comorbidities (Moni and Lio, 2015).

Methodological approaches point with a novel emphasis at the importance of molecules' spatial localization in the omic context. From polysome and ribosome profiling, RNA, and miRNA binding sites annotation and standardization (Dassi and Quattrone, 2014), to networks including 3D molecules' proximity thanks to Chromosome Conformation Capture (3C) and its omic version Hi-C (Merelli et al., 2015), spatial representation contributes with an important layer of information in this added multi-omic complexity.

Beyond spatial organization, temporal progression and causal inference are discussed to model the heterogeneity of CD4$^+$ T cells and their complex immune responses (Carbo et al., 2014), and to predict gene networks based on ChIP-seq and RNA-seq integration (Angelini and Costa, 2014).

Finally, meta analyses of genomes, be it for the exploration of microbiomes' compositions or disease genome-wide association studies (GWAS) still benefit from discussion in this research topic, on one side for the need of standardization of the workflow (Ladoukakis et al., 2014) in a relatively novel research area (omic microbiology) and on the other side to compensate with multi-omic layers to the limited statistical power and reproducibility of GWAS (Lin et al., 2014).

This collection is the tip of an iceberg that continues to grow and to evolve in multiple directions. From the continuously improving efficiency of existing high-throughput platforms that imply easier, cheaper and more frequent spatio-temporal sampling, to the input of novel technologies that will offer omic views on novel types of data (phenotypes, tissues, 3D proteins etc., all entailing the production and approval of dedicated standards for data storage) we are only at the beginning of almost endless possibilities of data integration.

However, to avoid getting lost in the sea of data, efficient algorithms as well as biologically meaningful directions in which to integrate information will be of importance. This will imply not only the implementation of powerful tools to give answers, but also the design of careful approaches to form questions.

We hope and foresee that these needs will foster the collaboration between biologists, medical doctors, statisticians, and computer scientists further, transforming the residual perception of this forced cooperation from a burden to a resource. The impact of completing this other type of integration among scientific expertise is difficult to predict at large, but can easily be assumed as a necessary and crucial starting point for the effective implementation of personalized medicine, where patients' and health practitioners' needs are translated into technology and report on systemic markers, offering patients the possibility to be treated as a whole and not as a mere assemblage of parts to be "adjusted."

Funding

This project is partialy funded by MoST international cooperation program no. 2013DFA30790, NSFC no. 31070748, EC FP7-PEOPLE-2011-IRSES program, project 294935 "KEPAMOD," and CAS fellow grant no. 2011Y1SA04.

References

Angelini, C., and Costa, V. (2014). Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front. Cell Dev. Biol.* 2:51. doi: 10.3389/fcell.2014.00051

Carbo, A., Hontecillas, R., Andrew, T., Eden, K., Mei, Y., Hoops, S., et al. (2014). Computational modeling of heterogeneity and function of CD4$^+$ T cells. *Front. Cell Dev. Biol.* 2:31. doi: 10.3389/fcell.2014.00031

Chowdhury, S., and Sarkar, R. R. (2015). Comparison of human cell signaling pathway databases–evolution, drawbacks and challenges. *Database (Oxford)* 2015. doi: 10.1145/2752746

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

Dassi, E., and Quattrone, A. (2014). Fingerprints of a message: integrating positional information on the transcriptome. *Front. Cell Dev. Biol.* 2:39. doi: 10.3389/fcell.2014.00039

Heinzel, A., Perco, P., Mayer, G., Oberbauer, R., Lukas, A., and Mayer, B. (2014). From molecular signatures to predictive biomarkers: modeling disease pathophysiology and drug mechanism of action. *Front. Cell Dev. Biol.* 2:37. doi: 10.3389/fcell.2014.00037

Ladoukakis, E., Kolisis, F. N., and Chatziioannou, A. A. (2014). Integrative workflows for metagenomic analysis. *Front. Cell Dev. Biol.* 2:70. doi: 10.3389/fcell.2014.00070

Lee, K. J., Yin, W., Arafat, D., Tang, Y., Uppal, K., Tran, V., et al. (2014). Comparative transcriptomics and metabolomics in a rhesus macaque drug administration study. *Front. Cell Dev. Biol.* 2:54. doi: 10.3389/fcell.2014.00054

Lin, D., Zhang, J., Li, J., He, H., Deng, H. W., and Wang, Y. P. (2014). Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. *Front. Cell Dev. Biol.* 2:62. doi: 10.3389/fcell.2014.00062

Merelli, I., Tordini, F., Drocco, M., Aldinucci, M., Lio, P., and Milanesi, L. (2015). Integrating multi-omic features exploiting chromosome conformation capture data. *Front. Genet.* 6:40. doi: 10.3389/fgene.2015.00040

Moni, M. A., and Lio, P. (2015). How to build personalized multi-omics comorbidity profiles. *Front. Cell Dev. Biol.* 3:28. doi: 10.3389/fcell.2015.00028

Noble, D. (2012). A theory of biological relativity: no privileged level of causation. *Interface Focus* 2, 55–64. doi: 10.1098/rsfs.2011.0067

Tieri, P., and Nardini, C. (2013). Signalling pathway database usability: lessons learned. *Mol. Biosyst.* 9, 2401–2407. doi: 10.1039/c3mb70242a

Tieri, P., Zhou, X., Zhu, L., and Nardini, C. (2014). Multi-omic landscape of rheumatoid arthritis: re-evaluation of drug adverse effects. *Front. Cell Dev. Biol.* 2:59. doi: 10.3389/fcell.2014.00059

Wilson, M. H., Holman, T. J., Sorensen, I., Cancho-Sanchez, E., Wells, D. M., Swarup, R., et al. (2015). Multi-omics analysis identifies genes mediating the extension of cell walls in the *Arabidopsis thaliana* root elongation zone. *Front. Cell Dev. Biol.* 3:10. doi: 10.3389/fcell.2015.00010

Witzany, G., and Baluska, F. (2012). Life's code script does not code itself. The machine metaphor for living organisms is outdated. *EMBO Rep.* 13, 1054–1056. doi: 10.1038/embor.2012.166

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
CELL AND DEVELOPMENTAL BIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 February 2015
doi: 10.3389/fcell.2015.00010

Multi-omics analysis identifies genes mediating the extension of cell walls in the *Arabidopsis thaliana* root elongation zone

Michael H. Wilson[1‡], Tara J. Holman[1‡], Iben Sørensen[2†‡], Ester Cancho-Sanchez[1], Darren M. Wells[1], Ranjan Swarup[1], J. Paul Knox[3], William G. T. Willats[2], Susana Ubeda-Tomás[1], Michael Holdsworth[1], Malcolm J. Bennett[1], Kris Vissenberg[4]* and T. Charlie Hodgman[1]*

[1] Centre for Plant Integrative Biology, School of Biosciences, University of Nottingham, Sutton Bonington, UK
[2] Plant Glycobiology Section, Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, Denmark
[3] Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds, UK
[4] Laboratory of Plant Growth and Development, Department of Biology, University of Antwerp, Antwerp, Belgium

Edited by:
Paolo Tieri, Consiglio Nazionale delle Ricerche, Italy

Reviewed by:
Ayse Meric Ovacik, Merck, USA
Elena R. Alvarez-Buylla, Universidad Nacional Autónoma de Mexico, Mexico
Teresa Colombo, Institute for Computing Applications "Mauro Picone," Italy

***Correspondence:**
Kris Vissenberg, Plant Growth and Development, Department of Biology, University of Antwerp, Groenenborgerlaan 171, Antwerp 2020, Belgium
e-mail: kris.vissenberg@uantwerpen.be;
T. Charlie Hodgman, Centre for Plant Integrative Biology, School of Biosciences, University of Nottingham, College Road, Sutton Bonington, LE12 5RD, UK
e-mail: charlie.hodgman@nottingham.ac.uk

†Present address:
Iben Sørensen, Department of Plant Biology, Cornell University, Ithaca, New York, USA

‡These authors have contributed equally to this work.

Plant cell wall composition is important for regulating growth rates, especially in roots. However, neither analyses of cell wall composition nor transcriptomes on their own can comprehensively reveal which genes and processes are mediating growth and cell elongation rates. This study reveals the benefits of carrying out multiple analyses in combination. Sections of roots from five anatomically and functionally defined zones in *Arabidopsis thaliana* were prepared and divided into three biological replicates. We used glycan microarrays and antibodies to identify the major classes of glycans and glycoproteins present in the cell walls of these sections, and identified the expected decrease in pectin and increase in xylan from the meristematic zone (MS), through the rapid and late elongation zones (REZ, LEZ) to the maturation zone and the rest of the root, including the emerging lateral roots. Other compositional changes included extensin and xyloglucan levels peaking in the REZ and increasing levels of arabinogalactan-proteins (AGP) epitopes from the MS to the LEZ, which remained high through the subsequent mature zones. Immuno-staining using the same antibodies identified the tissue and (sub)cellular localization of many epitopes. Extensins were localized in epidermal and cortex cell walls, while AGP glycans were specific to different tissues from root-hair cells to the stele. The transcriptome analysis found several gene families peaking in the REZ. These included a large family of peroxidases (which produce the reactive oxygen species (ROS) needed for cell expansion), and three xyloglucan endo-transglycosylase/hydrolase genes (XTH17, XTH18, and XTH19). The significance of the latter may be related to a role in breaking and re-joining xyloglucan cross-bridges between cellulose microfibrils, a process which is required for wall expansion. Knockdowns of these XTHs resulted in shorter root lengths, confirming a role of the corresponding proteins in root extension growth.

Keywords: root growth, plant cell walls, multiomics, transcriptomics, localisomics, epitomics, cell-wall polysaccharides, cell elongation

INTRODUCTION

The plant kingdom displays an enormous diversity in shapes and sizes, varying from unicellular algae with a simple rather spherical morphology to very complex multicellular organisms that can reach more than 100 m in height. Growth of plants is the sum of two processes, namely the increase in cell number by repeated cycles of cell division and the subsequent—sometimes major—increase in volume of these newly formed cells by expansion. Both processes are controlled by the action of plant hormones among which auxin plays a major role (Perrot-Rechenmann, 2010). In roots, cells pass sequentially through different developmental stages along the root axis. Growth occurs through rapid elongation of cells in a zone shootward to the root apical meristem, which is the site of cell division, and before further cell type differentiation, for example root hair emergence and lateral organ initiation (Verbelen et al., 2006).

Plant cell walls are rigid yet deformable materials, and growth is seen as the irreversible increase in surface area of cell walls. This process requires an internal turgor pressure, which arises from water uptake into the cell. Turgor exerts a force against surrounding cell walls, but is otherwise mediated by changes in the mechanical properties of the walls, resulting in stress relaxation (Ray et al., 1972; Cosgrove, 1986, 1993, 2005; Guerriero et al., 2014). Cell walls are made up of a fibrillar component, the cellulose microfibrils, that is embedded in a highly hydrated matrix of pectins, which principally comprise homogalacturonan (HGA), rhamnogalacturonan-I (RG-I), and rhamnogalacturonan-II (RG-II). The tethering of the adjacent cellulose microfibrils occurs primarily through xyloglucan in dicotyledonous and non-commelinid monocotyledonous walls (Hayashi, 1989), or by glucuronoarabinoxylan (Nishitani and Nevins, 1991; Carpita and Gibeaut, 1993) and mixed-linkage (1-3),(1-4)-β-D-glucans in the walls of Poales and Equisetales (Kato et al., 1982; Scheller and Ulvskov, 2010; Mohler et al., 2013).

As well as pectin and hemicelluloses, several different classes of glycoproteins and enzymes are also present (Albenne et al., 2009). This complex composition results in the mechanical properties of the cell wall and greatly influences its growth potential. The volume increase of a plant cell was described in the Lockhart equation (1965), $dV/dT = \phi\ (P–Y)$, where (P–Y) is the turgor above a yield threshold Y that must be exceeded before plastic wall extension can occur, and ϕ is the extensibility coefficient that represents the time-dependent yielding properties of the cell wall in the direction of growth (Schopfer, 2006). Several proteins that can influence the cell wall's yielding parameters have been described, including expansins (McQueen-Mason et al., 1992), xyloglucan endotransglucosylase/hydrolases (XET/XTHs; Nishitani and Vissenberg, 2007; Miedes et al., 2013), peroxidases (Passardi et al., 2006), β(1-4)-glucanases (Labrador and Nevins, 1989), yieldins (Okamoto-Nakazato et al., 2001), and lipid transfer proteins (LTPs; Nieuwland et al., 2005).

With the emergence of different molecular biological approaches and tools, many genes that encode enzymes with a role in the synthesis of the various cell wall polysaccharides and proteins found in cell walls have been identified and their mutants described (e.g., Harholt et al., 2010; Carpita, 2011; Mewalal et al., 2014). The synthesis of complete cell-wall components, their trafficking and final assembly in cell walls are, however, very complex (McCann and Rose, 2010) and there is often not a simple link between genotype and a growth phenotype. Furthermore, the content, architecture and biophysical characteristics of the walls of a cell change at any point along the growth axis as a consequence of both the cell's history (i.e., the multiple processes since the cell originated) and the needs of its current location.

As a result of this complexity, point measurements are extremely difficult to interpret and the use of a single "omics" technique to uncover cell-wall processes underpinning plant growth might not be sufficient (Somerville et al., 2004; Farrokhi et al., 2006). Therefore, model organisms, such as *Arabidopsis thaliana*, and appropriate research tools are needed. The *A. thaliana* root has a relatively simple anatomy and develops in a highly predictable manner (Dolan et al., 1993), lending itself to investigation of growth mechanisms, and their regulation as evidenced by numerous reports (e.g., Ubeda-Thomás et al., 2009; Band et al., 2011; Bruex et al., 2012; De Rybel et al., 2012). In addition, its genome sequence is published (Arabidopsis Genome Initiative, 2000) and many research tools already exist (e.g., Fukao et al., 2013; Jacques et al., 2013; Moussaieff et al., 2013).

We used a combination of point measurements and three techniques to characterize the different developmental zones along the *A. thaliana* root, looking at cell wall composition by means of quantitative assessment of cell-wall epitopes (epitomics), epitope localization (localisomics), and gene expression (transcriptomics), and combined the -omics data in this study to provide an integrated perspective. This revealed that individual omics-techniques are inadequate and can even result in misleading conclusions. In contrast, the multi-omics approach has identified three gene families that appear to play a role in regulating root growth, and mutant analysis for one of these families (XTHs) supports these findings.

MATERIALS AND METHODS

PLANT MATERIAL AND HANDLING

Seeds of *Arabidopsis thaliana* (L.) Heynh. (ecotype Columbia-0) were surface-sterilized by incubation in 5% (v/v) sodium hypochlorite for 5 min, washed three times in sterile water and sown on vertical 125 × 125 mm square Petri plates. Each plate contained 60 ml 1/2 strength Murashige and Skoog media (Sigma) solidified with 1% (w/v) agar. For material used for transcriptomic and glycan microarray profiling (epitomics), sterile 9 × 9 cm square sections of 100 µm nylon mesh (Clarcor) were placed onto the media surface before sowing to facilitate root dissection and harvesting of cut sections. After 2 days at 4°C, plates were transferred to controlled-environment chambers at 23°C under continuous light at a photon flux density of 150 µmol m^{-2} s^{-1} for 7 days.

Roots were dissected into five sections as shown in **Figure 1**: (1) meristem (from the root tip to the top of the lateral root cap, approximately 350 µm from the tip); (2) rapid elongation zone (from the top of the lateral root cap to the first visible root hair bulge, approximately 850 µm from the shootward boundary of zone 1); (3) late elongation (deceleration) zone (from the first root hair bulge to the first fully elongated root hair); (4) mature root (500 µm shootward of the first fully elongated root hair); and the lateral root zone (2.5 cm in length, from the shootward boundary of zone 4 in a shootward direction). Dissected samples were immediately frozen in liquid nitrogen.

EPITOMICS

Techniques involving glycan microarrays were used for this (Shin et al., 2005; Moller et al., 2007, 2008). Cell wall material was isolated from dissected material as alcohol-insoluble residue (AIR). Frozen material was ground in liquid nitrogen using a micro-pestle, 1 ml of 70% (v/v) ethanol was added to each tube and the mixture shaken at 4°C for 1 h. The mixture was centrifuged (5 min, 10,000 ×*g*) and the supernatant discarded. This process was repeated 4 times. Pellets were resuspended in 1 ml of acetone for 5 min and then air-dried overnight. A total of 600 roots were dissected, yielding 10–20 mg AIR for each of the

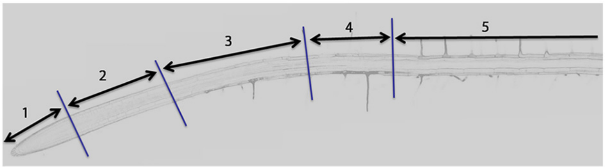


FIGURE 1 | Overview of the longitudinal sections used for the whole-genome transcript and epitomic analyses. (1) Meristem (MS); (2) rapid elongation zone (REZ); (3) late elongation zone (LEZ); (4) mature zone (MZ); (5) lateral root zone (LRZ). Image from De Rybel et al. (2010).

five sections, which were subjected to three sequential extractions as previously described (Sørensen and Willats, 2011). Briefly, the extractions were performed using 15 µl each of 50 mM diamino-cyclo-hexane-tetra-acetic acid (CDTA), 4 M sodium hydroxide (NaOH) with 0.1% v/v sodium borohydride ($NaBH_4$) and cadoxen [31% (v/v) 1,2-diaminoethane with 0.78 M cadmium oxide (CdO)]. These respectively enrich for pectin, hemicelluloses, and cellulose-associated molecules. The extracted fractions were printed as microarrays with six replicates and three dilutions using a Microgrid II microarray robot (Genomic Solutions, Ann Arbor, MI, USA) and the arrays were probed with a range of primary antibodies or carbohydrate-binding modules and appropriate alkaline phosphatase (AP) conjugated secondary antibodies before developing as previously described (Sørensen and Willats, 2011). All JIM- and LM-monoclonal antibodies and 2F4 were obtained from PlantProbes (http://www.plantprobes.net), CCRC antibodies from CarboSource (http://www.ccrc.uga.edu/~carbosource/CSS_home.html) and secondary antibodies from Sigma-Aldrich (http://www.sigmaaldrich.com). The arrays were scanned and analyzed using the microarray software ImaGene 6.0 (http://www.biodiscovery.com) to obtain raw signal values. These were then treated in the same way as fluorescent microarray data. Specifically, the value from the negative control where the secondary antibody was omitted was subtracted, and then the median and median-absolute deviation values were calculated. A heatmap was produced using Microsoft Excel.

LOCALISOMICS

Four-day-old seedlings were fixed and prepared for whole-mount immunolocalization analyses requiring some cell-wall permeabilization steps as described previously (Peret et al., 2012). Cell wall antibodies were used at 1:100 dilutions, whereas Alexafluor488 or Alexafluor543 coupled anti-rat or anti-mouse secondary antibodies were used at 1:200 dilutions to give green or red fluorescence, respectively. Counter staining was performed using either propidium iodide (for the AlexaFluor488 coupled secondary antibody) or Sytox Green (for AlexaFluor543 coupled secondary antibody). Seedlings were mounted in 50% glycerol and images were taken using a Leica SP2 confocal laser scanning microscope (Leica Microsystems UK Ltd). The specific antibodies, their epitopes and localization (primarily as determined in this study) are listed in **Table 1**.

TRANSCRIPTOMICS

Three biological replicates from separate pools of seeds were used. For each biological replicate, plants were grown and approximately 50 roots dissected as described in Section Plant Material

Table 1 | Cell wall epitopes assessed in the glycan microarray and their localization from *in situ* fluorescence studies.

Antibody	Epitope	Localization	References/image
CAL	(1→3) β-glucan	n/a	
CCRC-M1	Xyloglucan	All walls	**Figure 3** and Freshour et al., 2003
CBM22	Xylan	Secondary cell walls	McCartney et al., 2006
2F4	Calcium-stabilized homogalacturonan chains	n/a	
LM1	Extensin (HRGP)	No signal	
LM2	AGP (β-linked glucuronic acid)	Mainly lateral root cap and young epidermis; in meristem some cell-wall plates in epidermis and a stele cell file	Supplementary Figure S1, pp. 12–17
LM5	(1→4)-β-D-galactan	All cytoplasm and walls, especially of epidermis and stele	Supplementary Figure S1, p. 18 and McCartney et al., 2003
LM6	(1→5)-α-L-arabinan/AGPs	Epidermis and lateral root cap; higher up root localized to patches possibly forming diagonal stripes	**Figure 3** and Talboys et al., 2011
LM8	Xylogalacturonan	Mainly lateral root cap	Supplementary Figure S1, p. 21 and Willats et al., 2008
LM10	(1→4)-β-D-xylan	No signal	
LM15	XXXG motif of xyloglucans	Quiescent center and mature epidermis, especially at interface between cell files and root hairs	**Figure 3** and Larsen et al., 2014
JIM5	Partially methylesterified-homogalacturonan	Mainly lateral root cap, walls of quiescent center and initial cells, and some cell-division plates	**Figure 3**
JIM7	Partially methylesterified-homogalacturonan	No signal	
JIM8	AGP glycan	No signal	
JIM13	AGP glycan	Stele files especially zones 2–5, faint signal in epidermis	**Figure 3** and Dolan and Roberts, 1995
JIM19	Extensin	No signal	
JIM20	Extensin	Mainly epidermis and cortex from zone 2 upwards, some in lateral root cap	**Figure 3**
MAC207	AGP glycan	n/a	

and Handling. RNA was extracted using the Qiagen MicroRNA Kit following the manufacturer's instructions (Qiagen, Crawley, UK) and quantified using a Nanodrop ND100 spectrophotometer (Nanodrop, Wilimington, USA). All RNA samples were approximately 50 ng μl^{-1} in a total volume of 10 μl. Labeling of RNA samples was conducted using the Affymetrix IVT-Express Eukaryotic Target Labeling Assay kits following standard Affymetrix protocols (Affymetrix UK Ltd., High Wycombe, UK). RNA labeling and hybridization to Affymetrix ATH1 arrays were performed by the Nottingham Arabidopsis Stock Centre (NASC).

Data were normalized from.cel files with RMA and statistical tests performed using the Limma package in R/Bioconductor (Smyth, 2005) and a custom CDF file (ATH1121501_At_TAIRT v17; Dai et al., 2005). A gene was considered to be expressed if its expression was greater than 100 and differentially expressed if a t-test between two zones was significant at a q-value of 0.05 after Benjamini and Hochberg false discovery rate correction (Smyth, 2005). Further analyses were performed using Excel 2010 (Microsoft Corporation, Redmond, USA). Genes were further annotated into cell-wall functional subclasses using Cell Wall Navigator (Girke et al., 2004) and PlnTFDB (Pérez-Rodrłguez et al., 2010). Transcriptomics data used in these experiments have been made available through ArrayExpress (www.ebi.ac.uk) with accession number E-MEXP-2912.

xth MUTANT ANALYSIS

Atxth17-1(SALK_015077), *Atxth19-1* (SALK_034274), *Atxth20-1*(SAIL_575_H09), and XTH18-RNAi were kindly provided by Prof. K. Nishitani (Tohoku University, Japan). *Atxth17-2* (SALK_008429), *Atxth19-2* (SAIL_62_A10), and *Atxth20-2* (SALK_066689) were obtained from NASC. All lines are in the Colombia-0 background. To assess basal root growth, the root length of seedlings grown vertically for 7 days was measured from the hypocotyl to the root tip. Root lengths were measured using the NeuronJ plugin of ImageJ 1.4.1j (http://rsb.info.nih.gov/ij/). Two-tail Student t-tests were performed using Excel 2010 to determine significance (p-value < 0.05).

Confocal microscopy for imaging of *A. thaliana* roots was performed using a Leica SP5 confocal laser-scanning microscope (Leica, Milton-Keynes, UK). For cell quantification and cell length measurements, seedlings were treated with propidium iodide (10 μg ml^{-1}; Sigma) to visualize cell walls. Cell lengths were measured using the Cell-o-Tape image-analysis tool (French et al., 2012). Data are presented as the mean ± the standard error and two-tail Student t-tests, used to determine significance (p-value < 0.05), were performed using Microsoft Excel software.

RESULTS

EPITOMICS

Figure 2 depicts the epitope intensities in the different fractions and root zones, and highlights the zonal difference between antibodies to related epitopes. The signals corresponding to the pectin I and xyloglucan (XyG) binding antibodies, suggests that the CDTA and NaOH treatments were effective in terms of extracting the associated polysaccharides (pectin and hemicelluloses, respectively), since it was expected that some cellulose-associated XyG would be extracted using cadoxen due to the fact that XyG tethers adjacent cellulose microfibrils.

The more soluble (1→3)-β-glucans, recognized by CAL, peaked in zones 2 and 4 in the hemicellulosic fraction while in the pectin fraction they actually peaked in zone 5, consistent with the highest pectin-associated signal of crystalline cellulose. The xylans (recognized by LM10 and CBM22) were present in all three fractions. The CBM22 signal peaked in zones 1–3 in the pectin fraction, in zones 2–4 in the hemicellulosic fraction and in zone 5 in the cellulosic fraction. LM10, however, showed an increasing signal in the hemicellulosic fraction as the root matured, a trend that was also present in the cellulosic fraction where it was accompanied by a reduction in zones 2 and 3.

Xyloglucans (probed by CCRC-M1 and LM15) peaked in the REZ and were specific to the hemicellulosic fraction, with only minor binding in the cellulosic fraction, a pattern becoming more evident as the root matures. The homogalacturonan/pectin I epitope, probed by 2F4, JIM5, and JIM7, which differ in their sensitivity to the degree of esterification, shows a similar pattern and is specific to the pectin fraction. JIM7 peaked in zones 1 and 2, JIM5 peaked in zone 2, and 2F4 peaked in zones 2 and 3.

The pectin II epitopes show a broad range of patterns, consistent with the complexity of its polymer constituents. LM6, specific for arabinan, peaked in the hemicellulosic fraction in zone 1 and rapidly decreased as the root matures. LM8, specific for xylogalacturonan, peaked in zones 1 and 2 in the same fraction and to a lesser extent in the cellulosic fraction before trailing off rapidly after zone 3, whilst in the pectin fraction it was consistently present at a low level throughout. LM5, recognizing galactan, shows one of the most complex signal patterns, peaking in zones 2–3 in the pectin fraction, zone 2 in the hemicellulosic fraction while showing the inverse pattern in the cellulosic fraction.

Three antibodies detect variants from the extensin family. The LM1 and JIM20 epitopes are present at high levels in all zones, with the JIM20 signal being higher in the hemicellulosic fraction, peaking in zone 2, and the LM1 signal being nominally highest in zone 5. The JIM19 epitope, however, was absent from the meristem, was mostly pectin-associated in the elongation zones, and showed its highest signal in zones 4 and 5, but was specifically non-pectin associated. Several antibodies recognize different epitopes associated with AGPs. Their signal levels tended to be higher from zone 2 onwards, but they were also found in all extractions and developmental zones.

LOCALISOMICS

Figure 3 contains localization results for four classes of cell wall-related epitopes, namely XyGs, pectins, AGPs, and extensins. Crucially, the images show that antibodies were able to bind to epitopes throughout the root, rather than only at the surface, giving us confidence that cell-wall epitope localization can be obtained when using permeabilization procedures that will lead to the loss of some cell wall structures. However, five antibodies were not available for localization studies and a further five gave no signal, possibly due to the epitope being modified beyond recognition by the permeabilization process. The full image dataset can be found in the Supplementary Material

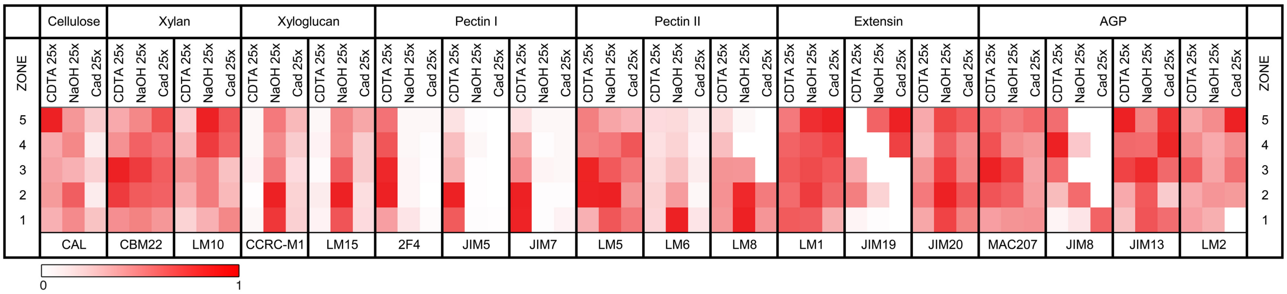


FIGURE 2 | Epitomic heatmap. The results for each antibody are scaled relative to the maximum signal for that antibody. The heatmap is supplemented by Supplementary Figure S2, which shows the actual signals in graph form for each antibody. Pectin I corresponds to homogalacturonan epitopes and Pectin II to rhamnogalacturonans.

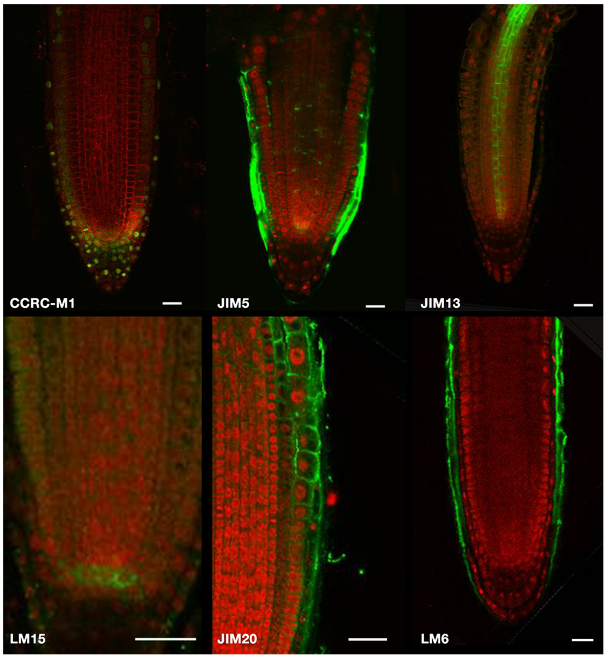


FIGURE 3 | Immunolocalization with key antibodies. CCRC-M1 and LM15 detect xyloglucans, JIM5 detects pectins, JIM20 detects extensins, and JIM13 and LM6 bind to AGPs. An AF488-linked secondary antibody was used for all antibodies except CCRC-M1 for which an AF543-linked antibody was used. The scale bars correspond to 100 μm for all images except LM15 for which it is 25 μm.

Table 2 | Transcriptomic data for the five zones of the *A. thaliana* root.

	Genes expressed (% of 21,331)	Significantly up-regulated	Significantly down-regulated
Zone 1	7741 (36.3%)		
Zone 1–2		1632	1939
Zone 2	7539 (35.3%)		
Zone 2–3		1343	1750
Zone 3	7801 (36.6%)		
Zone 3–4		155	369
Zone 4	7854 (36.8%)		
Zone 4–5		155	184
Zone 5	8044 (37.7%)		
Zone 5–1		1322	2627

A gene was considered to be expressed if its expression was greater than 100 and differentially expressed if a t-test between two zones was significant at a q-value of 0.05.

(Supplementary Figure S1) and epitope localization arising from this dataset is described in **Table 1**.

The XyGs recognized by LM15 were present at low levels throughout, but with a particularly high signal in the quiescent center, although no such specificity was seen with the CCRC-M1 antibody, recognizing fucosylated XyG. JIM5 showed the partially methylesterified pectins to be localized to the lateral root cap and radial walls in the inner tissues, while the extensins detected by JIM20 were most prominent in the epidermis and cortex of the REZ and to a lesser degree in the lateral root cap. JIM13 and LM6 binding patterns suggested that individual AGPs might be specific to different locations, suggesting that they play specific roles in the cells where they are expressed but that their general role in root growth may be difficult to interpret. Finally, binding of the LM8 antibody, recognizing xylogalacturonan, was restricted to the lateral root cap in a manner reminiscent of LM6 (see Supplementary Material).

TRANSCRIPTOMICS

A large percentage of the genome is expressed in the root with about 7500–8000 genes detectably expressed in any given root zone (See **Table 2** and Supplementary Table 1). Zones 1–3, despite their physical proximity, show large inter-zonal differences consistent with their very different developmental roles. Zones 3–5 have more similar gene expression levels, highlighting the general developmental quiescence of the mature root. However, ∼10% of significantly differentially expressed genes in these zones are annotated as cell-wall genes, which is twice the number of genes changing between the earlier zones. It is the changes between zones 1–2 and 2–3 that are likely to inform the expression changes that allow rapid elongation and deceleration, respectively. At their peak, 52% of reactive oxygen species (ROS) and 63% of aquaporin genes are expressed in the root. Only 25% of annotated transcription factors are detectably expressed in the root, which is lower than other annotated gene families—possibly reflecting tight developmental differentiation.

DISCUSSION

The *A. thaliana* root develops in a highly predictable manner. Cells pass through consecutive developmental phases during which the post-mitotic elongation of cells contributes the majority of the increase in the root length. Cell wall metabolism is very important in allowing and controlling cellular expansion. The synthesis, trafficking, deposition, integration, and remodeling pathways of cell-wall components are, however, very complex and not completely resolved, so multiple omics-approaches are needed to establish which genes are contributing to the observed changes in composition and consequent mechanical properties.

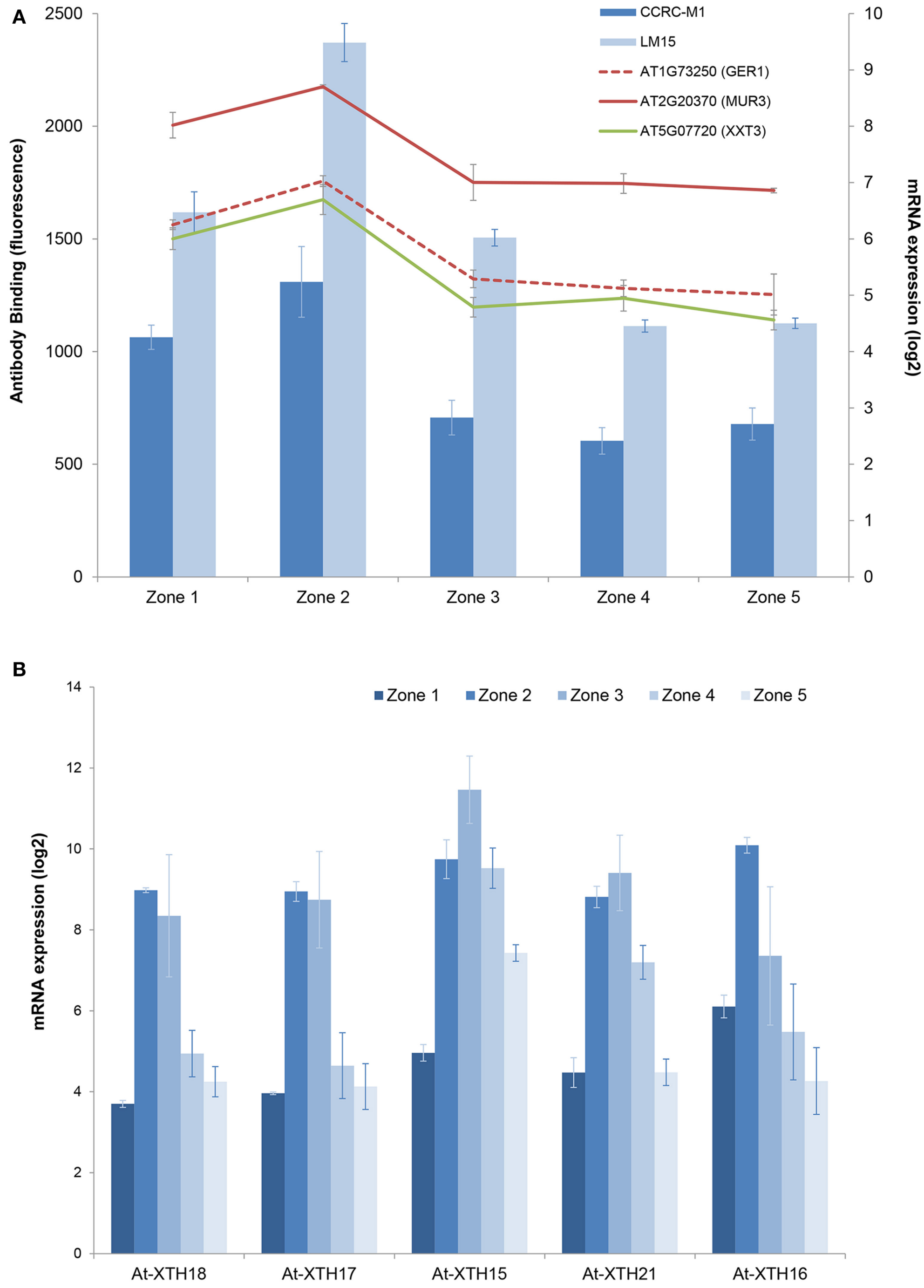


FIGURE 4 | Combined xyloglucan, XyG biosynthesis genes, and XTH expression profiles. (A) epitomic (columns) and XyG biosynthesis mRNA expression (lines). **(B)** zonal transcriptomic expression profiles for selected members of the XTH family. Error bars are ±1 SD.

These issues are complicated further by the observation that the different tissues differentially contribute to the mechanics of root elongation (Dyson et al., 2014).

EPITOMICS ANALYSIS

Epitomics (also known as glycan microarray analysis and previously described as Comprehensive Microarray Polymer Profiling) is a high-throughput microarray-derived technique that allows the handling of multiple samples. In this study, it was used to indicate which polysaccharides and glycoproteins were found in specific root developmental-zones. Many of the results support current knowledge of root cell-wall composition, for example that pectin is synthesized and deposited into the existing cell wall in a highly esterified form (Liners and Van Cutsem, 1992) and that pectinesterases modify the pectins while the cells age (Micheli, 2001). Carbohydrate polymer synthesis is complicated and does not necessarily correlate well with wall composition. Little is known of extant synthesis pathways, let alone of any differences in synthesis along developmental zones. In addition, the spatial resolution of this experimental approach is rather poor.

Table 3 | Mean root lengths in wild type and *xth* mutants, measured at 7 days after germination, asterisks denote significance at a $p > 0.05$.

	Length (mm ± *SEM*)	% vs. Col-0	*n*
SINGLE MUTANTS			
Col-0	37.87 (± 0.39)	–	71
xth17-1	34.01 (± 0.40)*	89.8%	44
xth17-2	34.09 (± 0.46)*	90.0%	39
xth18-RNAi	32.75 (± 0.65)*	86.5%	26
xth18-2	38.74 (± 0.45)	102.3%	40
xth19-1	41.31 (± 0.29)*	109.1%	47
xth19-2	32.36 (± 0.48)*	85.5%	19
xth20-1	35.17 (± 0.48)*	92.9%	41
xth20-2	33.73 (± 0.57)*	89.1%	36
DOUBLE MUTANTS			
Col-0	38.02 (± 0.39)	–	69
17-1×18-1	32.09 (± 0.60)*	84.4%	25
17-2×18-1	27.77 (± 0.61)*	73.0%	31
17-2×19-1	33.63 (± 0.54)*	88.4%	25
19-1×20-2	34.02 (± 1.21)*	89.5%	20
17-1×20-1	35.29 (± 0.40)*	92.8%	40
18-1×20-1	36.13 (± 0.73)*	95.0%	24

LOCALISOMICS ANALYSIS

In contrast to epitomics, whole-mount immunolocalization provides high spatial resolution and revealed cell and tissue-specific locations for some cell wall epitopes that would not have been clear from the epitomic data alone. The AGPs appear to be the most remarkable in this respect, suggesting that individual proteins play a specific and subtly different role. The localization of extensins to zone 2 epidermis and cortex is intriguing, given the predictions of Dyson et al. (2014) that the outer layers of the root have most influence on growth. Unfortunately, some of the antibodies were not available for this technique and others did not yield any signal in the cell walls, probably because of masking of the epitope by other cell wall components or more likely due to the use of enzymes in the permeabilization procedures required for whole-mount preparations. This -omics has refined the roles of certain epitopes, but cannot be used to link to specific genes.

TRANSCRIPTOMICS ANALYSIS

Analysis of the transcriptomic data in isolation can also be misleading. Three gene families account for nearly as much of the expression as all the other cell-wall-related families together. The largest cell wall family is the AGPs with 61 members representing more than 18% of cell wall-related gene expression, but the role of individual genes in expansion and maturation is unclear as the localization shows that different epitopes are found in different places.

Aquaporins play a role in vacuolar filling and contribute to turgor pressure, the driving force of expansion. Their expression is present in the elongation zone (zones 2 and 3), naively suggesting that expansion is effected by an increase in turgor pressure. However, recent work (Dyson et al., 2014) shows that pressures remain constant, implying that the role of the aquaporins is to ensure that turgor pressure is not lost by the rapid expansion in cell size.

The third highly-expressed gene family encodes peroxidases, which contains members that are involved in the generation of ROS. These have long been implicated in root growth and development (Gapper and Dolan, 2006; Manzano et al., 2014). However, our data (Supplementary Table 1) show that their expression rises in zone 2, peak in zone 3 and remain high in the later zones. This suggests a role in maturation rather than

Table 4 | Effect of reduced XTH17 and XTH18 expression on root growth.

	Growth rate (mm/h)					Average mature cell length (μm)
	3 dag	4 dag	5 dag	6 dag	7 dag	
Col-0	0.086 (± 0.003)	0.219 (± 0.013)	0.314 (± 0.018)	0.370 (± 0.013)	0.407 (± 0.021)	199.18 (± 1.10)
xth17-2	0.075 (± 0.006)*	0.206 (± 0.016)	0.255 (± 0.024)*	0.354 (± 0.026)	0.367 (± 0.030)	184.42 (± 7.88)*
XTH18-RNAi	0.070 (± 0.005)*	0.172 (± 0.016)*	0.286 (± 0.020)	0.359 (± 0.015)	0.357 (± 0.016)*	189.66 (± 6.78)*
xth17-2 × XTH18-RNAi	0.054 (± 0.004)*	0.193 (± 0.017)*	0.240 (± 0.025)*	0.241 (± 0.019)*	0.319 (± 0.052)*	182.06 (± 7.66)*

Growth rates were measured at the indicated days after germination (dag) using NeuronJ. Cortical cell lengths were measured from confocal microscope images, asterisks denote significance at a $p > 0.05$.

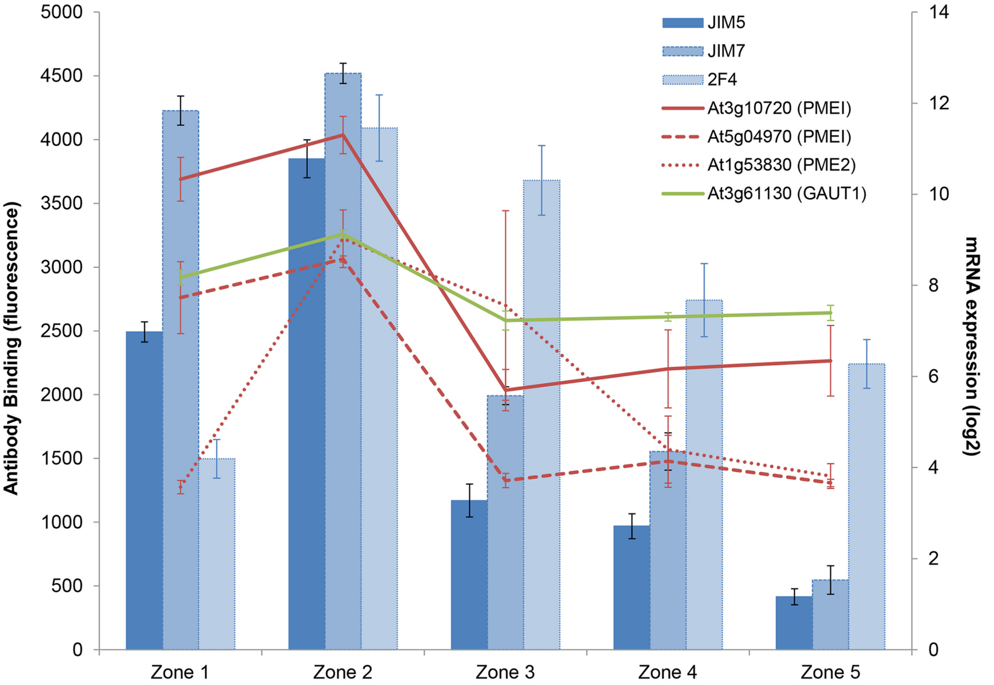


FIGURE 5 | Combined profiles of pectin epitopes, GAUT1, PME, and PMEI expressions. Epitomic (columns) and biosynthesis (GAUT1) and modification (PME and PMEI) mRNA expression (lines). Error bars are ±1 SD.

elongation, and also possibly in lignification and Casparian strip formation. Although the other -omics data indicated a role for extensins, the genes for these proteins are difficult to define because sequences/functions overlap with other gene families. Hence they were not considered further in this context. Initial carbohydrate biosynthesis peaks in zone 2, where large amounts are needed for both accelerating and subsequent decelerating expansion, but in order to get enough RNA these transcriptomic data face the same issue of resolution as the epitomics work.

MULTI-OMICS ANALYSIS

As mentioned above, all these techniques in isolation have their benefits and drawbacks. We therefore combined data from all analysis in a multi-omics approach to identify genes that play an important role in the elongation of *A. thaliana* root cells.

To investigate xyloglucans, we looked for expression patterns in the transcriptome that correlated with the epitope pattern shown by LM15 (peaking in the zone 2). The transcriptomic dataset showed a large number of genes with a similar pattern (Supplementary Table 1). However, filtering the data to include only genes known to affect XyG biosynthesis, we found genes involved in fucose biosynthesis (*MUR3*, AT2G20370 and *GER1*, AT1G73250) and a xyloglucan xylosyltransferase, (*XXT3*, AT5G07720) with R^2-values of 0.96–0.98, suggesting that these members from large gene families may be responsible for the observed LM15-XyG signal (**Figure 4**).

XyG is highest in zone 2, the rapid elongation zone, correlating with studies showing that *in vivo*, XET activity by XTHs is also highest in this zone (Vissenberg et al., 2000). The general XTH expression signal peaks in zone 2 and 3 (**Figure 4**), suggesting an important role for XTHs in root elongation. XTHs remodel the cell wall and some are believed to promote growth acceleration (Van Sandt et al., 2007), while others could aid in the deceleration. Different family members have distinct activity dependencies and pH optima, suggesting that some can act as loosening factors, while others could do the opposite (Maris et al., 2009, 2011). If they were only associated with accelerated cell wall actions, then the plant might override the loosening in the deceleration zone, probably by pectin modifications (Micheli, 2001) or peroxidase-mediated cross-linking of other cell-wall components such as structural proteins (Ma et al., 2004; Passardi et al., 2004; De Cnodder et al., 2005). The key XTHs controlling expansion might be expected to be expressed as early as possible to tightly regulate wall extensibility and disruption of these XTHs may therefore show growth defects.

Looking for XTHs with a peak very early in root development and associated with gibberellic acid (GA), a known regulator of cell expansion (Middleton et al., 2012), showed that XTH17 and 18 increase more than 30 fold between zones 1 and 2. One of these is GA-induced (XTH17) and both belong to a subclade of group 1 XTHs (Rose et al., 2002) consisting of 4 genes (XTH17, 18, 19, and 20). XTH17, 18, and 19 are expressed in the elongation zones of the root, and XTH20 is expressed in the vascular tissue of the mature root (Vissenberg et al., 2005). We therefore investigated phenotypes in mutant lines for this subclade. Putative knock out (KO) lines were identified for three of the genes and an RNAi line was created for the fourth. All these lines showed significant growth defects with shorter mature root lengths, shorter mature

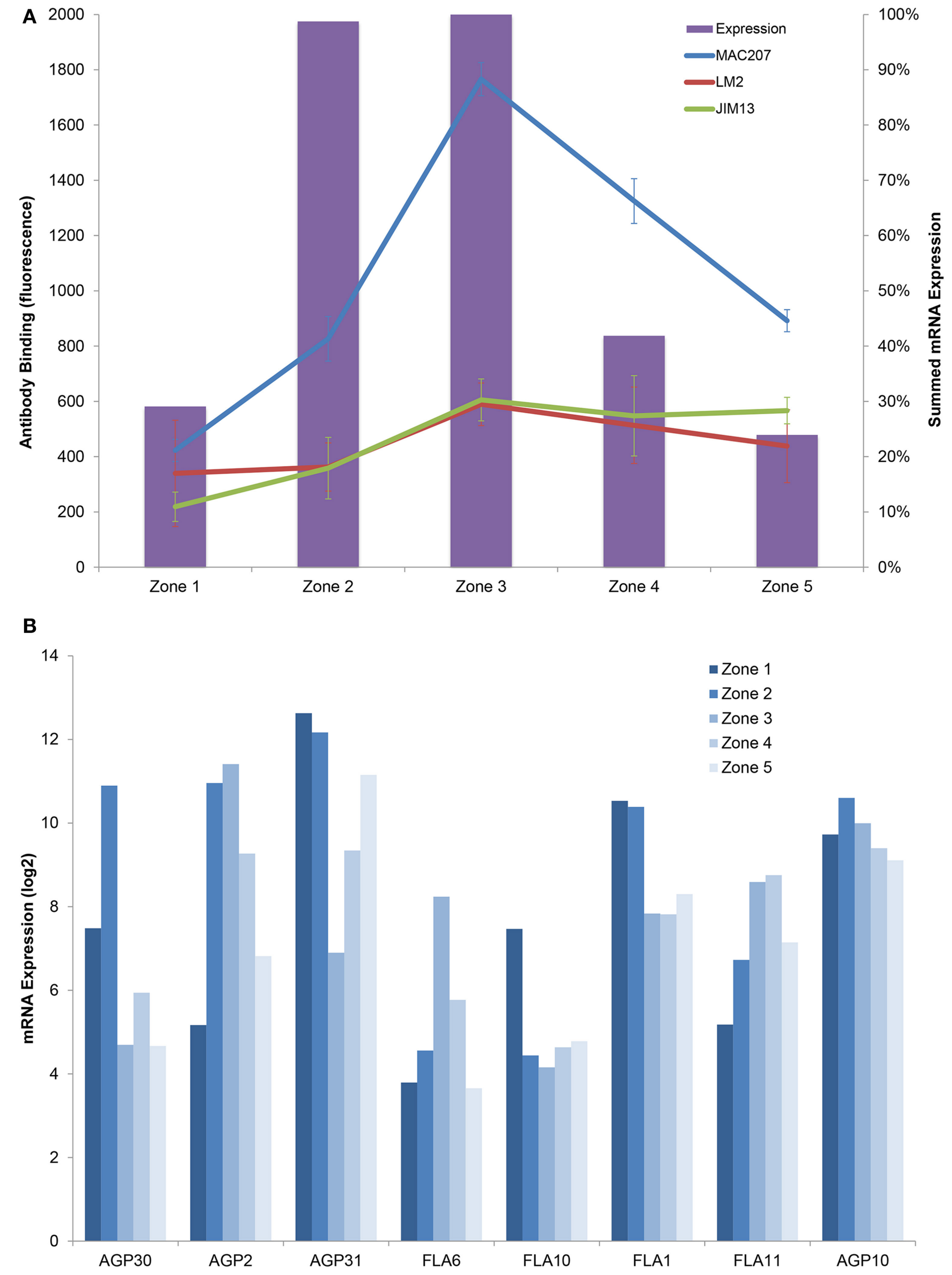


FIGURE 6 | Combined profiles of AGP epitopes and expression. (A) epitomic (lines) and mRNA expression (columns), expression is summed over all expressed AGPs and given as a percentage of the expression in each zone, error bars are ±1 SD. **(B)** Exemplars of the principal expression profiles for members of the AGP family.

root cells and reduced growth rates (**Tables 3**, **4**). The one exception to this was *xth19-1* which, in contrast to the *xth19-2* line, showed an increase in root length. The insertion in the *xth19-1* line is 3′ to the gene, which may stabilize the mRNA and hence act as an over-expression line. The reduction in growth phenotype was increased in the *xth17-2*xXTH18-RNAi double mutant (**Table 4**).

Regarding pectin (or more specifically homogalacturonan), JIM5, and JIM7 recognize esterified pectin, which is present at highest levels in zones 1 and 2 and drops off in the mature root. Pectin is deposited into the cell wall in a highly esterified form, and is typically a component of a loosened cell wall. 2F4 recognizes cross-linked pectin with no more than 40% esterification (Liners et al., 1992) and this is low in zone 1 but rises to high levels in zone 2 and beyond. This suggests zone 2 has a mixture of esterified and non-esterified forms and the distribution is increasingly biased toward non-esterified along the shootward axis. This follows the expression of several pectin methylesterases [PMEs; e.g., PME2 (AT1G53830)], some of which are highly correlated with the epitomic pattern (**Figure 5**).

Several PME Inhibitors (e.g., AT5G04970, AT3G10720) also correlate with this pattern (**Figure 5**), suggesting that the root uses a tight balance of these two groups of enzymes to control the rate of de-esterification (i.e., stiffening) rather than using one to shut the other off. With regard to pectin biosynthesis, the galacturonosyltransferase 1 enzyme (GAUT1, At3g61130) is expressed in zones 1 and 2 and then decreases, which correlates ($R^2 > 0.9$) with the pattern of the three antibody epitopes recognizing homogalacturonan (**Figure 5** and Supplementary Figure S2).

The transcriptomics data revealed multiple patterns for AGP expression, dominated by maximum levels in zones 2 and 3 (**Figure 6**). In contrast, the epitomic signals peak in zone 3, which could be accounted for by several hypotheses. One possibility is that there are delays in synthesis and transport either because of long synthesis and transport times of the proteins, or accumulation for use in zone 3 walls. Alternatively the epitomic profiles could simply be reflecting AGP accumulation over time.

Different AGPs may play different roles in the cell wall as revealed by the localisome, which was also mentioned before (Ellis et al., 2010). For example, LM6 (AGP or pectic arabinan) localizes to the lateral root cap and epidermis and its epitope may play a role in expansion, while the JIM13 epitope is specific to stele cell files and could be part of the process of vascular patterning.

OVERARCHING CONCLUSIONS

It is evident that the individual -omics approaches provide an incomplete picture, and a combination of multiple analyses aids in establishing a clearer picture of the processes involved. In general terms, the transcriptomic dataset suggests the location of cell wall synthesis, whereas the glycan microarray analyses show the accumulation and dilution as these polymers are modified or additional material added. This is best shown with regard to pectins and XyGs. Deceleration in the root elongation rate appears to be linked to a change in the ratio of esterified to non-esterified pectins. The mutation studies have confirmed a role for XTH17, XTH18, and possibly XTH19 in root growth, probably affecting yield threshold, as this was the case in dark-grown hypocotyl cells (Miedes et al., 2013).

POTENTIAL APPLICATION TO OTHER SYSTEMS

Root growth is a system in which the genes for synthesis of cell wall material are expressed in one location, while the molecules themselves might not appear in the wall until later. Different molecules contribute to different aspects of the cell-wall mechanics, which are further affected by subsequent modification and interaction. The multi-omics approach used in this study could be used for other plant structures and translated to other systems where the chronology of gene expression, macromolecular synthesis, and modification contribute to growth or mechanical properties of an organ as a whole. Potential applications include musculoskeletal growth, strength, and brittleness in health, aging and disease states, as well as plant lodging (i.e., the bending or even falling over of stalks leading to reduced crop yields).

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support of the UK BBSRC and EPSRC grant BB/D019613/1 (Centre for Plant Integrative Biology), Andrew French for valuable discussions about statistical data analysis, the University of Antwerp, the Research Foundation - Flanders (FWO-Vlaanderen), the Interuniversity Attraction Poles Programme – Belgian State – Belgian Science Policy [IUAP VI/33].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fcell.2015.00010/abstract

REFERENCES

Albenne, C., Canut, H., Boudart, G., Zhang, Y., San Clemente, H., Pont-Lezica, R., et al. (2009). Plant cell wall proteomics: mass spectrometry data, a trove for research on protein structure/function relationships. *Mol. Plant.* 2, 977–989. doi: 10.1093/mp/ssp059

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* 408, 796–815. doi: 10.1038/35048692

Band, L. R., Wells, D. M., Larrieu, A., Sun, J., Middleton, A. M., French, A. P., et al. (2011). Root gravitropism is regulated by a transient lateral auxin gradient controlled by a tipping-point mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4668–4673. doi: 10.1073/pnas.1201498109

Bruex, A., Kainkaryam, R. M., Wieckowksi, Y., Kang, Y. H., Bernhardt, C., Xia, Y., et al. (2012). A gene regulatory network for root epidermis cell differentiation in Arabidopsis. *PLOS Genet.* 8:e1002446. doi: 10.1371/journal.pgen.1002446

Carpita, N. C. (2011). Update on mechanisms of plant cell wall biosynthesis: how plants make cellulose and other (1→4)-β-d-glycans. *Plant Physiol.* 155, 171–184. doi: 10.1104/pp.110.163360

Carpita, N. C., and Gibeaut, D. M. (1993). Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J.* 3, 1–30. doi: 10.1111/j.1365-313X.1993.tb00007.x

Cosgrove, D. J. (1986). Biophysical control of plant cell growth. *Annu. Rev. Plant Physiol.* 37, 377–405. doi: 10.1146/annurev.pp.37.060186.002113

Cosgrove, D. J. (1993). Water uptake by growing cells: an assessment of the controlling roles of wall relaxation, solute uptake, and hydraulic conductance. *Int. J. Plant Sci.* 154, 10–21. doi: 10.1086/297087

Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nat. Rev. Mol. Cell Biol.* 6, 850–861. doi: 10.1038/nrm1746

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* 15, e175 doi: 10.1093/nar/gni179

De Cnodder, T., Vissenberg, K., Van Der Straeten, D., and Verbelen, J.-P. (2005). Regulation of cell length in the *Arabidopsis thaliana* root by the ethylene precursor 1-aminocyclopropane-1-carboxylic acid: a matter of apoplastic reactions. *New Phytol.* 168, 541–550. doi: 10.1111/j.1469-8137.2005.01540.x

De Rybel, B., Audenaert, D., Xuan, W., Overvoorde, P., Strader, L. C., Kepinski, S., et al. (2012). A role for the root cap in root branching revealed by the non-auxin probe naxillin. *Nat. Chem. Biol.* 8, 798–805. doi: 10.1038/nchembio.1044

De Rybel, B., Vassileva, V., Parizot, B., Demeulenaere, M., Grunewald, W., Audenaert, D., et al. (2010). A novel aux/IAA28 signaling cascade activates GATA23-dependent specification of lateral root founder cell identity. *Curr. Biol.* 20, 1697–1706. doi: 10.1016/j.cub.2010.09.007

Dolan, L., Janmaat, K., Willemsen, V., Linstead, P., Poethig, S., Roberts, K., et al. (1993). Cellular organisation of the *Arabidopsis thaliana* root. *Development* 119, 71–84.

Dolan, L., and Roberts, K. (1995). Secondary thickening in roots of *Arabidopsis thaliana*: anatomy and cell surface changes. *New Phytol.* 131, 121–128. doi: 10.1111/j.1469-8137.1995.tb03061.x

Dyson, R. J., Vizcay-Barrena, G., Band, L. R., Fernandes, A. N., French, A. P., Fozard, J. A., et al. (2014). Mechanical modelling quantifies the functional importance of outer tissue layers during root elongation and bending. *New Phytol.* 202, 1212–1222. doi: 10.1111/nph.12764

Ellis, M., Egelund, J., Schultz, C. J., and Bacic, A. (2010). Arabinogalactan-proteins: key regulators at the cell surface? *Plant Physiol.* 153, 403–419. doi: 10.1104/pp.110.156000

Farrokhi, N., Burton, R. A., Brownfield, L., Hrmova, M., Wilson, S. M., Bacic, A., et al. (2006). Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. *Plant Biotechnol. J.* 4, 145–167. doi: 10.1111/j.1467-7652.2005.00169.x

French, A. P., Wilson, M. H., Kenobi, K., Dietrich, D., Voß, U., Ubeda-Tomás, S., et al. (2012). Identifying biological landmarks using a novel cell measuring image analysis tool: Cell-o-Tape. *Plant Methods* 8:7. doi: 10.1186/1746-4811-8-7

Freshour, G., Bonin, C. P., Reiter, W. D., Albersheim, P., Darvill, A. G., and Hahn, M. G. (2003). Distribution of fucose-containing xyloglucans in cell walls of the mur1 mutant of Arabidopsis. *Plant Physiol.* 131, 1602–1612. doi: 10.1104/pp.102.016444

Fukao, Y., Yoshida, M., Kurata, R., Kobayashi, M., Nakanishi, M., Fujiwara, M., et al. (2013). Peptide separation methodologies for in-depth proteomics in Arabidopsis. *Plant Cell Physiol.* 54, 808–815. doi: 10.1093/pcp/pct033

Gapper, C., and Dolan, L. (2006). Control of plant development by reactive oxygen species. *Plant Physiol.* 141, 341–345. doi: 10.1104/pp.106.079079

Girke, T., Lauricha, J., Tran, H., Keegstra, K., and Raikhel, N. (2004). The cell wall navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. *Plant Physiol.* 136, 3003–3008. doi: 10.1104/pp.104.049965

Guerriero, G., Hausman, J.-F., and Cai, G. (2014). No stress! Relax! Mechanisms governing growth and shape in plant cells. *Int. J. Mol. Sci.* 15, 5094–5114. doi: 10.3390/ijms15035094

Harholt, J., Suttangkakul, A., and Scheller, H. V. (2010). Biosynthesis of pectin. *Plant Physiol.* 153, 384–395. doi: 10.1104/pp.110.156588

Hayashi, T. (1989). Xyloglucans in the primary cell wall. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 40, 139–168. doi: 10.1146/annurev.pp.40.060189.001035

Jacques, E., Buytaert, J., Wells, D. M., Lewandowski, M., Bennett, M. J., Dirckx, J., et al. (2013). MicroFilament Analyzer, an image analysis tool for quantifying fibrillar orientation, reveals changes in microtubule organization during gravitropism. *Plant J.* 74, 1045–1058. doi: 10.1111/tpj.12174

Kato, Y., Ito, S., Iki, K., and Matsuda, K. (1982). Xyloglucan and *r*β-D-glucan in cell walls of rice seedlings. *Plant Cell Physiol.* 23, 351–364.

Labrador, E., and Nevins, D. J. (1989). An exo-/β-D-glucan derived from Zea coleoptile walls with a capacity to elicit cell elongation. *Physiol. Plant.* 77, 479–486. doi: 10.1111/j.1399-3054.1989.tb05380.x

Larsen, E. R., Domzych, D. S., and Tierney, M. L. (2014). SNARE VTI13 plays a unique role in endosomal trafficking pathways associated with the vauole and is essential for cell wall organization and root hair growth in arabidopsis. *Ann. Bot.* 114, 1147–1159. doi: 10.1093/aob/mcu041

Liners, F., Thibault, J.-F., and Van Cutsem, P. (1992). Influence of the degree of polymerization of oligogalacturonates and of esterification pattern of pectin on their recognition by monoclonal antibodies. *Plant Physiol.* 99, 1099–1104. doi: 10.1104/pp.99.3.1099

Liners, F., and Van Cutsem, P. (1992). Distribution of pectic polysaccharides throughout walls of suspension-cultured carrot cells. *Protoplasma* 170, 10–21. doi: 10.1007/BF01384453

Lockhart, J. A. (1965). An analysis of irreversible plant cell growth. *J. Theor. Biol.* 8, 264–275. doi: 10.1016/0022-5193(65)90077-9

Ma, H., Tan, L., Kamyab, A., Hare, M., Shpak, E., and Kieliszewski, M. J. (2004). Di-isodityrosine is the intermolecular cross-link of isodityrosine-rich extensin analogs cross-linked *in vitro*. *J. Biol. Chem.* 279, 55474–55482. doi: 10.1074/jbc.M408396200

Manzano, C., Pallero-Baena, M., Casimiro, I., De Rybel, B., Orman-Ligeza, B., Van Isterdael, G., et al. (2014). The emerging roles of ROS signalling during lateral root development. *Plant Physiol.* 30, 1105–1119. doi: 10.1104/pp.114.238873

Maris, A., Kaewthai, N., Eklöf, J. M., Miller, J. G., Brumer, H., Fry, S. C., et al. (2011). Characterization of five recombinant xyloglucan endotransglucosylase/hydrolase (XTH) proteins of Arabidopsis reveals specific enzymatic properties. *J. Exp. Bot.* 62, 261–271. doi: 10.1093/jxb/erq263

Maris, A., Suslov, D., Fry, S. C., Verbelen, J.-P., and Vissenberg, K. (2009). Enzymic characterization of two recombinant xyloglucan endotransglucosylase/hydrolase (XTH) proteins of Arabidopsis and their effect on root growth and cell wall extension. *J. Exp. Bot.* 60, 3959–3972. doi: 10.1093/jxb/erp229

McCann, M., and Rose, J. (2010). Blueprints for building plant cell walls. *Plant Physiol.* 153, 365. doi: 10.1104/pp.110.900324

McCartney, L., Blake, A. W., Flint, J., Bolam, D. N., Boraston, A. B., Gilbert, H. J., et al. (2006). Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4765–4770. doi: 10.1073/pnas.0508887103

McCartney, L., Steele-King, C. G., Jordan, E., and Knox, J. P. (2003). Cell wall pectic (1→4)-β-D-galactan marks the acceleration of cell elongation in the *Arabidopsis* seedling root meristem. *Plant J.* 33, 447–454. doi: 10.1046/j.1365-313X.2003.01640.x

McQueen-Mason, S., Durachko, D. M., and Cosgrove, D. J. (1992). 2 Endogenous proteins that induce cell-wall extension in plants. *Plant Cell* 4, 1425–1433. doi: 10.1105/tpc.4.11.1425

Mewalal, R., Mizrachi, E., Mansfield, S. D., and Myburg, A. A. (2014). Cell wall-related proteins of unknown function: missing links in plant cell wall development. *Plant Cell Physiol.* 55, 1031–1043. doi: 10.1093/pcp/pcu050

Micheli, F. (2001). Pectin methylesterases: cell wall enzymes with important roles in plant physiology. *Trends Plant Sci.* 6, 414–419. doi: 10.1016/S1360-1385(01)02045-3

Middleton, A. M., Ubeda-Tomás, S., Griffiths, J., Holman, T., Hedden, P., Thomas, S. G., et al. (2012). Mathematical modeling elucidates the role of transcriptional feedback in gibberellin signaling. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7571–7576. doi: 10.1073/pnas.1113666109

Miedes, E., Suslov, D., Vandenbussche, F., Kenobi, K., Ivakov, A., Van Der Straeten, D., et al. (2013). Xyloglucan endotransglucosylase/hydrolase (XTH) overexpression affects growth and cell wall mechanics in etiolated Arabidopsis hypocotyls. *J. Exp. Bot.* 64, 2481–2497. doi: 10.1093/jxb/ert107

Mohler, K. E., Simmons, T. J., and Fry, S. C. (2013). Mixed-linkage glucan:xyloglucan endotransglucosylase (MXE) re-models hemicelluloses in Equisetum shoots but not in barley shoots or Equisetum callus. *New Phytol.* 197, 111–122. doi: 10.1111/j.1469-8137.2012.04371.x

Moller, I., Marcus, S. E., Haeger, A., Verhertbruggen, Y., Verhoef, R., Schols, H., et al. (2008). High-throughput screening of monoclonal antibodies against plant cell wall glycans by hierarchical clustering of their carbohydrate microarray binding profiles. *Glycoconj. J.* 25, 37–48. doi: 10.1007/s10719-007-9059-7

Moller, I., Sørensen, I., Bernal, A. J., Blaukopf, C., Lee, K., Øbro, J., et al. (2007). High-throughput mapping of cell-wall polymers within and between plants using novel microarrays. *Plant J.* 50, 1118–1128. doi: 10.1111/j.1365-313X.2007.03114.x

Moussaieff, A., Rogachev, I., Brodsky, L., Malitsky, S., Toal, T. W., Belcher, H., et al. (2013). High-resolution metabolic mapping of cell types in plant roots. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1232–E1241. doi: 10.1073/pnas.1302019110

Nieuwland, J., Feron, R., Huisman, B. A. H., Fasolino, A., Hilbers, C. W., Derksen, J., et al. (2005). Lipid transfer proteins enhance cell wall extension in tobacco. *Plant Cell* 17, 2009–2019. doi: 10.1105/tpc.105.032094

Nishitani, K., and Nevins, D. J. (1991). Glucuronoxylan xylanohydrolase. A unique xylanase with the requirement for appendant glucuronosyl units. *J. Biol. Chem.* 266, 6539–6543.

Nishitani, K., and Vissenberg, K. (2007). "Roles of the XTH protein family in the expanding cell," in *The Expanding Cell*. Plant Cell Monographs, Vol. 5, eds J.-P. Verbelen and K. Vissenberg (Berlin; Heidelberg; New York: Springer), 89–116.

Okamoto-Nakazato, A., Takahashi, K., Katoh-Semba, R., and Katou, K. (2001). Distribution of yieldin a regulatory protein of the cell wall yield threshold in etiolated cowpea seedlings. *Plant Cell Physiol.* 42, 952–958. doi: 10.1093/pcp/pce121

Passardi, F., Penel, C., and Dunand, C. (2004). Performing the paradoxical: how plant peroxidases modify the cell wall. *Trends Plant Sci.* 9, 534–540. doi: 10.1016/j.tplants.2004.09.002

Passardi, F., Tognolli, M., De Meyer, M., Penel, C., and Dunand, C. (2006). Two cell wall associated peroxidases from Arabidopsis influence root elongation. *Planta* 223, 965–974. doi: 10.1007/s00425-005-0153-4

Peret, B., Li, G., Zhao, J., Band, L. R., Voss, U., Postaire, O., et al. (2012). Auxin regulates aquaporin function to facilitate lateral root emergence. *Nat. Cell Biol.* 14, 991–998. doi: 10.1038/ncb2573

Pérez-Rodrłguez, P., Riaño-Pachón, D. M., Corrêa, L. G., Rensing, S. A., Kersten, B., and Mueller-Roeber, B. (2010). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucl. Acids Res.* 38, D822–D827. doi: 10.1093/nar/gkp805

Perrot-Rechenmann, C. (2010). Cellular responses to auxin: division versus expansion. *Cold Spring Harb. Perspect. Biol.* 2:a001446. doi: 10.1101/cshperspect.a001446

Ray, P. M., Green, P. B., and Cleland, R. (1972). Role of turgor in plant cell growth. *Nature* 239, 163–164. doi: 10.1038/239163a0

Rose, J. K., Braam, J., Fry, S. C., and Nishitani, K. (2002). The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: current perspectives and a new unifying nomenclature. *Plant Cell Physiol.* 43, 1421–1435. doi: 10.1093/pcp/pcf171

Scheller, H. V., and Ulvskov, P. (2010). Hemicelluloses. *Annu. Rev. Plant Biol.* 61, 263–289. doi: 10.1146/annurev-arplant-042809-112315

Schopfer, P. (2006). Biomechanics of plant growth. *Am. J. Bot.* 93, 1415–1425. doi: 10.3732/ajb.93.10.1415

Shin, I., Park, S., and Lee, M-R. (2005). Carbohydrate microarrays: an advanced technology for functional studies of glycans. *Chem. Eur. J.* 11, 2894–2901. doi: 10.1002/chem.200401030

Smyth, G. K. (2005). "Limma: linear models for microarray data" in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber (New York, NY: Springer), 397–420. doi: 10.1007/0-387-29362-0_23

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., et al. (2004). Toward a systems approach to understanding plant cell walls. *Science* 2306, 2206–2211. doi: 10.1126/science.1102765

Sørensen, I., and Willats, W. G. T. (2011). "Screening and characterization of plant cell walls using carbohydrate microarrays," in *Methods in Molecular Biology*, Vol. 715, ed Z. Popper (John Walker) (New York, NY: Humana Press), 115–121. doi: 10.1007/978-1-61779-008-9_8

Talboys, P. J., Zhang, H. M., and Knox, J. P. (2011). ABA signaling modulates the detection of the LM6 arabinan cell wall epitope at the surface of *Arabidopsis thaliana* seedling root apices. *New Phytol.* 190, 618–626. doi: 10.1111/j.1469-8137.2010.03625.x

Ubeda-Thomás, S., Federici, F., Casimiro, I., Beemster, G. T. S., Bhalerao, R., Swarup, R., et al. (2009). Gibberellin signaling in the endodermis controls Arabidopsis root meristem size. *Curr. Biol.* 19, 1194–1199. doi: 10.1016/j.cub.2009.06.023

Van Sandt, V., Suslov, D., Verbelen, J.-P., and Vissenberg, K. (2007). Xyloglucan endotransglucosylase activity loosens a plant cell wall. *Ann. Bot.* 100, 1467–1473. doi: 10.1093/aob/mcm248

Verbelen, J.-P., De Cnodder, T., Le, J., Vissenberg, K., and Baluška, F. (2006). Root apex of *Arabidopsis thaliana* consists of four distinct zones of growth activities: meristematic zone, transition zone, fast elongation zone, and growth terminating zone. *Plant Signal. Behav.* 1, 296–304. doi: 10.4161/psb.1.6.3511

Vissenberg, K., Martinez-Vilchez, I. M., Verbelen, J.-P., Miller, J. G., and Fry, S. C. (2000). *In vivo* colocalisation of xyloglucan endotransglycosylase activity and its donor substrate in the elongation zone of Arabidopsis roots. *Plant Cell* 12, 1229–1237. doi: 10.1105/tpc.12.7.1229

Vissenberg, K., Oyama, M., Osato, Y., Yokoyama, R., Verbelen, J. P., and Nishitani, K. (2005). Differential expression of AtXTH17, AtXTH18, AtXTH19 and AtXTH20 genes in *Arabidopsis* roots. Physiological roles in specification in cell wall construction. *Plant Cell Physiol.* 46, 192–200. doi: 10.1093/pcp/pci013

Willats, W. G., McCartney, L., Steele-King, C. G., Marcus, S. E., Mort, A., Huisman, M., et al. (2008). A xylogalacturonan epitope is specifically associated with plant cell detachment. *Planta* 218, 673–681. doi: 10.1007/s00425-003-1147-8

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

Received: 29 August 2014; accepted: 02 February 2015; published online: 20 February 2015.

*Citation: Wilson MH, Holman TJ, Sørensen I, Cancho-Sanchez E, Wells DM, Swarup R, Knox JP, Willats WGT, Ubeda-Tomás S, Holdsworth M, Bennett MJ, Vissenberg K and Hodgman TC (2015) Multi-omics analysis identifies genes mediating the extension of cell walls in the Arabidopsis thaliana root elongation zone. Front. Cell Dev. Biol. **3**:10. doi: 10.3389/fcell.2015.00010*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 08 October 2014
doi: 10.3389/fcell.2014.00054

Comparative transcriptomics and metabolomics in a rhesus macaque drug administration study

Kevin J. Lee[1], Weiwei Yin[2], Dalia Arafat[1], Yan Tang[1], Karan Uppal[3], ViLinh Tran[3], Monica Cabrera-Mora[4], Stacey Lapp[4], Alberto Moreno[4,5], Esmeralda Meyer[4], Jeremy D. DeBarry[6], Suman Pakala[7], Vishal Nayak[7], Jessica C. Kissinger[6,7], Dean P. Jones[3], Mary Galinski[4,5], Mark P. Styczynski[2] and Greg Gibson[1]*

[1] *Center for Integrative Genomics, School of Biology, Georgia Institute of Technology, Atlanta, GA, USA*
[2] *School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA*
[3] *Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, School of Medicine, Emory University, Atlanta, GA, USA*
[4] *Emory Vaccine Center and Yerkes National Primate Research Center, Emory University, Atlanta, GA, USA*
[5] *Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, GA, USA*
[6] *Center for Topical and Emerging Global Diseases, University of Georgia, Athens, GA, USA*
[7] *Institute of Bioinformatics, University of Georgia, Athens, GA, USA*

Edited by:
Paolo Tieri, Consiglio Nazionale delle Ricerche, Italy

Reviewed by:
Caroline Rae, University of New South Wales, Australia
David A. C. Beck, University of Washington, USA

****Correspondence:***
Greg Gibson, Center for Integrative Genomics, School of Biology, Georgia Institute of Technology, Boggs Building, 770 State Street, Atlanta, GA 30332, USA
e-mail: greg.gibson@biology.gatech.edu

We describe a multi-omic approach to understanding the effects that the anti-malarial drug pyrimethamine has on immune physiology in rhesus macaques (*Macaca mulatta*). Whole blood and bone marrow (BM) RNA-Seq and plasma metabolome profiles (each with over 15,000 features) have been generated for five naïve individuals at up to seven timepoints before, during and after three rounds of drug administration. Linear modeling and Bayesian network analyses are both considered, alongside investigations of the impact of statistical modeling strategies on biological inference. Individual macaques were found to be a major source of variance for both omic data types, and factoring individuals into subsequent modeling increases power to detect temporal effects. A major component of the whole blood transcriptome follows the BM with a time-delay, while other components of variation are unique to each compartment. We demonstrate that pyrimethamine administration does impact both compartments throughout the experiment, but very limited perturbation of transcript or metabolite abundance was observed following each round of drug exposure. New insights into the mode of action of the drug are presented in the context of pyrimethamine's predicted effect on suppression of cell division and metabolism in the immune system.

Keywords: pyrimethamine, bone marrow, peripheral blood, axes of variation, bayesian network inference, principal component analysis (PCA)

INTRODUCTION

The Malaria Host-Pathogen Interaction Center (MaHPIC) has initiated a systems biology program to understand the course of events and mechanistic processes that occur in the biology of infected non-human primates (NHPs) and *Plasmodium* parasites over the course of malaria episodes. The long-term goal is to advance the development of interventions for this major global parasitic disease (WHO World Malaria Report, 2013). This research program investigates how NHP-infective species of *Plasmodium* that model human malaria caused by *P. falciparum* and *P. vivax* elicit various host responses, develop immunity, adopt immune-avoidance strategies, and cope with anti-malarial drugs (Galinski et al., 2013; Wright and Rayner, 2014). We are integrating diverse data types, including transcriptomics, metabolomics, lipidomics, proteomics, and innate and adaptive immune profiles and performing cross-species comparisons with multiple different host-parasite infection model combinations. There are many gaps in knowledge relating to *Plasmodium* infections including the immune response, the mechanisms of malaria pathogenesis and multiorgan dysfunction, the adverse impact on the bone marrow (BM) progenitors and the dynamics of co-infections (Hafalla et al., 2011; Schwenk and Richie, 2011; Frevert and Nacer, 2013; Stanisic et al., 2013). The NHP models being studied enable more rigorous experimentation and in-depth analyses than are possible from direct investigations in humans (Deye et al., 2012; Tachibana et al., 2012; Moreno et al., 2013) and they are well-suited for systems biology approaches.

In this study, we establish logistics and procedures that lay the foundation for studies of rhesus macaques (*Macaca mulatta*) inoculated with infectious *Plasmodium* parasites, following which intermittent antimalarial drug intervention may be required. The data presented here serve as pilot data, with inoculations consisting of Anopheline mosquito salivary gland preparations lacking sporozoites, and the analyses begin to show how multiple diverse datasets can be integrated. We present multi-omic data analyses using top-down approaches to the integration of RNA-Seq derived transcriptome data from the BM and peripheral blood

(PB), as well as plasma metabolite data, and complete blood cell count (CBC) parameters. These data types were obtained during a 100-day period, at specific timepoints before and after pyrimethamine administration to five rhesus macaques.

By top-down integration, we mean statistical and machine-learning strategies that are naïve to the known biochemical annotation of the transcript and metabolite features (Bang et al., 2008; Giuliani et al., 2014). Our approach is to use principal components analysis (PCA) to describe the major sources of variance (among individual animals or temporal) in each data type, and then to seek correlations between the major components across data types (Boedigheimer et al., 2008). We perform standard differential gene expression analysis, also asking how the statistical modeling strategy and data reduction influence identification of drug-responsive genes, and employ gene set enrichment analysis to identify pathways of interest. In an attempt to overcome the limitations of orthogonal PCA, particularly in the context of a relatively small experiment, we ask whether biologically derived axes of variation that are known to consistently capture PB variation in humans, are conserved in macaques and covary with drug treatment. A bottom-up strategy, starting with known cellular and biochemical pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG: Kanehisa and Goto, 2000; Kanehisa et al., 2014), is contrasted and used to help draw more inferences about the physiological impact of pyrimethamine particularly on the BM. Finally, Bayesian network analysis (Bumgarner and Yeung, 2009; Pei and Shin, 2012) is also applied as an orthogonal approach with promise for overcoming the conditional dependence of the transcriptome and metabolome in a dataset with high dimensionality and a small number of samples. The work flow is shown in **Figure 1** including the questions posed by each mode of analysis and the major conclusions.

The null hypotheses are first that neither host nor drug administration impact gene expression, and second that the major variance components of the BM and PB transcriptomes and plasma metabolome are uncorrelated. Transcriptome data was collected by RNA-Seq (Wilhelm and Landry, 2009) with a mean of 40 million paired-end reads for each of 35 BM and 35 PB samples studied (five macaques each with seven collection timepoints), focusing on transcript abundance for ∼15,000 genes. Metabolome data was collected by Orbitrap mass spectrometry following liquid chromatography (Jones et al., 2012; Soltow et al., 2013) on two different columns (AE and C18) with ∼6000 and ∼14,500 m/z features, respectively. Our expectation was that drug administration would have global effects on each of the four omic measures (BM and PB transcriptomes, and AE and C18 generated metabolomes), and that among-individual differences would be relatively minor. However, we had no pre-conception of the fraction of genes that would be differentially expressed, or of the degree of correspondence we would find between the transcriptomes and metabolome. Since the PB consists of cells generated in the BM, we expected a temporal delay between these two compartments with considerable overlap in variance components, which would also reflect differences in the counts of major blood cell types obtained by standard CBC analysis. Herein we quantify departures from each of these expectations as well as a general failure to reject the null hypothesis that the blood transcriptomes and metabolomes are uncorrelated, and discuss the implications for the mode of action of pyrimethamine.

MATERIALS AND METHODS

EXPERIMENTAL DESIGN

The experimental design of this experiment involving rhesus macaques (*Macaca mulatta*) was approved by the Emory University Institutional Animal Care and Use Committee (IACUC) and is as follows. Five males (*RCs13, RWr13, RUn13, RZe13,* and *RTi13*) approximately 2 years of age were injected intravenously with a preparation of *Anopheles dirus* salivary gland material (prepared similarly to how infectious *Plasmodium* sporozoites would be purified; Kennedy et al., 2012) and then profiled for clinical and omic measurements over the course of a 100-day experiment. The animals were moved into experimental pair housing (RCs13/RWr13 and RUn13/RZe13) 10 days prior to the baseline sampling point at Day 0, namely timepoint 1 (TP1). The fifth macaque (RTi13) was housed alone. Capillary blood samples collected daily from ear pricks into EDTA-tubes were used to obtain complete blood cell counts (CBCs), with the exception of days 51 to 53 when an equipment failure occurred. On days 21, 27, 52, 59, 90, and 98, PB and BM samples were collected comprising TPs 2-7. These collections, and that of TP1, were taken under chemical restraint with ketamine delivered intramuscularly at 10 mg/kg. This dissociative anesthetic has a short elimination half-life (20–40 min) and, to our knowledge, has no known drug interaction with pyrimethamine. The experimental design does not, however, allow for distinguishing the effects of the drug or anesthetic. BM aspirates were obtained from the right or left iliac crests in an alternating manner for consecutive timepoints and performed using 18G needles. Immediately after collection BM samples were transferred into Vacutainer EDTA tubes. PB samples were collected from the femoral artery into Vacutainer EDTA tubes. The transcriptomes and metabolomes were interrogated at seven (TP1-7) and five (TP3-7) timepoints, respectively, as shown in **Figure 1**. Pyrimethamine (Sigma-P7771) was delivered (1 mg/kg) intramuscularly once on day 20, and for 3 successive days starting at days 52 and 90 (TP2, 4, and 6), corresponding to predicted periods for sub-curative and curative experimental treatment regimens for malaria infection of macaques.

LIBRARY PREPARATION FOR RNA-SEQ

BM (1 ml) was collected into 1.5 ml tubes with EDTA, and the mononuclear cells were purified by density gradient centrifugation on Lymphoprep (Stem Cell Technologies) solution and preserved in RLT buffer (Qiagen) to stabilize mRNA. Whole blood (3 ml) was collected in Tempus tubes (Applied Biosystems) which preserve mRNA; these samples include erythrocytes, platelets and granulocytes, and mononuclear lymphocytes. RNA was extracted from the BM samples using Qiagen RNEasy Mini-Plus kits following the manufacturer-recommended procedures, and from PB samples using Tempus-Spin RNA isolation kits (Life Technologies). The quality of all RNA samples was confirmed using a Bioanalyzer, with an RNA Integrity Number (RIN; Schroeder et al., 2006) greater than 8 recorded for all samples.

Approximately 1 μg of total RNA per sample was converted to double-stranded cDNA using poly-A beads to enrich

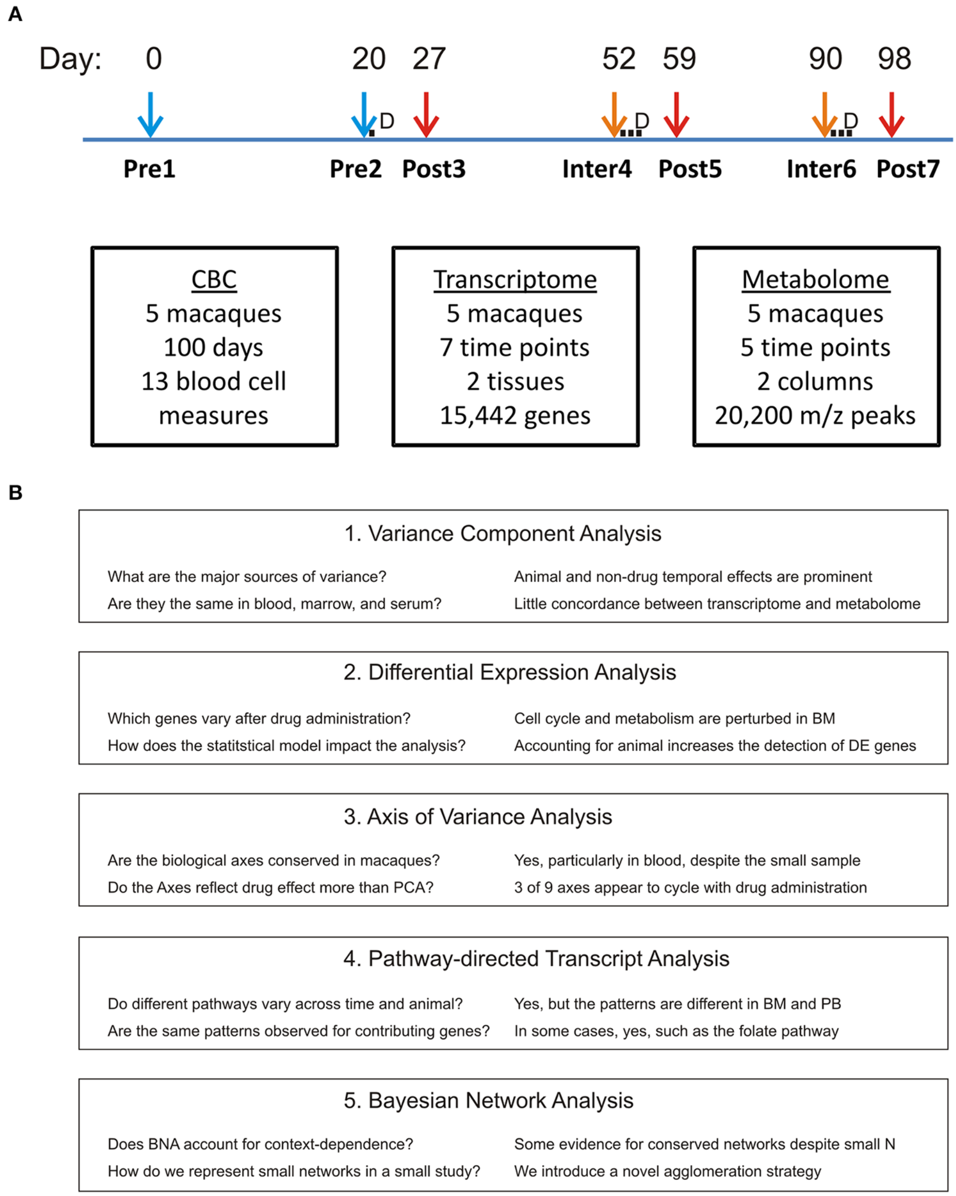

A
Day: 0
20
27
52
59
90
98
D
D
D
Pre1
Pre2
Post3
Inter4
Post5
Inter6
Post7
CBC
5 macaques
100 days
13 blood cell measures
Transcriptome
5 macaques
7 time points
2 tissues
15,442 genes
Metabolome
5 macaques
5 time points
2 columns
20,200 m/z peaks
B
1. Variance Component Analysis
What are the major sources of variance?
Are they the same in blood, marrow, and serum?
Animal and non-drug temporal effects are prominent
Little concordance between transcriptome and metabolome
2. Differential Expression Analysis
Which genes vary after drug administration?
How does the statitstical model impact the analysis?
Cell cycle and metabolism are perturbed in BM
Accounting for animal increases the detection of DE genes
3. Axis of Variance Analysis
Are the biological axes conserved in macaques?
Do the Axes reflect drug effect more than PCA?
Yes, particularly in blood, despite the small sample
3 of 9 axes appear to cycle with drug administration
4. Pathway-directed Transcript Analysis
Do different pathways vary across time and animal?
Are the same patterns observed for contributing genes?
Yes, but the patterns are different in BM and PB
In some cases, yes, such as the folate pathway
5. Bayesian Network Analysis
Does BNA account for context-dependence?
How do we represent small networks in a small study?
Some evidence for conserved networks despite small N
We introduce a novel agglomeration strategy

FIGURE 1 | Experimental Design. (A) Five macaques were each delivered a sub-curative dose of pyrimethamine at Day 21, and 3-day curative doses commencing at Days 52 and 90, in each case immediately following peripheral blood sampling. This results in two pre-drug, three post-drug, and two inter-drug treatments as indicated. Metabolome data was not generated for the first two timepoints. **(B)** Flow of analytical approaches including major questions asked and inferences drawn.

for mRNA, and Illumina TruSeq Stranded mRNA Sample Prep kits to generate strand-specific libraries. As a quality control, 96 spike-in RNAs of known concentration and GC proportions (ERCC Spike-In Control, Life Technologies; Devonshire et al., 2010) were added to constitute approximately 1% of the total RNA for each library. Adapters were ligated to facilitate 3-plex sequencing on an Illumina HiSeq2000 at the Yerkes National Primate Center Genomics Core, aiming for 80 million paired-end 100 base pair (bp) reads per library. Average insert sizes were in the range of 300–400 bp.

SHORT READ MAPPING AND GENE EXPRESSION QUANTIFICATION

To quantify gene expression, the RNA-Seq reads were mapped to an early version of a new assembly of the rhesus macaque (MacaM assembly, Version 4.0, GenBank accession number PRJNA214746 ID: 214746, created by Aleksey Zimin at the University of Maryland, Rob Norgren at the University of Nebraska Medical Center, and their colleagues) using Tophat2 (Trapnell et al., 2012; Kim et al., 2013). Default options were used with the exception that the command—library-type fr-secondstrand was invoked since the reads were generated using a stranded library preparation method. This allowed us to differentiate between sense and antisense transcripts. Rob Norgren and his colleagues also provided a GTF file (version 4.12) of the annotated MacaM assembly indicating the exon boundaries of rhesus genes that was used in our transcriptome analyses to improve the mapping accuracy across splice junctions. Only reads that map to a single location in the genome were included, to ensure high-confidence mapping. All downstream analyses were performed at the level of annotated gene: this study does not consider exon-specific or transcript isoform relative abundance. Transcription was detected for 15,442 genes. The dataset has been deposited to the Gene Expression Omnibus archive (GEO) under accession number GSE58340.

Several quality control steps were used to verify the reliability of the data: linear correlation of estimated abundance of ERCC spike-in controls with known concentration; confirmation of 99.9% strand-specificity of the controls; less than 0.1% control fusion transcripts; and absence of 3′ bias in the controls was confirmed with RSeqC v2.3.8 software (http://rseqc.sourceforge.net; Wang et al., 2012). Transcript abundance levels were inferred using HTSeq v0.5.4p5 (http://www-huber.embl.de/users/anders/HTSeq/doc; Anders et al., 2014). HTSeq takes the short-read mapping.bam file from tophat2 and the gene annotation file which contains the locations of all annotated genes. Since some libraries were sequenced more deeply than others, the libraries were normalized before determining differential gene expression using the gene level expression files with the default parameters of DESeq version 1.10.1 (http://www.bioconductor.org/packages/release/bioc/html/DESeq.html; Anders and Huber, 2010).

METABOLOMIC FEATURE QUANTIFICATION

High resolution metabolomics (m/z range 85–2000) was performed using a liquid chromatography/mass spectrometry (LC/MS) approach on a Thermo Orbitrap-Velos Mass Spectrometer (Thermo Fisher, San Diego, CA) via positive-ion electrospray ionization (ESI). Two different columns were used for the LC separation stage: C18 and anion exchange (AE). Each distinct biological sample was run in triplicate in order to ensure high reliability of the data, with randomization within batches (Soltow et al., 2013). MS peaks were called using xMSanalyzer v1.3.2 (Uppal et al., 2013) with apLCMS v5.9.4 (Yu et al., 2009). Standard quality control measures were performed, such that features with greater than 30% missing values were removed from the analysis. Since the frequency distributions of all samples were comparable, no additional normalization was performed, but an abundance cutoff of 256 peak area units was adopted and all features below this were excluded. All downstream analyses utilized the median values of three technical replicate samples, namely a single measure per biological sample. The AE and C18 columns generated 5861 and 14,339 m/z and retention time features respectively, the majority of which are either not yet annotated or have ambiguous annotation to multiple possible organic compounds. The m/z features are thought to include the majority of known components of central metabolism, as well as xenobiotics.

STATISTICAL ANALYSIS

After data normalization, the transcriptome and metabolome levels were log-2-transformed and imported into JMP Genomics (version 6.0, SAS Institute, Cary, NC). The log-2 transformation was performed both to ensure that the data is more normally distributed and to facilitate simple comparison of the magnitude of differential expression in a symmetrical manner with respect to up- and down-regulation, as is standard in microarray analysis: plus or minus 1 unit corresponds to a 2-fold change for each of the datasets. To determine how much of the variance in each of our datasets is explained by our two measured factors (animal and timepoint), we performed a principal components (PC)-variance component analysis using JMP v6.0 (SAS) for the transcriptomes, metabolomes, and the CBC data (Boedigheimer et al., 2008). This consists of generation of all PC explaining up to 90% of the total variance (12–15 for the transcriptomes and ∼30 for the metabolomes), regressing each PC on "animal" or "timepoint," and generating a weighted average of the squared correlation coefficient (percent variance explained) across all of the PC scores. Since the low abundance features for metabolomics and transcriptomics both have high coefficients of variation, we set thresholds of 5 log2 units for transcripts and 17 log2 units for metabolites based simply on visual inspection of plots of the coefficient of variance against average abundance. After estimating the effect of lower-abundance features on the variance components, they were removed for all downstream analyses. No attempt was made to optimize the threshold or systematically evaluate its impact, but the major conclusions of the study are unlikely to be affected.

To assess whether the major PC capture similar aspects of the data, the first 10 PC were calculated for the four omics datasets using JMP. All 780 pairwise correlations of these PC values were determined, and a Bonferroni multiple comparison adjustment was used to assess the significance of each pair of PC. Exploratory partial least square regression analyses were also performed with MixOmics (González et al., 2012; http://cran.r-project.org/web/packages/mixOmics/index.html) in an attempt to select variables that co-vary, but did not reveal significant associations.

BLOOD INFORMATIVE TRANSCRIPT (BIT) AXES

In addition to PCA, we employed a second method, blood informative transcript (BIT) axes analysis (Preininger et al., 2013). Briefly, 10 highly co-regulated transcripts in blood (the BIT) capture each of 9 common axes of variation that are observed in all human PB gene expression datasets. PC1 for each of these 9 sets of 10 transcripts provide Axis scores for each individual sample, and were generated independently for both the PB and BM samples, using the normalized expression data. We then examined the dynamics of the axes scores (or their residuals after fitting "Animal") over time and used ANOVA to evaluate differences among the timepoints or animals.

DIFFERENTIAL GENE EXPRESSION

The next step in our analysis was the identification of genes that are differentially expressed across the experimental conditions (Soneson and Delorenzi, 2013). For between-TP differences, an ANOVA was performed on each transcript separately using "animal" as a random effect with five levels and "timepoint" with seven levels, or "drug" with three levels as the fixed effect. For the drug exposure factor, we define our three experimental conditions as before drug exposure (pre-drug; TP1 and TP2), 7 days after the most recent dose (post-drug; TP3, TP5, and TP7), and 30 days after most recent dose and immediately before the next dose (inter-drug; TP4 and TP6), as shown in **Figure 1**. A Benjamini-Hochberg false discovery rate cutoff of 5% was used to define differentially expressed genes. These were examined using hierarchical clustering of the standardized least squares means, and volcano plots of significance against fold difference between specific conditions (Wolfinger et al., 2001). The significantly differentially expressed genes are reported as a Supplementary flat file that consists of a list of gene names with their corresponding F-statistics at http://www.cig.gatech.edu/supplementary-data.

Gene set enrichment analyses were performed using pre-existing human gene set annotations from the Broad Institute (Subramanian et al., 2005), considering that the majority of known genes in the macaque genome have very closely related syntenic human orthologs (Zhang et al., 2014). We used the ranked gene list method of GSEA v2.0.14 (http://www.broadinstitute.org/gsea/index.jsp) to perform the contrast of interest (pre-vs.-[post plus inter] drug treatment), testing for enrichment of t-statistics in KEGG pathways and/or GO terms. Gene sets with a nominal $p < 0.001$ and an FDR $q < 5\%$ were considered as significant per the recommendations of the GSEA software manual. Default parameters were used, excluding gene sets with more than 500 or fewer than 20 genes.

BAYESIAN NETWORK ANALYSIS

For the Bayesian network analyses, only the 1000 most differentially expressed genes (largest F-ratios for the Drug effect) or 500 metabolites were used, so as to ensure computational tractability of the clustering software while incorporating biologically relevant genes. The transcript abundance measures were the residuals after fitting "animal" in the ANOVA to remove this large overall source of variance, while raw median metabolite abundance measures were used based on the relatively small contribution of animal and timepoint to the variance. Custom scripts were written in MATLAB and R to perform quality threshold clustering (Heyer et al., 1999; De Smet et al., 2002) on the mean-centered expression values, namely the residuals after fitting "animal" to each gene. A $d = 0.3$ cluster similarity threshold was employed as suggested by Heyer et al. (1999). Since the data are normalized and since that similarity threshold is based on a specific type of correlation metric, it is reasonable to expect that such a value may be an excellent starting point across transcriptomic studies. We also performed small perturbations of the cluster similarity threshold and found that the main differences were in the merging portions of some smaller clusters into some bigger clusters. The number and membership of the larger clusters (used for analysis here) remained similar (data not shown). Only clusters with at least 10 transcripts (or metabolites) were retained for further analysis. Significance and robustness of these clusters were assessed via permutation tests.

Since robust and accurate Bayesian network inference is typically very difficult with only 35 observations (five animals and seven timepoints), we treated genes within clusters as separate observations of those clusters. We ranked each gene relative to its correlation with the centroid of the cluster across all 35 samples (25 for metabolites since TP1 and TP2 were missing) and then concatenated the top 10 genes into a list of 350 (250) observations for each of 26 transcript and four metabolite clusters that satisfied the clustering criteria. We chose to use just the top 10 genes so as to ensure that each node in the network has the same number of observations (no missing data) and because Bayesian network inference benchmarking literature has shown that having a few hundred observations can provide reasonably robust inference. The selection of 10 genes thus balances robustness with the number of different clusters that can be analyzed, as increasing that threshold necessarily eliminates more clusters. Next, the data was discretized using a mutual information content-preserving algorithm (Hartemink, 2001), as the Bayesian network analysis is expected to be more robust for discrete data. Briefly, for each variable, observations are stepwise coalesced into discretized bins such that the loss in mutual information content between that variable and all other variables is minimized. The estimated elbow-point in the remaining mutual information as a function of the number of discretization levels was then selected as the desired number of levels (seven for transcriptional and five for metabolite data). Networks were subsequently generated using the Sparse Candidate Algorithm (Friedman et al., 1999) of Causal Explorer in MATLAB (http://www.dsl-lab.org/causal_explorer; Aliferis et al., 2003). The most robust connections between clusters were identified using subsampling and permutation tests. We used three shuffled datasets, within which the order of the 10 genes in each cluster was permuted independently, to minimize the possibility of over-fitting the available data: if the 10 genes are good representatives of the cluster, then there should not be much mutual information based solely on a given gene in one cluster being compared to a specific gene in another cluster, and so the most robust edges (and least likely to be due to over-fitting) are the ones that occur in multiple shuffled datasets. For each of these shuffled datasets we assessed the sensitivity of the inference method to perturbations in the amount of available data by performing network inference with 90% subsampling of 1000 replicates of the dataset.

To assess whether the BM clusters are valid in the PB, the PB data for each of the clusters was used to determine their centroids. The distribution of Pearson correlation coefficients for each member of the cluster to the centroid was calculated. These distributions were compared to analogous calculations for random samples of the same number of genes for each cluster, using a one-tailed Kolmogorov–Smirnov test.

To assess whether the BM clusters form a network in the PB, we used the same 26 clusters from the BM data, identified the 10 genes closest to the centroid using the PB data for each cluster, and used concatenated gene data for each cluster to generate new Bayesian networks. To evaluate whether there are interactions between transcriptomic and metabolic networks, we repeated the Bayesian network inference using the 26 BM transcript clusters and four plasma metabolite clusters from the C18 column MS data analyzed with essentially the same pipeline. The same methods as described above were used to form this integrated network, except that for the transcriptional data only the five timepoints corresponding to the metabolomics timepoints were kept, the data discretization was performed jointly on the combined datasets, and metabolomics data was unit normalized.

RESULTS

VARIANCE COMPONENTS OF OMIC DATA

The experimental design consisted of a 100 day mock-infection cycle of *M. mulatta* that follows a similar time course as will be used for a series of *Plasmodium* infections in later MaHPIC studies (**Figure 1**). Each of five monkeys was transferred, in two pairs and a single, to indoor cages at the Yerkes National Primate Research Center in Atlanta, Georgia, 10 days prior to the baseline (TP1) PB and BM draws. The second timepoint (TP2) samples were taken 20 days later, immediately prior to administration of the anti-malarial drug pyrimethamine for the first time and 7 days before sampling of the first post-Drug timepoint (TP3). After a further 3 weeks, a second round of drug treatment immediately followed the TP4 sampling. Consequently, TP1 and TP2 represent pre-drug samples that nevertheless differ in their gene expression and metabolic profiles, we suspect due to an acclimation period of each animal in the new experimental environment. TP3, TP5, and TP7 represent post-drug samples, and TP4 and TP6 represent inter-drug samples. Plasma metabolites were only profiled following the drug administration period (TP3-TP7), but CBCs were generated for all timepoints except TP4.

Our first objective was to define the variance components of gene expression and metabolite abundance, namely the contributions of among animal and among timepoint differences to the overall variance. This was accomplished by generating the PC that collectively account for 90% of the variance of each omic data type and computing the weighted average of the variance of each PC explained by animal or timepoint. The variance explained by each PC is shown in **Table 1** and typically ranges from 25% for PC1 to less than 3% for PC5 (subsequent PC contribute too little to the overall variance to significantly impact the evaluation of contributions of animal and time). **Figure 2A** shows that for the BM and PB transcript data, as well as the CBC data, approximately 30% of the variance is among animal and 10% among timepoints. For the metabolomes by contrast, only 15% of the variance is among animals with a slightly larger proportion due to timepoint.

The unexplained residual variance could be due to undefined biological sources, animal-by-timepoint interactions, random sampling variance, or technical error. To control for contributions of the latter, we reduced the datasets by removing the low-abundance features with the greatest coefficients of variation. Consistent with published findings (Rapaport et al., 2013), both RNA-Seq and MS have a strong relationship between abundance and variability, and based on the plots we adopted heuristic cutoffs of 5 log2 units for the transcripts and 17 log2 units for the metabolites. **Figure 2B** shows the variance components analysis based on the remaining features. In the PB, almost 70% of the variance is among animals, and in the BM approximately 50%. The temporal contribution drops to less than 5% for the PB, but increases to 20% for the BM. These results confirm that measurement error is a major contributor to estimation for low abundance transcripts with RNA-Seq. By contrast, the variance components for both metabolite columns is relatively unaffected by the data reduction, with both animal and time each continuing to explain approximately 15% of the overall variance. This may reflect filtration of the metabolites with the highest technical variance during peak calling.

Table 1 | Principle components of variation.

PC	PVE[a]	Animal[b]	Timepoint[b]	Sig. drug[c](effect)
BM1	14.5%	0.67	0.18	3×10^{-5}(Pre high)
BM2	10.2%	0.92	0.02	0.09^{ns}
BM3	7.1%	0.15	0.74	2×10^{-4}(inter low)
BM4	6.8%	0.91	0.05	0.0052 (pre high)
BM5	5.8%	0.80	0.13	0.011 (post low)
PB1	12.0%	0.94	0.02	0.59^{ns}
PB2	8.2%	0.97	0.01	0.48^{ns}
PB3	7.4%	0.87	0.01	0.75^{ns}
PB4	7.0%	0.97	0.01	0.98^{ns}
PB5	5.5%	0.12	0.14	0.36^{ns}
AE1	17.3%	0.23	0.41	0.35^{ns}
AE2	8.9%	0.13	0.47	0.14^{ns}
AE3	7.3%	0.23	0.47	0.58^{ns}
AE4	6.2%	0.41	0.34	0.13^{ns}
AE5	4.9%	0.22	0.37	0.07^{ns}
C18_1	18.5%	0.28	0.23	0.17^{ns}
C18_2	9.3%	0.20	0.59	0.11^{ns}
C18_3	6.6%	0.40	0.17	0.15^{ns}
C18_4	6.3%	0.28	0.31	0.49^{ns}
C18_5	4.8%	0.22	0.38	0.38^{ns}

[a] *Amount of total variance explained by the PC.*

[b] *Amount of variance explained by Animal or Timepoint.*

[c] *Significance of Drug effect (pre vs. post vs. inter for transcriptome; post vs. inter for metabolome), also showing which effect was differentiated.*

HIERARCHICAL CLUSTERING OF THE OMIC DATA

The preceding analysis tells us that both animal and time influence gene expression, but not which animals or timepoints are more similar. A quick means of visualizing these relationships is by two-way hierarchical clustering (**Figure 3**; Eisen et al., 1998). Applied to the raw data, in a joint analysis of the BM and PB, the gene expression of the two tissue types is clearly distinct, and the greater contribution of animal to the PB than the BM is seen by the perfect clustering of each of the seven timepoints within each animal grouping (**Figure 3A**). In the BM, there is some mixing of samples across animals, but it is also striking that TP4 is somewhat distinct since the data from four of the five animals cluster together. After standardization to *z*-scores for each gene, which removes the effect of overall abundance level for each transcript on the hierarchical clustering, these relationships are largely maintained (**Figure 3B**). The separation of TP4 in the BM is enhanced, although the clustering within animals is disrupted for a few PB samples. The situation for the metabolome is very different (**Figure 3C**) as neither the animals nor timepoints form discrete clusters. Both LC columns have almost identical topologies (data not shown), and technical variability does not explain these results since almost without exception all three technical replicates of each metabolite sample cluster adjacent to one another.

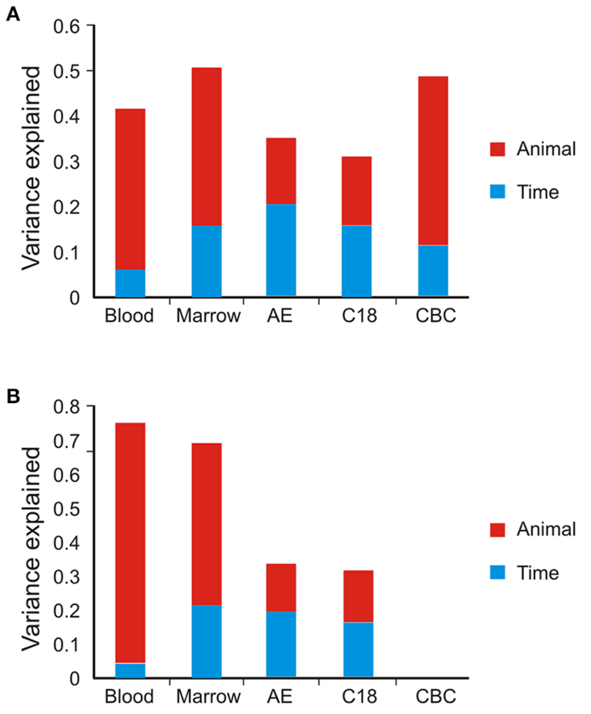


FIGURE 2 | Principal component variance component analyses. Bar graphs show the weighted average contribution of among animal (red) and among timepoint (blue) variance to gene expression, metabolite, and complete blood counts. **(A)** Full data set. **(B)** Reduced dataset after removal of low abundance transcripts or metabolites.

With respect to blood cell counts, the hierarchical clustering topology of monocytes, lymphocytes, granulocytes, RBC, and platelets did not correspond to either the transcriptome or metabolome topologies. The final two timepoints (TP6 and TP7) cluster to the exclusion of the earlier timepoints with a couple of exceptions, and within the two large clusters the individual animals are adjacent. However, there is no strong relationship between blood cell counts and overall gene expression (**Figure 3D**). Since each blood cell type has a characteristic expression profile which allows each cell to perform its specified role(s), including the limited mRNA complement in a nuclear RBC, we hypothesized that the macaques that clustered together in the expression profile would also have similar levels of the major cell types. However, we do not observe such a trend: macaques *RCs13* and *RWr13* (relative to *RTi13*, *RUn13*, and *RZe13*) form two sets of transcript profiles, whereas *RTi13* and *RWr13* are most similar for CBC with *RZe13* the most variable. Therefore, we conclude that the CBC is capturing information about the system that is non-redundant with the transcriptome. This result is particularly striking when considering that both the transcriptome datasets as well as the CBC datasets have the variance component of animal explaining more than 30% of the variance.

INTEGRATION OF THE TRANSCRIPTOME AND METABOLOME PROFILES

These analyses suggest that the transcriptomes and metabolome are poorly correlated overall across all timepoints and animals, but do not exclude the possibility that subsets of features in the BM, the PB, or the plasma may be co-regulated. To test this using our top-down strategy, we evaluated the covariance between the major PC of each of the four omic datasets. **Figure 4** shows the pairwise regression coefficients for each of the first 10 PC, allowing for the possibility that minor PC involving strong covariance of a small number of transcripts or metabolites may contribute.

The pattern that emerges is informative in many ways. Firstly, it shows that the two metabolomic datasets are highly correlated. This is to be expected since metabolites are being measured from the same plasma sample; the difference between the two datasets is the use of different liquid chromatography columns to optimize peak resolution across different classes of metabolites (broadly, sugars and amino acids on the AE anion exchange column, and lipids on the C18 column). PC1 and PC2 scores for the two columns are highly significantly correlated; many lower PC scores are also correlated. Unlike the metabolomic datasets, the two transcriptomic datasets, representing the BM and PB, do not show as much correlation (**Figure 4**, top left quadrant). Statistical analysis however does indicate that PC1, PC2, PC3, and PC4 in the PB are significantly correlated with PC2, PC4, PC1, and PC5 from the BM, respectively (Bonferroni corrected $p < 0.05$). Such a result is not unexpected considering that the two compartments have different functions yet one (blood) is composed of cell populations derived from the other (marrow). In some cases the sign of the regression is negative, but this is simply a function of PCA which commonly reverses signs and order of PC due to sampling variance. One difference between the compartments is that the marrow contains many cell types that are rapidly dividing whereas most of the cells in the blood are likely to be post-mitotic and terminally differentiated.

Strikingly, there is no significant correlation between the transcriptome PC scores and the metabolome PC scores across animals and timepoints. **Figure 4** (top right box) seems to show some relationships, but none of these are significant after multiple testing correction. We also explored 2-block partial least square analysis (González et al., 2012) to identify minor variance components that may correlate in a joint analysis, but did not observe any significant enrichment between the two data types. This could be explained by the fact that the transcriptome of these two immune compartments is contained within the cell whereas the metabolome that we are interrogating is in the plasma. Furthermore, the plasma is not only influenced by metabolites from blood cells, but by metabolites secreted from all tissues in the body and taken up from the environment.

The correlation of the major PC between the PB and the BM datasets is dominated by among animal differences from the variance component analysis, but also includes a temporal component. The pre-drug samples in the BM are distinct from the post- and inter-drug samples, but TP4 is the most differentiated. In the PB, the baseline sample (TP1) is most differentiated, but TP4, and, even more strongly, TP5 are also somewhat divergent from the remaining samples. To assess the significance of the overlap, we extracted the genes that were up-regulated in the TP4 samples from the BM and performed a binomial sign-test of whether the same genes were up-regulated at TP5 in the PB. The result was highly significant ($p < 3 \times 10^{-16}$). A similar

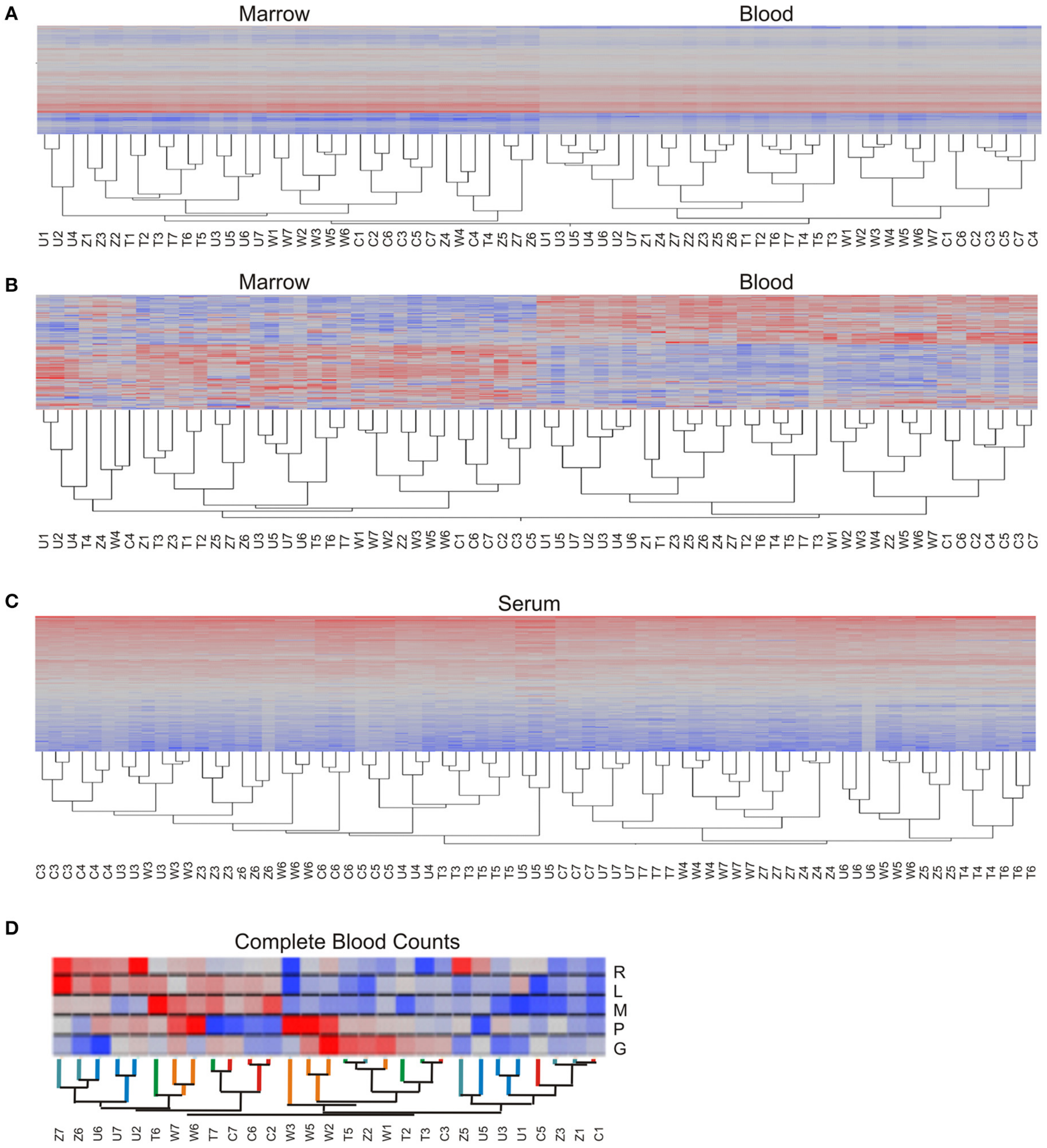


FIGURE 3 | Two-way hierarchical clustering. Each heat map shows the abundance of each transcript or metabolite (rows) in each sample (column) with red indicating high expression, blue low, gray intermediate. **(A)** Transcriptome, based on absolute log-2 intensity estimates, and **(B)** based on standardized log-2 intensities, in both cases combining both BM and PB in the same clustering. **(C)** Plasma metabolome, where each column is a technical replicate, showing almost perfect alignment of each of the three replicates of each sample. **(D)** Complete blood counts are clustered with the branches of the dendrogram colored according to the identity of the animal, and cell types ordered as Red Blood Cells, Lymphocytes, Monocytes, Platelets, and Granulocytes (R, L, M, P, G respectively). Each macaque is abbreviated as C, T, U, W, or Z for RCs13, RTi13, RUn113, RWr13, or RZE13 respectively, and numbers refer to timepoints.

result was obtained for the down-regulated genes, but the control comparison of TP6 and TP7 in the PB for the same up- and down-regulated genes and did not result in any enrichment ($p = 0.74$). These data show that differential gene expression in the BM is reflected in the PB with a time lag (though contamination of BM with PB cannot be excluded). Note as well that in both tissues the TP4/5 differentially expressed genes are similar to the TP1 genes, but with opposite sign of effect, implying that genes up-regulated at baseline are down-regulated at TP4/5, and vice versa.

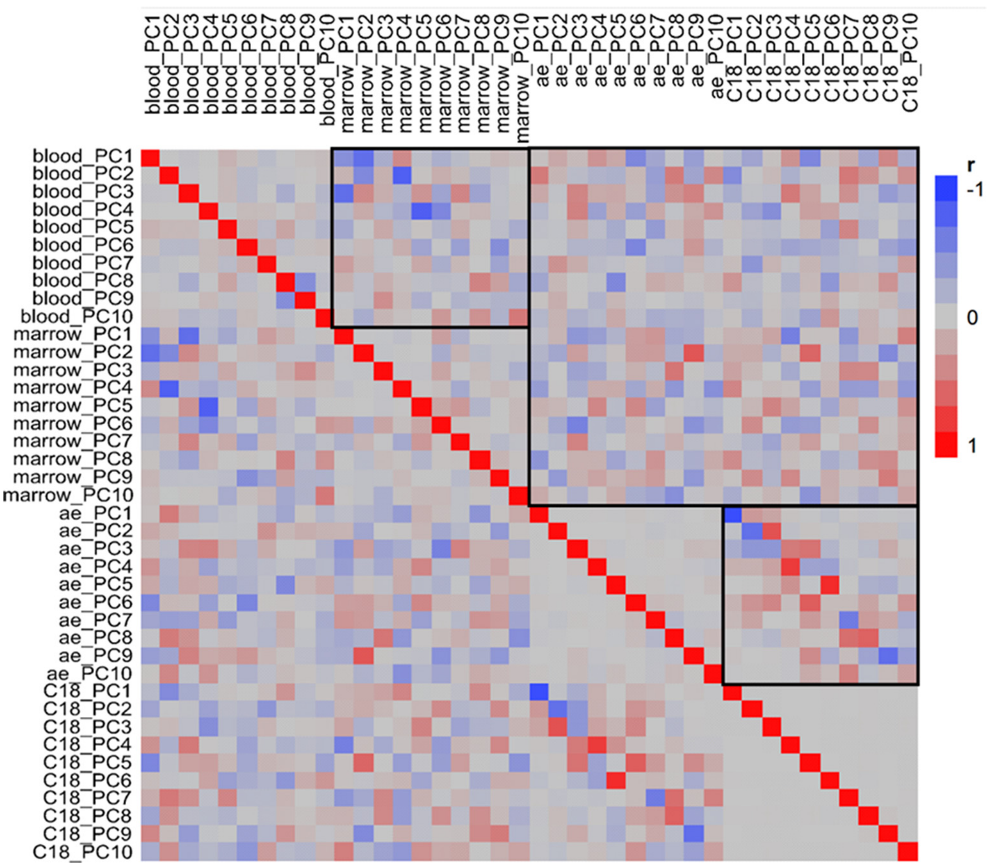


FIGURE 4 | Pairwise correlation of principal components. Heatmap shows correlations of the first 10 principal components for each of the four omic datasets ordered by peripheral blood, bone marrow, AE and C18 (blue negative, red positive, stronger hues indicate stronger correlation). The highlighted boxes in the top left and bottom right quadrants show the correspondence between PC scores for the two transcriptomes and two metabolome columns, respectively; stronger colors along the diagonal indicate that those PC are capturing similar signals. Although there are scattered stronger colors in the top right quadrant comparing transcriptome and metabolome, the actual correlations are not significant.

IDENTIFICATION OF DRUG-RESPONSIVE GENES AND METABOLITES

The major temporal component of variation is not a cyclical response to drug administration, which would have produced a pattern where TP3, TP5, and TP7 were distinct from TP4 and TP6 and again from TP1 and TP2. None of the first 10 PC in PB showed a significant effect of drug administration on gene expression by analysis of variance with pre, inter and post levels of drug (**Table 1**). We nevertheless employed three strategies to identify potential drug-responsive genes: axis of variance analysis, pathway-oriented analysis, and gene set enrichment analysis.

The requirement that PCs are orthogonal to one another introduces a statistical bias that is well-known to obscure underlying biology (Biswas et al., 2008). Consequently, we employed an alternate partitioning of the transcriptional variance based on conserved patterns of covariance of axes of gene expression that are observed in all large human PB transcriptome datasets (Preininger et al., 2013). Each of 9 axes is defined by 10 BITs that are highly co-regulated, and the first PC of these BITs is used as a measure of activity of genes in the axis. Each axis is thought to represent an aspect of immune function, such as T- or B-cell signaling (Axes 1 and 3), innate immune activation (Axis 5) or an interferon-related axis (Axis 7). We confirmed that the BIT are co-regulated in macaques, in both the PB and the BM, and asked whether they vary by animal or timepoint. **Figure 5** shows highly significant among animal effects in the BM for Axes 3, 5, and 7. Importantly these data also capture a drug response effect that is not evident from the standard principal components, as Axes 7 and 9 are clearly differentially expressed at TP3, TP5, and TP7 in the BM, representing the post-drug samples. Axis 7 is also significantly divergent in the PB, as is Axis 3, while Axis 9 shows a non-significant trend (**Table 2**).

Analysis of variance at the level of individual genes was also effective at recovering timepoint specific responses in the BM, but not initially in the PB: **Table 3** lists the number of features significant at a False Discovery Rate of 5% (Benjamini and Hochberg, 1995). Recognizing that animal effects may obscure the temporal differences, we also ran the model with "animal" included as a random statistical effect, and recovered almost twice as many timepoint-responsive transcripts in the BM, and 292 in the PB. Similarly, ANOVA of the metabolome yielded many more significant timepoint-responsive metabolites after inclusion of "animal" as a random effect.

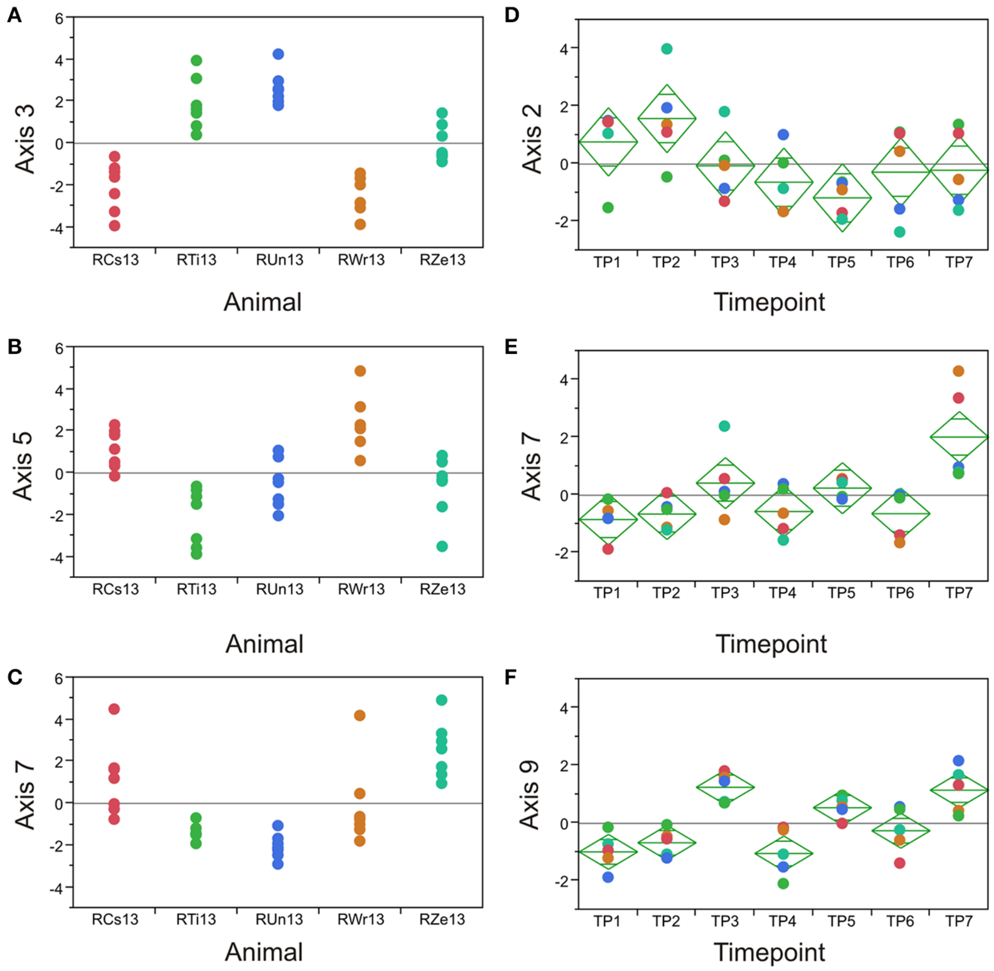


FIGURE 5 | Axis of variance analysis. Each plot shows the indicated Axis score (PC1 of the 10 BIT for the Axis) in the five animals **(A–C)** or at the seven timepoints **(D–F)**. In bone marrow, Axes 7 and 9 are significantly differentiated at TP3, 5, and 7, the post-drug timepoints.

Table 2 | Axes of variance analysis.

Axis	**Bone marrow**			**Peripheral blood**		
	PVE by PC1[a]	**Sig Animal**[b]	**Sig Drug**[c]	**PVE by PC1**[a]	**Sig Animal**[b]	**Sig Drug**[c]
1	45	0.0034	0.0006	43	ns	ns
2	77	8×10^{-6}	0.0061	68	0.0002	0.0025*
3	63	1×10^{-10}	0.0169*	91	2×10^{-13}	ns
4	25	ns	ns	41	0.0005	ns
5	55	5×10^{-6}	0.0022	73	0.0003	ns
6	35	0.0066	0.0005	29	0.0002	ns
7	54	3×10^{-6}	0.0008*	76	3×10^{-5}	0.0267*
8	45	ns	0.0183	58	2×10^{-5}	ns
9	35	0.0002	1×10^{-6}*	44	0.0234	ns

[a] *The percent of variation in the BIT explained by PC1 (>35% implies strong covariance).*

[b] *The signficance of the among-animal effect.*

[c] *The significance of the pre-/post-/inter-drug treatment after removing the animal effect.*

**Implies the post-drug treatment effect was extreme relative to pre- and inter-drug.*

The most interesting timepoint effect with respect to drug exposure is where each of the post-drug timepoints is greater (or less) than the immediately preceding pre-drug timepoint, namely TP3 > TP2, TP5 > TP4, and TP7 > TP6. Again, this situation was only observed in the BM: 73 genes were consistently greater post-drug, and 25 consistently less strongly expressed post-drug, but no genes satisfied this criterion in PB (paired t-test, $p < 0.05$ at each of the three comparisons in 5 animals). A list is provided in the Supplementary data, and is notable for multiple immune-related genes, including TLR4, IL1RAP, IL1RAP, IL10RB, and MAP2K1.

PATHWAY-ORIENTED ENRICHMENT ANALYSES

We next visualized the broad distribution of gene expression in pathways across the time course of pyrimethamine treatment by performing hierarchical clustering of a summary measure of each of 270 KEGG pathways. For each pathway with at least five transcripts expressed in both BM and PB, we generated the first principal component (PC1) of all of the transcripts annotated to the pathway, measured in all five monkeys at seven timepoints. These pathway PC1 values were averaged across the monkeys, and clustered by Ward's method in JMP. **Figure 6** shows heat maps of the average PC1 scores for BM (A) and PB (B) with red corresponding to a high positive score, generally high transcript abundance, and blue a negative score, generally lower abundance on average.

Table 3 | Differential gene expression between timepoints.

Data type	Tissue	Without animal in model	With animal in model
RNA-Seq	Bone marrow	3678	6483
RNA-Seq	Whole blood	0	292
AE MS	Plasma	651	927
C18 MS	Plasma	1254	1992
CBC	Whole blood	10	13

Table shows number of genes significant at 5% FDR rate for each data type, contrasting seven timepoints for RNA-Seq, and 5 timepoints for metabolomics.

In the BM, we observed seven clusters of pathway PC1 scores, with the major division of timepoints grouping TP1, TP2, and TP3 separately from TP4 through TP7. There was no clustering of pathways at the three post-drug timepoints (TP3, TP5, and TP7). In the PB, we observed just six clusters of pathway PC1 scores, with the major division of timepoints separating TP1, TP4, and TP6 from the remainder. Again, there was no evident clustering of the post-drug timepoints. The grouping of TP4 and TP6 corresponds to expected absence of drug, as does the baseline TP1, but this does not seem to relate to drug exposure since TP2 sampled immediately prior to the first drug administration, groups with the post-drug samples.

Comparing both sample types, 10% of the cluster identities in the PB are explained by the cluster identities in the BM (Pearson Chi-square, $p < 10^{-6}$). However, this also means that the majority of the pathways change their average PC1 profile between the BM and the PB. There are nevertheless some interesting clusters.

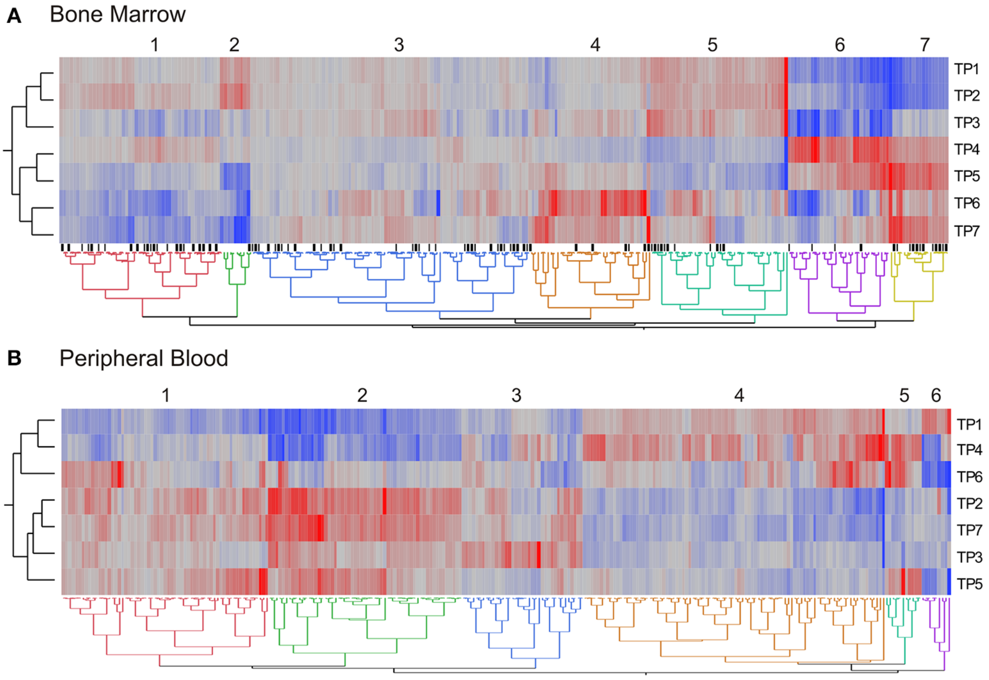


FIGURE 6 | Differential representation of pathways across timepoints. The two-way hierarchical heat maps summarize co-expression of genes within pathways in the bone marrow **(A)** and peripheral blood **(B)**. Data points are the mean PC1 score for each of 270 KEGG pathways. Rows are timepoints, and the major clusters of PC1 scores are indicated. Black tick marks below the heatmap in **(A)** indicate pathways that are significantly different for the contrast of post- versus pre-drug treatment.

For example, the small green cluster 2 to the left in **Figure 6A** that is high prior to drug administration and low at the final three timepoints includes Ras and Rap 1 signaling, purine metabolism, and infectious disease response. By contrast, the yellow cluster 7 that is more highly expressed uniformly after first drug administration includes inflammatory autoimmune pathways, as well as extracellular matrix and cell adhesion. Most of the DNA repair and recombination pathways show the inverse pattern (clusters 4 and 5) implying down-regulation after persistent exposure to pyrimethamine, as might be expected due to reduction of cell division in response to folate inhibition. In the blood, the small blue-green cluster 5 that is high at TP4, TP5, and TP6 involves diverse pathways indicating perturbation of a variety of aspects of cellular physiology during that interval of time.

These trends were not necessarily consistent across all five monkeys. Similar hierarchical clustering of the 270 pathway PC1 scores of all 5 animals showed that two (RCs13 and RWr13) have quite similar profiles, while another two (RUn13 and RZe13) are only similar if TP4 is withdrawn from the analysis. Intriguingly, these pairs of monkeys were each housed together in the same cage, but there is no way of knowing whether that is coincidence or reflects an effect of shared environment. The result does however underline the conclusion that any effect of drug administration is to a large extent animal-specific.

TARGETED AND GENE SET ENRICHMENT ANALYSIS

A disadvantage of the pathway-oriented approach is that it assumes that the covariance of genes within pathways that is captured by PC1 represents the most relevant aspect of perturbed gene expression. A more common approach is to identify differentially expressed genes and then ask whether they are enriched in particular pathways. We thus applied Gene Set Enrichment Analysis (GSEA) to the dataset, focusing on genes that are globally altered at the 5% FDR level following drug treatment, namely different between the two pre-drug samples and all five post- and inter-drug samples. In the BM, analysis of 4178 genes revealed 13 pathways down-regulated following pyrimethamine exposure, and 12 pathways up-regulated, at $p < 0.001$ and FDR $q < 0.01$; these are listed in **Table 4**. The down-regulated pathways reflect functions in the cell-cycle and metabolism including nucleotide biosynthesis and DNA repair, as well as oxidative phosphorylation (and glycolysis/gluconeogenesis trends in the same direction). The up-regulated pathways are all involved in immune signal transduction. Notably, in many cases the gene expression appears to be intermediate at TP3, suggesting a gradual transition in response to first drug administration that was reinforced with subsequent administrations and lasted several months.

Table 4 | Gene set enrichment analysis.

KEGG ID	Pathway name	Size	*p*	FDR *p*
DOWN REGULATED AFTER PYRIMETHAMINE				
3030	DNA replication	26	<0.001	<0.001
4110	Cell cycle	57	<0.001	<0.001
3410	Base excision rep.	18	0.002	<0.001
3420	Nucleotide excis'n	22	0.004	0.006
0072	Ox phosphoryl'n	45	<0.001	<0.001
0010	Glycolysis	23	0.011	0.019
0480	Glutathione	24	<0.001	<0.001
0240	Pyrimidine	37	<0.001	<0.001
0230	Purine	60	0.003	0.006
3040	Spliceosome	46	<0.001	<0.001
5016	Huntington's	56	<0.001	<0.001
5012	Parkinson's	44	<0.001	<0.001
5010	Alzheimer's	53	<0.001	<0.001
5322	SLE	34	0.005	0.008
UP REGULATED AFTER PYRIMETHAMINE				
4660	TCR signaling	44	<0.001	<0.001
4650	NK-mediated cytotox	34	<0.001	<0.001
4630	JAK-STAT signaling	36	<0.001	0.001
4070	PI signaling	19	<0.001	0.001
4210	Apoptosis	23	<0.001	0.011
4662	B cell receptor signaling	21	<0.001	0.014
4370	VEGF signaling	19	0.002	0.013
4150	mTOR signaling	19	0.002	0.015
4310	WNT signaling	42	0.007	0.022
4722	Neurotrophin signaling	40	0.007	0.025
4012	ErbB signaling	24	0.007	0.028
4514	Cell adhesion	37	0.007	0.031

A good example of this is provided by focused analysis of the one-carbon pool by folate pathway, which we expected to be influenced by pyrimethamine, since the drug functions by inhibiting the enzyme dihydrofolate reductase (DHFR). The pathway was too small to include in the GSEA, but nevertheless 12 of 17 genes expressed in the macaque and annotated to KEGG map00670 are positively co-regulated in both BM and PB samples, with PC1 capturing 51% and 40% of the variance, respectively (**Figures 7A,B**). The trajectory of this score trends downward beginning at TP3 in the BM and over half the variance is among timepoints (ANOVA $p < 0.0001$), whereas in the PB there is no differential expression (**Figures 7C,D**).

In the PB, purine (KEGG map00230) and pyrimidine (KEGG map00240) metabolism pathways both show a very large coordinated reduction in PC1 after TP1, namely before the first drug administration, and remain low throughout the experiment. Similarly, oxidative phosphorylation (KEGG map00190) is dominated by a transition that precedes drug administration, as is glycolysis (KEGG map00010), although it occurs in the opposite direction (i.e., gene expression increases). These results suggest that the animals experienced a shift in their major mode of energy production in the circulating blood cells after introduction into the experimental cages. Fatty acid biosynthesis also shows interesting patterns that we do not have space to describe in detail. All of these observations await confirmation at the metabolite level once the annotation of the m/z features on the platform is more advanced.

BAYESIAN NETWORK ANALYSIS OF THE TRANSCRIPTOME

Finally, we adopted an orthogonal exploratory approach to describe networks of highly co-regulated genes. Each of the 1000 genes most differentially expressed relative to drug administration in the BM samples (that is, in the comparison of post- vs. inter-drug timepoints) were carried forward to quality-thresholded

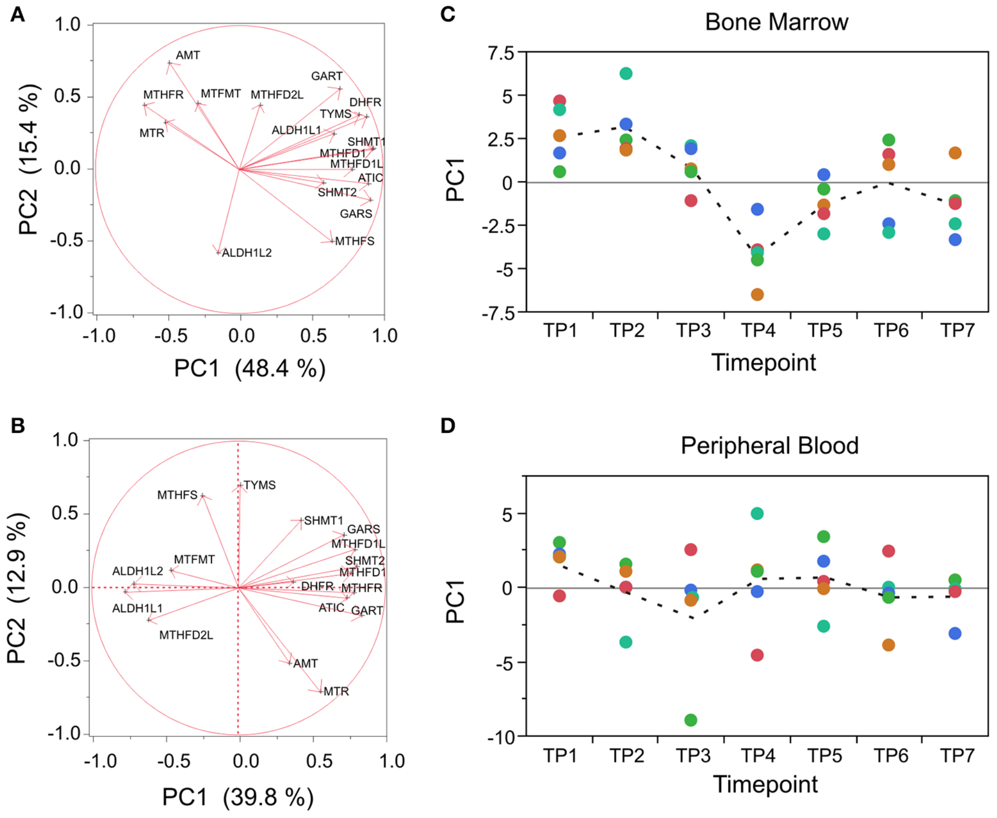


FIGURE 7 | Targeted analysis of the folate pathway. (A,B) Loadings of the first two PC for each of 17 genes in KEGG map00670 (One carbon pool by folate) in BM and PB, also indicating the percent variation explained by each PC. Note that MTHFR switches direction effect between the cell sources. **(C,D)** Corresponding profiles across the timecourse, showing decline in PC1 generally in BM following drug exposure, but no significant differential expression in PB. Colors correspond to the five monkeys as in **Figure 5**.

clustering (De Smet et al., 2002). We identified 26 clusters of 10 or more transcripts, the first 4 of which have at least 50 transcripts each (**Figure 8A**). Permutation of sample labels across timepoints or the entire data set never identified this degree of covariance: full permutation of sample labels for each gene independently recovered zero clusters, while permutation of timepoints within animals for each gene independently yielded just one cluster with two genes, indicating that the clusters found were not artifacts of the underlying data distributions while also increasing confidence that the clusters are biologically motivated. Permutation of animals within timepoints recovered a similar number and size of clusters as the true data, indicating that animal effects had been largely (but not completely) removed in generating the residuals. Since Bayesian network inference is typically much more robust and reliable with many more than the 35 samples available in this study, we increased the effective number of observations per variable by treating each gene (at each timepoint) as an observation of the behavior if its cluster. To do this, we concatenated the top 10 genes most closely correlated with the cluster centroid, yielding 350 observations for each of the 26 clusters. The most robust and likely connections in the emergent networks were then determined by subsampling and permutation as described in the Methods.

An efficient and powerful method for Bayesian network structure learning, the Sparse Candidate Algorithm (Friedman et al., 1999), was used to uncover the potential connections between the clusters. Networks were inferred for 1000 randomly generated subsamples of 90% of the data for each gene to ensure the robustness of the learning results; all connections shown in **Figure 8B** satisfy the criterion that each connection must exist in at least 50% of all of the resampling simulations for the original data, and in each of the three permutations of that data; we found an average overlap of 66.7% of interactions conserved between the original dataset and each of the three shuffles for the BM data, with 13 connections consistently detected in all datasets (see detailed descriptions of robustness testing in the Methods). Core features of this BM network were further investigated by inspection, and validated by gene set enrichment analysis (Subramanian et al., 2005). For instance, cluster 1 shows complimentary patterns to clusters 8 and 22, while it is most similar to clusters 20 and 24. The core genes in each of these clusters suggest functions in immune T-cell responses. Clusters 1 and 3 are "hubs" with a relatively high degree of connectivity in a graph that is otherwise quite sparse.

Although there was little evidence for significant differential expression among the three drug response classes (pre-, post-, and inter-) in the PB, we nevertheless assessed whether the BM cluster modularity may be present in the PB. Projected onto the PB, many of the BM clusters appeared to be co-regulated. To statistically validate this inference, for each cluster we computed the correlation of each gene with the centroid in the PB data, and compared the observed distribution with that of 100 random samples of the same size as that cluster, taken from the 660 transcripts contained in all of the clusters. **Figure 8C** shows the proportion of permutations showing a significant deviation in the direction of stronger

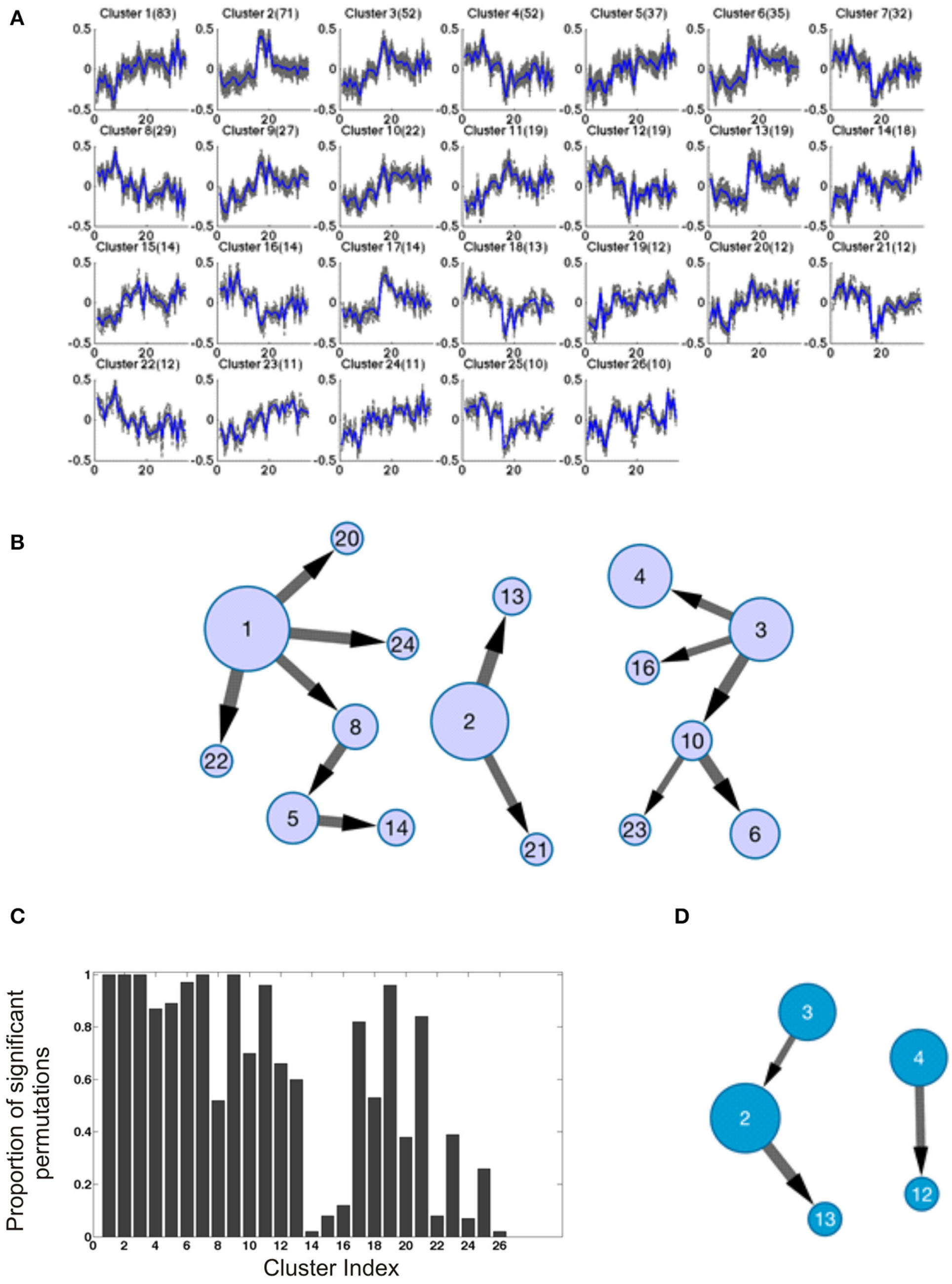


FIGURE 8 | Bayesian Network analysis. (A) The results of quality-based clustering on BM transcriptional data provide tight clusters of coregulated genes used as input for Bayesian network inference. **(B)** The resulting robust network, defined as those connections present in at least 50% of all subsample analyses for each of four different permutations of the dataset. The size of nodes indicates the size of the clusters (also included in **A**), and the size of edges connecting nodes reflects the relative likelihood of a connection based on its overall frequency of occurrence across subsample replicates. **(C)** Statistical testing of the significance of correlations within clusters of PB data derived from BM data clustering. For each cluster, 100 random samples of genes of the same size of the cluster were compared to the PB data using the BM clustering of genes. The distributions of gene profile correlations to the centroid of their cluster were compared using a one-tailed Kolmogorov–Smirnov test. Histogram bars represent the number of random samplings showing statistically significant increases in correlation of the actual data compared to random data. **(D)** The PB network derived using BM clusters; of note there is one conserved connection between this and the BM network.

concordance in the observed data, providing good evidence for cluster conservation in the PB for 12 of the 13 largest clusters, as well as several of the smaller ones. Furthermore, the Bayesian network approach identified a number of robust connections in the PB data, showing an average 37.5% overlap between the original data and any of the permutations, and three connections were observed in common across all permutations (**Figure 8D**). One of those three connections was observed in both the PB and the BM, implying robust dependence of cluster 13 on cluster 2 between the two compartments. Of note, the genes in both of these clusters have been implicated in 6 and 24 h responses to the anti-tumor aminopeptidase inhibitor Tosedostat (Krige et al., 2008). The other two connections (cluster 3 to 2, and cluster 4 to 12) are found in PB but not in BM, suggesting that not only is there conservation of modularity between the compartments, but that new relationships using the modularity of one compartment can be observed in the other.

Application of a similar pipeline to the metabolome data also revealed novel structure to the data. Since there are fewer differentially abundant m/z features in the plasma, this analysis was performed on 500 features for each column with relatively high false discovery rates on the post- vs. inter-drug samples, namely 28% for C18 ($p < 0.01$) and 36% for AE ($p < 0.031$). Quality-thresholded clustering identified 11 and 10 clusters respectively with more than 5 m/z features, and 4 and 3 with more than 10 m/z features. The profiles of the larger clusters in each of the two columns are concordant (**Figures 9A,B**), and most were also recovered in a joint analysis of both columns, with approximately double the number of features. This suggests that the two columns reproduce the same tight clusters of metabolites across animals and times, although the actual m/z values do not match, suggesting that different ionizations or adducts may have been included in the selected features for each column. After Bayesian Network analysis, three robust connections were observed between AE clusters, but none with the C18 data (**Figures 9C,D**).

To investigate possible integration of the metabolic and transcriptional data types, we first simply performed correlation analysis between the centroids of all of the clusters. The strongest interaction that was identified had a correlation coefficient of −0.61. Using Bayesian network inference on all clusters of size greater than 10 between BM transcriptional data and AE column metabolomics data (using only timepoints 3–7 for all based on availability of metabolomics data), we found little in the way of robust connections between the two data types (data not shown). There were no connections conserved across 50% of the subsampling analyses in each of the original and four permuted datasets; however, relaxing this criterion slightly to include any connection present for 50% of all subsampling runs across all four datasets (not 50% in each individual dataset) revealed one potential connection between the two data types.

DISCUSSION

This study with naïve healthy rhesus macaques precedes others that will involve specific infection and treatment regimens, and, importantly, it has served to establish logistics and methodologies for systems biological approaches requiring the monitoring

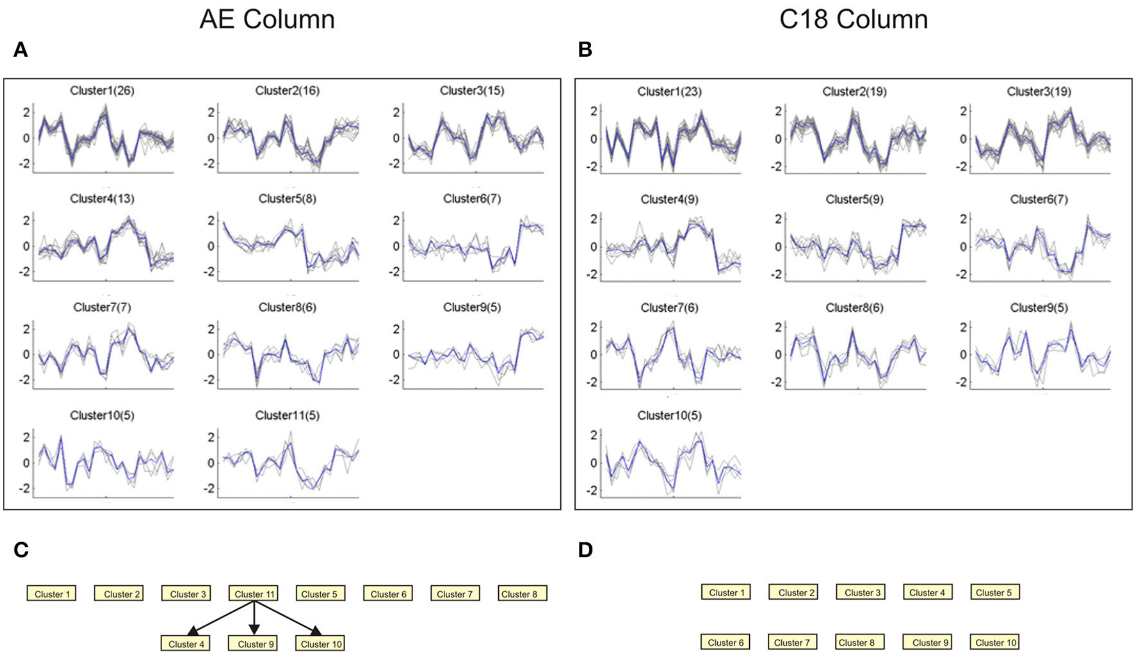


FIGURE 9 | Quality-thresholded clustering of metabolomics data. Each plot shows the standardized levels of the indicated number in parentheses of m/z features in Qt-clusters that have at least 5 features. The order of samples along the x-axis is {RCs18, RTi18, RUn18, RWr18, and RZe18} at TP3, 4, 5, 6, and 7, and the solid blue line indicates the centroid of the cluster. **(A)** AE column. **(B)** C18 column. **(C,D)** Bayesian networks assembled on clusters with five metabolites and seven levels of discretization, and 80% subsampling threshold, similar to the analyses used to generate the transcriptomic networks in **Figure 8B**.

and evaluation of clinical data, collection of PB and BM samples, and the integrative analysis of multiple omics and other datasets. The biological objective of this study was to use the combined power of transcriptomic and metabolomic profiling to investigate the effects of an anti-malarial drug on the physiology of the PB and BM. Five rhesus macaques were injected with a preparation of uninfected *Anopheles dirus* salivary glands, to mimic an inoculation of *Plasmodium* sporozoites, and then followed for 100 days with intermittent administration of pyrimethamine, a drug known to have an effect on the BM.

Ideally the unbiased top-down analytical approach that we adopted would identify components of variation in both the transcriptional and metabolomic domains that covary with drug administration, and enrichment analysis of both would point to a common aspect of metabolic regulation such as nucleotide biosynthesis. To some extent we were thwarted in this objective by three findings: (i) there is very low correspondence between the transcriptome and metabolome and no major components of variation correlate with repeated pyrimethamine administration; (ii) among animal effects dominate the transcriptome raising the possibility that pyrimethamine responses are variable among individuals and obscure any common response; and (iii) although the metabolomics platform reports thousands of features, annotation is not yet robust enough to support global enrichment analysis in this dataset (but see Li et al., 2013, for encouraging developments).

Additionally, we must acknowledge that this is a relatively small study, with just five monkeys and seven timepoints. The failure to detect strong drug responses or covariance of the blood and transcriptome may simply be a function of lack of statistical power. For example, although we can attribute the largest PC to specific sources of variance, those explaining much less than 10% of the variance might be regarded as noise, and are unlikely to replicate. That is one reason why we turned to the Axis of Variation analysis, since the axes have a more biological basis that is not as dependent on sample size. The pathway-oriented analysis also highlights how interpretation of single gene effects must be placed within the context of the major sources of variance, in this case animal effects and some temporal shifts that may not relate to drug administration. It is likely that much larger studies would be required to detect strong transcriptome-metabolome covariance: for example, our analysis of 20 strains of Drosophila profiled on four diets with considerable technical replication did suggest some specific examples of covariance despite general absence of correspondence of the major PC axes, and even in the presence of large genotype-by-diet interactions (Reed et al., 2014). Studies with hundreds of NHPs are impractical, so we must make do with analytical methods such as those reported here, but recognizing that there is low power and the potential to over-interpret those associations that are detected.

Nevertheless, several key findings contradict our expectations and highlight aspects of the biology that emerge from multi-omic analyses. Most striking is the magnitude of the among-animal differentiation of the transcriptome. This is even stronger in the PB than the BM, with almost perfect clustering of all seven timepoint samples within animals. Each of the major PC and Axes of variation are significantly different among animals. In the BM, an unknown variable caused TP4 to generate a markedly different profile common to all five macaques, yet the individual profiles return to the animal-specific baseline within weeks. Persistent among-individual differential expression in the PB has also been reported in humans (Whitney et al., 2003; GG unpublished), but here we demonstrate for the first time that the differential expression is initiated in the BM and the data suggests that it precedes and is independent of individual-specific environmental influences faced in the PB. Persistent inter-individual variation is less marked in the metabolome, but nevertheless present as previously observed by Park et al. (2009) in a human dietary intervention study.

The temporal component of transcriptional variance shows some sign of cycling with drug administration, but it is dominated by timepoints that are not primarily associated with recent administration of the drug. In the PB, TP1 is most divergent, possibly indicating incomplete acclimatization of the animals to their new experimental housing and experimental procedures. Indeed, one of the pathways elevated at TP1 denotes a generalized stress response, as we have previously observed in captive relative to free range red wolves (Kennerly et al., 2008). This effect is much less in the BM, and since TP1 and TP2 differ from the remaining timepoints for 2 of the first 5 PC and 3 of 9 Axes (e.g., **Figures 5D,F**). Consistent with the observation that pyrimethamine has an effect on the BM (Wickramasinghe and Litwinczuk, 1981) our data indicates that in the BM there is a global impact of pyrimethamine that persists throughout the experiment following the first administration. Then at TP4 in the BM and TP4 and especially TP5 in the PB, there is further differentiation of gene expression consistent with a heightened response to the drug. TP4 is an inter-treatment timepoint, over 20 days after the previous administration at a time when pyrimethamine should no longer be in circulation based on its half-life of 140 h (Almond et al., 2000). Based on this figure there should nevertheless be around one third of the administered dose still available at the post-drug timepoints (TP3, 5, and 7), but we do not know to what extent it would be directly available to cells in the BM or circulating in the blood. Consequently, it is possible that the relatively weak drug effects are because the animals are no longer functionally exposed to pyrimethamine at the sampled timepoints. We have been unable to correlate the change at TP4 with any variable such as a change in handler or cage conditions. The null hypothesis of no differential expression across time is rejected, but we do not have a clear alternate hypothesis for the effect.

In the metabolome, there is very good correspondence between the PC and the hierarchical cluster profiles of the two columns, but the major variance components do not correlate with either animal or drug response. Since retention times differ between the columns, and m/z peaks included in feature selection may be from different adducts for several metabolites, it is not straightforward to combine the analysis of both columns. The major temporal effect is at TP7, which shows a correlated response across all five animals. It is unclear whether this represents a long-term effect of more than two months of drug treatment, or some other unidentified stimulus, but it has no correlate in the transcriptome. Several hundred metabolite features are globally different in the post-drug samples, even though the major PC also differentiate the post- and inter-drug timepoints.

The antimalarial drug pyrimethamine interferes with folate by inhibiting the enzyme dihydrofolate reductase and disrupts the parasite life cycle by interfering with nucleotide metabolism and replication. It also affects host metabolism, and in fact folate supplementation is often used to sustain healthy erythropoiesis in pregnant women and infants (Titaley et al., 2010). Gene set enrichment analysis (Subramanian et al., 2005) of the transcriptome provides some evidence for effects on metabolism and cell division. Briefly, contrasting the pre- with the post- and inter-drug timepoints, some common pathways between the BM and PB, as well as BM-specific changes, are observed. The former are of a metabolic nature, including oxidative phosphorylation, pentose phosphate, glyoxylate, butanoate, and linoleic acid metabolism; the latter include multiple KEGG pathways related to the cell cycle such as DNA replication, recombination, and repair. Our Bayesian network approach also focuses attention on regulation of cell division since one of the key enrichments in the BM is with targets of Tosedostat, an anti-cancer drug that antagonizes aminopeptidase activity (DiNardo and Cortes, 2014). Results such as this generate hypotheses that can be tested by targeted metabolomics and manipulation of gene expression, suggesting a new integrative genomics approach to pharmacogenetics.

Our top down analyses also provide some important lessons regarding the joint use of different data integration strategies in MaHPIC (or similar) experiments where a relatively small number of individuals will be followed longitudinally during an intervention. While the principal components approach efficiently defines the major sources of variation, it misses important biological results and is not obviously the best strategy for integration of multiple omic and immunological measures. In particular, the axis of variation analysis picks up effects of drug administration on broad aspects of immune function, most notably interferon-related gene activity highlighted by Axis 7 in both PB and BM samples. It is unlikely in this case that the elevation of this axis is due to viral activity, but this result and weaker evidence for dysregulation of Axes 2 and 9 in the week after pyrimethamine administration show that the network of immune interactions is perturbed and that drug activity is not narrowly restricted to the immediate effects of folate. Finally, given the small number of animals and timepoints in this experiment, statistical power is low for formal hypothesis testing, but we begin to show how Bayesian Network analysis can tease out interaction effects that are not evident in univariate analysis or in analyses designed to capture the largest overall components of variance. The two immune compartments share clusters of co-regulated gene modules, but the connectivity of these differs between BM and PB samples. *Plasmodium* infection will have a much larger impact on the animals' physiology than the mock-inoculations described here, providing ample opportunity for exploring network-based modeling of the host-parasite interactions that underlie malaria infections, immunity, pathogenesis, and severe disease.

ACKNOWLEDGMENTS

We thank the following individuals for their essential contributions. The Yerkes veterinarian and research resources team and Trenton Hoffman assisted with the clinical follow-up of the animals. Zachary Johnson performed the HiSeq2000 sequencing at the Yerkes Genomics Core. We are especially grateful to Aleksey Zimin and Rob Norgren for providing revisions to the annotated *M. mulatta* genome sequence.

FUNDING

This project has been funded in part with funds from the National Institute of Allergy and Infectious Diseases; National Institutes of Health, Department of Health and Human Services [Contract No. HHSN272201200031C], and in part by ORIP/OD P51OD011132 (formerly NCRR P51RR000165).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fcell.2014.00054/abstract

REFERENCES

Aliferis, C., Tsamardinos, I., and Statnikov, A. (2003). *Causal Explorer: a probabilistic Network Learning Toolkit for Biomedical Discovery*. METMBS. Available online at: http://www.dsl-lab.org/ml_tutorial/Publications/Causal_Explorer.pdf

Almond, D. S., Szwandt, I. S., Edwards, G., Lee, M. G., and Winstanley, P. A. (2000). Disposition of intravenous pyrimethamine in healthy volunteers. *Antimicrob. Agents Chemother.* 44, 1691–1693. doi: 10.1128/AAC.44.6.1691-1693.2000

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

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq - a Python framework to work with high-throughput sequencing data. *bioRxiv* doi: 10.1101/002824. Available online at: http://biorxiv.org/content/biorxiv/early/2014/08/19/002824.full.pdf

Bang, J. W., Crockford, D. J., Holmes, E., Pazos, F., Sternberg, M. J., Muggleton, S. H., et al. (2008). Integrative top-down system metabolic modeling in experimental disease states via data-driven Bayesian methods. *J. Proteome Res.* 7, 497–503. doi: 10.1021/pr070350l

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

Biswas, S., Storey, J. D., and Akey, J. M. (2008). Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics* 9:244. doi: 10.1186/1471-2105-9-244

Boedigheimer, M. J., Wolfinger, R. D., Bass, M. B., Bushel, P. R., Chou, J. W., Cooper, M., et al. (2008). Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics* 9:285. doi: 10.1186/1471-2164-9-285

Bumgarner, R. E., and Yeung, K. Y. (2009). Methods for the inference of biological pathways and networks. *Methods Mol. Biol.* 541, 225–245. doi: 10.1007/978-1-59745-243-4_11

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., and Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* 18, 735–746. doi: 10.1093/bioinformatics/18.5.735

Devonshire, A. S., Elaswarapu, R., and Foy, C. A. (2010). Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. *BMC Genomics* 11:662. doi: 10.1186/1471-2164-11-662

Deye, G. A., Gettayacamin, M., Hansukjariya, P., Im-erbsin, R., Sattabongkot, J., Rothstein, Y., et al. (2012). Use of a rhesus Plasmodium cynomolgi model to screen for anti-hypnozoite activity of pharmaceutical substances. *Am. J. Trop. Med. Hyg.* 86, 931–935. doi: 10.4269/ajtmh.2012.11-0552

DiNardo, C. D., and Cortes, J. E. (2014). Tosedostat for the treatment of relapsed and refractory acute myeloid leukemia. *Expert Opin. Investig. Drugs* 23, 265–272. doi: 10.1517/13543784.2014.864276

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863–14868. doi: 10.1073/pnas.95.25.14863

Frevert, U., and Nacer, A. (2013). Immunobiology of *Plasmodium* in liver and brain. *Parasite Immunol.* 35, 267–282. doi: 10.1111/pim.12039

Friedman, N., Nachman, I., and Peer, D. (1999). "Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99)* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 206–215.

Galinski, M. R., Meyer, E. V., and Barnwell, J. W. (2013). *Plasmodium vivax*: modern strategies to study a persistent parasite's life cycle. *Adv. Parasitol.* 81, 1–26. doi: 10.1016/B978-0-12-407826-0.00001-1

Giuliani, A., Filippi, S., and Bertolaso, M. (2014). Why network approach can promote a new way of thinking in biology. *Front. Genet.* 5:83. doi: 10.3389/fgene.2014.00083

González, I., Cao, K. A., Davis, M. J., and Déjean, S. (2012). Visualising associations between paired "omics" data sets. *BioData Min.* 5:19. doi: 10.1186/1756-0381-5-19

Hafalla, J. C., Silvie, O., and Matuschewski, K. (2011). Cell biology and immunology of malaria. *Immunol. Rev.* 240, 297–316. doi: 10.1111/j.1600-065X.2010.00988.x

Hartemink, A. J. (2001). "Discretization of genomic expression data," in *Principled Computational Methods for Validation and Discovery of Genetic Regulatory Networks.* PhD Thesis, Massachussets Institute of Technology.

Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of co-expressed genes. *Genome Res.* 9, 1106–1115. doi: 10.1101/gr.9.11.1106

Jones, D. P., Park, Y., and Ziegler, T. R. (2012). Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu. Rev. Nutr.* 32, 183–202. doi: 10.1146/annurev-nutr-072610-145159

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kennedy, M., Fishbaugher, M. E., Vaughan, A. M., Patrapuvich, R., Boonhok, R., Yimamnuaychok, N., et al. (2012). A rapid and scalable density gradient purification method for *Plasmodium* sporozoites. *Malar. J.* 11:421. doi: 10.1186/1475-2875-11-421

Kennerly, E., Ballmann, A., Martin, S., Wolfinger, R., Gregory, S., Stoskopf, M., et al. (2008). A gene expression signature of confinement in peripheral blood of red wolves (*Canis rufus*). *Mol. Ecol.* 17, 2782–2791. doi: 10.1111/j.1365-294X.2008.03775.x

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Krige, D., Needham, L. A., Bawden, L. J., Flores, N., Farmer, H., Miles, L. E., et al. (2008). CHR-2797: an antiproliferative aminopeptidase inhibitor that leads to amino acid deprivation in human leukemic cells. *Cancer Res.* 68, 6669–6679. doi: 10.1158/0008-5472.CAN-07-6627

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., et al. (2013). Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* 9:e1003123. doi: 10.1371/journal.pcbi.1003123

Moreno, A., Cabrera-Mora, M., Garcia, A., Orkin, J., Strobert, E., Barnwell, J. W., et al. (2013). *Plasmodium coatneyi* in rhesus macaques replicates the multisystemic dysfunction of severe malaria in humans. *Infect. Immun.* 81, 1889–1904. doi: 10.1128/IAI.00027-13

Park, Y., Kim, S. B., Wang, B., Blanco, R. A., Le, N. A., Wu, S., et al. (2009). Individual variation in macronutrient regulation measured by proton magnetic resonance spectroscopy of human plasma. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 297, R202–R209. doi: 10.1152/ajpregu.90757.2008

Pei, B., and Shin, D. G. (2012). Reconstruction of biological networks by incorporating prior knowledge into Bayesian network models. *J. Comput. Biol.* 19, 1324–1334. doi: 10.1089/cmb.2011.0194

Preininger, M., Arafat, D., Kim, J., Nath, A. P., Idaghdour, Y., Brigham, K. L., et al. (2013). Blood-informative transcripts define nine common axes of peripheral blood gene expression. *PLoS Genet.* 9:e1003362. doi: 10.1371/journal.pgen.1003362

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biol.* 14:R95. doi: 10.1186/gb-2013-14-9-r95

Reed, L. K., Lee, K., Zhang, Z., Rashid, L., Poe, A., Hsieh, B., et al. (2014). Systems genomics of metabolic phenotypes in wild-type *Drosophila melanogaster*. *Genetics* 197, 781–793. doi: 10.1534/genetics.114.163857

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., et al. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7:3. doi: 10.1186/1471-2199-7-3

Schwenk, R. J., and Richie, T. L. (2011). Protective immunity to pre-erythrocytic stage malaria. *Trends Parasitol.* 27, 306–314. doi: 10.1016/j.pt.2011.02.002

Soltow, Q. A., Strobel, F. H., Mansfield, K. G., Wachtman, L., Park, Y., and Jones, D. P. (2013). High-performance metabolic profiling with dual chromatography-Fourier-transform mass spectrometry (DC-FTMS) for study of the exposome. *Metabolomics* 9, S132–S143. doi: 10.1007/s11306-011-0332-1

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics* 14:91. doi: 10.1186/1471-2105-14-91

Stanisic, D. I., Barry, A. E., and Good, M. F. (2013). Escaping the immune system: how the malaria parasite makes vaccine development a challenge. *Trends Parasitol.* 29, 612–622. doi: 10.1016/j.pt.2013.10.001

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

Tachibana, S., Sullivan, S. A., Kawai, S., Nakamura, S., Kim, H. R., Goto, N., et al. (2012). *Plasmodium cynomolgi* genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat. Genet.* 44, 1051–1055. doi: 10.1038/ng.2375

Titaley, C. R., Dibley, M. J., Roberts, C. L., and Agho, K. (2010). Combined iron/folic acid supplements and malaria prophylaxis reduce neonatal mortality in 19 sub-Saharan African countries. *Am. J. Clin. Nutr.* 92, 235–243. doi: 10.3945/ajcn.2009.29093

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Uppal, K., Soltow, Q. A., Strobel, F. H., Pittard, W. S., Gernert, K. M., Yu, T., et al. (2013). xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics* 14:15. doi: 10.1186/1471-2105-14-15

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28, 2184–2185. doi: 10.1093/bioinformatics/bts356

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, 1896–1901. doi: 10.1073/pnas.252784499

WHO World Malaria Report. (2013). Available online at: http://www.who.int/malaria/publications/world_malaria_report_2013/en/

Wickramasinghe, S. N., and Litwinczuk, R. A. (1981). Effects of low concentrations of pyrimethamine on human bone marrow cells *in vitro*: possible implications for malaria prophylaxis. *J. Trop. Med. Hyg.* 84, 233–238.

Wilhelm, B. T., and Landry, J. R. (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48, 249–257. doi: 10.1016/j.ymeth.2009.03.016

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8, 625–637. doi: 10.1089/106652701753307520

Wright, G. J., and Rayner, J. C. (2014). *Plasmodium falciparum* erythrocyte invasion: combining function with immune evasion. *PLoS Pathog.* 10:e1003943. doi: 10.1371/journal.ppat.1003943

Yu, T., Park, Y., Johnson, J. M., and Jones, D. P. (2009). apLCMS - adaptive processing of high-resolution LC/MS data. *Bioinformatics* 25, 1930–1936. doi: 10.1093/bioinformatics/btp291

Zhang, S. J., Liu, C. J., Yu, P., Zhong, X., Chen, J. Y., Yang, X., et al. (2014). Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. *Mol. Biol. Evol.* 31, 1309–1324. doi: 10.1093/molbev/msu084

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

Received: 30 May 2014; accepted: 08 September 2014; published online: 08 October 2014.

Citation: Lee KJ, Yin W, Arafat D, Tang Y, Uppal K, Tran V, Cabrera-Mora M, Lapp S, Moreno A, Meyer E, DeBarry JD, Pakala S, Nayak V, Kissinger JC, Jones DP, Galinski M, Styczynski MP and Gibson G (2014) Comparative transcriptomics and metabolomics in a rhesus macaque drug administration study. Front. Cell Dev. Biol. ***2****:54. doi: 10.3389/fcell.2014.00054*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 04 November 2014
doi: 10.3389/fcell.2014.00059

Multi-omic landscape of rheumatoid arthritis: re-evaluation of drug adverse effects

Paolo Tieri[1,2]*†, XiaoYuan Zhou[2]†, Lisha Zhu[2] and Christine Nardini[2]*

[1] IAC - Istituto per le Applicazioni del Calcolo "Mauro Picone," CNR - Consiglio Nazionale delle Ricerche, Rome, Italy
[2] Group of Clinical Genomic Networks, Key Laboratory of Computational Biology, Chinese Academy of Sciences - Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, China

Edited by:
Radhakrishnan Nagarajan, University of Kentucky, USA

Reviewed by:
Hong-Wen Deng, Tulane University School of Public Health and Tropical Medicine, New Orleans, USA
Ayse Meric Ovacik, Merck, USA

***Correspondence:**
Paolo Tieri, Istituto per le Applicazioni del Calcolo, Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Rome, Italy
e-mail: p.tieri@iac.cnr.it;
Christine Nardini, Group of Clinical Genomic Networks, Key Laboratory of Computational Biology, Chinese Academy of Sciences - Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, China
e-mail:
christine.nardini.rsrc@gmail.com

† These authors have contributed equally to this work.

Objective: To provide a frame to estimate the systemic impact (side/adverse events) of (novel) therapeutic targets by taking into consideration drugs potential on the numerous districts involved in rheumatoid arthritis (RA) from the inflammatory and immune response to the gut-intestinal (GI) microbiome.

Methods: We curated the collection of molecules from high-throughput screens of diverse (multi-*omic*) biochemical origin, experimentally associated to RA. Starting from such collection we generated RA-related protein-protein interaction (PPI) networks (*interactomes)* based on experimental PPI data. Pharmacological treatment simulation, topological and functional analyses were further run to gain insight into the proteins most affected by therapy and by multi-*omic* modeling.

Results: Simulation on the administration of MTX results in the activation of expected (apoptosis) and adverse (nitrogenous metabolism alteration) effects. Growth factor receptor-bound protein 2 (GRB2) and Interleukin-1 Receptor Associated Kinase-4 (IRAK4, already an RA target) emerge as relevant nodes. The former controls the activation of inflammatory, proliferative and degenerative pathways in host and pathogens. The latter controls immune alterations and blocks innate response to pathogens.

Conclusions: This multi-omic map properly recollects in a single analytical picture known, yet complex, information like the adverse/side effects of MTX, and provides a reliable platform for *in silico* hypothesis testing or recommendation on novel therapies. These results can support the development of RA translational research in the design of validation experiments and clinical trials, as such we identify GRB2 as a robust potential new target for RA for its ability to control both synovial degeneracy and dysbiosis, and, conversely, warn on the usage of IRAK4-inhibitors recently promoted, as this involves potential adverse effects in the form of impaired innate response to pathogens.

Keywords: rheumatoid arthritis, multi-*omic* data integration, host-microbiome interface, protein-protein interaction, network topology

INTRODUCTION

Rheumatoid arthritis (RA) is a multifaceted autoimmune, chronic and inflammatory disease with, to date, unclear etiology. As a consequence of its complexity, the definition of efficient and effective therapies remains a remarkable challenge due to the difficulties in controlling side effects and adverse events in relation to known (like genetic susceptibility, Stahl et al., 2010) and emergent (epigenomic factors, Nakano et al., 2012, dysbiosis, Scher and Abramson, 2011) RA-associated con-causes.

Recently, translational research has welcomed into medicine a number of novel perspectives. Among these, sequencing technologies (*omic* screens) and computational intensive approaches (systems biology) now coagulate into a practice where technology and mathematical modeling serve basic research in the production of selected hypotheses, which testing *in vitro*, *in vivo* and ultimately in clinical studies can support medical research and practice (Okada et al., 2014; You et al., 2014). The recent acknowledgment of the importance and complexity of the gut intestinal (GI) microbiome in the onset, progression and regression of RA (Scher and Abramson, 2011; Scher et al., 2012, 2013) and other autoimmune diseases, requires to incorporate the effects on the GI microbiome for any novel therapy. While protocols and medical best practice recommendations become mature in this direction, we propose the use of network approaches and *omics* from diverse origins (i.e., different biochemical districts/compartments/layers) including genomics, epigenomics, transcriptomics, post-transcriptomics, proteomics, and host-microbiome interface to GI metagenomics, to appropriately monitor the complexity of the disease. The novelty of the present work, therefore, lies not only in its application to RA, but also in the number of *omic* layers we have used, from genomic to proteomic and including the host-microbiome interface. These novelties allow to draw a single analytical picture of the fragmented molecular information available to date on RA, an easily consultable and extendable reference map for the researchers in the field, and—importantly—a systemic evaluation on the impact of a recently proposed RA therapeutic target (IRAK4), valuable *per se* and as an exemplar application of this approach. Overall, this work contributes to the general debate about data integration by offering details on our methodology, and to the area of complex inflammatory diseases, by providing specific examples of data choice and operational results.

METHODS

MAP CONSTRUCTION

The datasets used to construct the map are gathered from 13 different sources from databases and literature (**Table 1**). We included molecules experimentally associated to RA from manual curation of literature sources (*core* dataset, *CD*, 377 proteins, Data Sheet 1, Tables S1–S6), and additional molecules and pathways strongly yet not explicitly associated to RA (*extended* dataset, *ED*, 4709 proteins, Data Sheet 1, Tables S3A–E, S7–S13). A summary of all datasets and proteins' Uniprot IDs is provided in Data Sheet 1, Table S14. While the *core* set constitutes a more specific RA map, its extension offers a more systemic and practically usable map, notably in terms of the significance of the statistics that can be run on the extended map. The map presented here assembles genomic, epigenomic, transcriptomic, post-transcriptomic, proteomic, and host-microbiome interface data related to RA, as detailed below, and integrates such information at the functional level of protein-protein interactions (PPIs). The PPI framework is an assessed integrative approach (Hodgman, 2007; Dittrich et al., 2008; Jin et al., 2008; Kim et al., 2010; Iskar et al., 2012) that has already been used in computational biology to understand diseases' pathogenesis (Huang et al., 2009b); to implement tools for the interpretation of inferred gene and protein lists (Berger et al., 2007; Antonov et al., 2009); to prioritize cancer-associated genes (Wu et al., 2012); to predict functional linkages among genes (Lehner and Lee, 2008); to show the implication of protein networks topology in genetics, personal genomics, and therapy (Lee et al., 2013); to implement data integration workflows showcased in obstructive nephropathy in children (Moulos et al., 2011).

CORE DATASET

The CD is composed of 377 proteins retrieved from six data sources (Data Sheet 1, Tables S1–S6):

1) RA genome-wide association studies (GWAS) gathered and integrated from five different databases (BioGPS (Wu et al., 2009), HuGE (Yu et al., 2008), NHGRI, OMIM, PharmGKB (Klein et al., 2001); see Data Sheet 1, Table S1 for the specific query processes);
2) RA-associated proteins from the Universal Protein Resource (Uniprot) (Consortium, 2010), retrieved using as search parameters "*rheumatoid arthritis*" and "*human*" and then manually screened (Data Sheet 1, Table S2);
3) Genes and proteins retrieved from a comprehensive review of the literature, in particular genes appearing in **Tables 1**, **2** of Review (Mcinnes and Schett, 2011) and cited references (Data Sheet 1, Table S3);
4) Genes that show epigenetic changes in relation to RA, as specified in Trenkmann et al. (2010); Karouzakis et al. (2011) (Data Sheet 1, Table S4);

Table 1 | Data sources, subsets and number of elements of the RA map.

Subset Id.	Source of subset	Main dataset destination	No. of proteins in subset	Total no. of proteins in main dataset	No. of proteins (and PPIs) in the interactome map	No. of proteins (and PPIs) in the interactome map: main cluster
1	GWAS	*Core*	223	377	303 (597)	161 (542)
2	UNIPROT	*Core*	49			
3	Literature review	*Core*	53			
4	Methylation	*Core*	37			
5	Exp. valid. microb. interface	*Core*	54			
6	NF-κB consensus	*Core*	16			
3A	T cell activation pathways	*Extended*	1248	4709	3783 (24457)	3466 (24364)
3B	Other pathways	*Extended*	283			
3C	Cytokines	*Extended*	1536			
3D	Growth and differentiation	*Extended*	472			
3E	Intracell signaling and TFs	*Extended*	1837			
7	Transcriptional RA map	*Extended*	212			
8	RA-miRNA reg. proteins	*Extended*	1652			
9A	Downreg. genes in RA	*Extended*	451			
9B	Upreg. genes in RA	*Extended*	210			
10	Inflammasomes	*Extended*	152			
11	Adenosine receptors	*Extended*	569			
12	GPCRs	*Extended*	364			
13	Microbiome interface	*Extended*	171			

5) Proteins that are at the interface between the host and the oral microbiome, in particular proteins experimentally known to be differentially expressed in presence of *Porphyromonas Gingivalis* (Zhou and Amar, 2006), a periodontitis-causing bacterium that has been strongly linked to the insurgence of RA (Mikuls et al., 2012; Scher et al., 2012; Smit et al., 2012; Bingham and Moni, 2013; Ogrendik, 2013; Okada et al., 2013) (Data Sheet 1, Table S5);
6) The key elements of the NF-κB system, the master regulator of inflammation (Oeckinghaus et al., 2011; Smale, 2011; Hayden and Ghosh, 2012) at the center of a complex regulatory interactome (Tieri et al., 2012) prominently implicated in the onset and development of RA (Miagkov et al., 1998; Makarov, 2001; Feldmann et al., 2002; Okamoto, 2006; Roman-Blas and Jimenez, 2006, 2008; Simmonds and Foxwell, 2008; Van Loo and Beyaert, 2011): we included 16 "consensus" proteins that appear at the intersection of the three main NF-κB-related datasets described in Tieri et al. (2012) (Data Sheet 1, Table S6).

EXTENDED DATASET

The *extended* dataset (ED, that includes CD) is composed of 4709 proteins, which are involved in a broader sense in the onset and development of RA, such as proteins participating in signaling pathways or cascades of recognized importance for RA. This extension provides a more general setting for the molecular framing of RA, and offers a larger network to operate on, with more relevant statistics and analyses, giving account for contributions coming from entities that may have been neglected or that are not experimentally related to RA, but that participate to the inception of the disease. In addition to the proteins of the *core* dataset, we added eight main subsets, as follows (Data Sheet 1, Tables S3A–E, S7–S13):

3A-B-C-D-E) in retrieving data from Mcinnes and Schett (2011) and references cited there, we considered that some of the key proteins can be "hidden" inside the signaling pathways involved in the disease. In order to take into account such potentially important and usually neglected elements, we expanded subset 3 of CD by a pathway enrichment analysis process, using the genes listed in Mcinnes and Schett (2011) **Tables 1**, **2**. To populate these five subsets, the selected genes have been input in the *pathway over-representation analysis* (ORA) tool of InnateDB, one of the most comprehensive sources of pathways data available (Lynn et al., 2008; Breuer et al., 2013). Pathway ORA has been performed on InnateDB using hypergeometric distribution for *p*-value computation and Benjamini–Hochberg correction method for multiple hypothesis testing. All the proteins participating to such over-represented pathways were then included. We retrieved respectively: 39 enriched pathways accounting for 1248 proteins (subset 3A), 14 pathways and 283 proteins (3B), 46 pathways and 1536 proteins (3C), 5 pathways and 472 proteins (3D), and 92 pathways and 1837 proteins (3E), all collected in Data Sheet 1, Tables S3A–E;

7) Genes derived from the transcriptional RA map in Wu et al. (2010) (Data Sheet 1, Table S7);
8) RA-related miRNA-regulated genes: experimentally validated target genes of all miRNAs that are associated to RA in the database miRWalk (Dweep et al., 2011) (search mode: *holistic view of validated disease-miRNA interactions*; web reference: http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/disease.html; query keywords: *Arthritis* AND *Rheumatic diseases*) (Data Sheet 1, Table S8);

9A,B) gene expression profiles of RA patients and healthy controls were searched on Gene Expression Omnibus (GEO, (Barrett et al., 2011) http://www.ncbi.nlm.nih.gov/geo/) with the query ["rheumatoid arthritis" AND "(synovi* OR blood)"] (i.e., in synovial tissue and/or blood). In order to include only highly consistent information, datasets without pre-treatment samples, with no details about the therapy and no raw data were filtered out. Human PBMCs collected and processed by Affymetrix technology were selected, leaving only one dataset out of the initial 61, GSE7524, which contains transcriptomic profiles of 2 healthy controls, 2 before and 2 after anti-TNFα treatment samples. Affymetrix Human Genome U133A Array was used to measure the expression levels of ∼14,500 well-characterized human genes. The raw data were preprocessed using *affy* package (Gautier et al., 2004) in R (http://www.r-project.org/), normalized using robust multi-array average (*rma*) (Irizarry et al., 2003) and for multiple probes corresponding to the same gene, the probe with the highest standard variation across all samples was used to represent the gene. Differentially expressed genes [fold-change (Murie et al., 2009) =2] were identified with the comparison between the 2 healthy controls and the 2 before anti-TNFα treatment samples resulting in 646 genes differentially expressed, among which 440 genes (451 proteins) were down-regulated and 206 genes (210 proteins) were up-regulated (Data Sheet 1, Tables S9A,B);

10) Proteins related to the inflammasome, a multiprotein oligomer responsible for activation of inflammatory processes proteins, which is also known to be activated from the bacterium *P. Gingivalis*, among others, and recognized to play a relevant role in RA (Sidiropoulos et al., 2008; Kolly et al., 2010; Farquharson et al., 2012; Mathews et al., 2013) (Data Sheet 1, Table S10). This set was retrieved using *ORA* as described in 3A-B-C-D-E;
11) Adenosine receptors and related proteins, known to be involved in RA (Varani et al., 2010, 2011; Vincenzi et al., 2013) and possibly at the basis of the mechanism of action of methotrexate, first-line therapy for the treatment of RA (Stamp et al., 2012) (Data Sheet 1, Table S11). This set was retrieved using *ORA* as in 3A-B-C-D-E and 10;
12) The large family of G Protein Coupled Receptors (GPCRs) (Hutchings et al., 2010; Lozupone et al., 2012; Maynard et al., 2012; Tremaroli and Backhed, 2012), pertaining to host-microbiome interface proteins (grouped in a separate set from 13 due to their numerosity), retrieved from http://www.iuphar-db.org/DATABASE/ReceptorFamiliesForward?type=GPCR (Sharman et al., 2013) (Data Sheet 1, Table S12);
13) The set of host-microbiome interacting proteins, manually curated from recent reviews (Lozupone et al., 2012;

Maynard et al., 2012; Tremaroli and Backhed, 2012), to describe the bridge between innate immunity (altered in RA) and the GI microbiome [known to be involved in immune diseases in general and in RA in particular (Scher and Abramson, 2011)]. Globally this dataset accounts for the Toll-like Receptor family (TLRs), the mucin proteins family, selected Immunoglobulins (Ig) and their receptors, among others (Data Sheet 1, Table S13).

Datasets are integrated at the PPI level as peers to avoid introducing any bias *a priori* in the network construction and to warrant that these data are connected in a biologically meaningful way. Protein-protein interactions were retrieved in Cytoscape from the Agile Protein Interaction DataAnalyzer database (APID, Prieto and De Las Rivas, 2006) that includes all known experimentally validated protein-protein interactions from BIND, BioGRID, DIP, HPRD, IntAct and MINT databases, accessed via the APID2NET (Hernandez-Toro et al., 2007) plugin. This process lead to the definitions of, respectively, the core interactome (*CI*, 303 proteins, 597 interactions, high resolution Image S1) and the extended interactome (*EI*, 3783 proteins, 24457 interactions, high resolution Image S2). Discussion on caveats and choices of original sources can be found in Tieri and Nardini (2013).

TOPOLOGICAL ANALYSIS

Topological analysis was run separately on the main *connected component* of each interactome (i.e., excluding the proteins for which no PPI was retrieved, i.e., that remained isolated) to evaluate a number of network parameters (Assenov et al., 2008): *degree*, or *connectivity,* i.e., the number of nodes linked to the node of interest (number of edges); and *betweenness centrality* (BC), a measure of the amount of control that a node exerts over the interactions of other nodes in the network. This measure favors nodes that join communities such as dense subnetworks, rather than nodes that lie inside a community, and has been shown to characterize essential proteins (Platzer et al., 2007). All calculated network parameters and rankings are listed in Data Sheet 2, Tables S15, S16 or can be recalculated from the Cytoscape CI_EI.cys (Data Sheet 3) file available at http://www.picb.ac.cn/ClinicalGenomicNTW/RAmultiomic.html.

PHARMACOLOGICAL TREATMENT SIMULATION

To simulate the pharmacological treatment, a virtual node knockout experiment has been performed by controlling (manual removal of the nodes and Cytoscape plugin *Interference* (Scardoni et al., 2014) 20 MTX controlled targets identified in literature (Cutolo et al., 2001; Chan and Cronstein, 2002) present in EI (Data Sheet 2, Table S17). Betweenness centrality (and, to add robustness to the analysis, stress, S, i.e., an alternative centrality functional form) were then re-calculated to assess the impact of such therapy on the topology and hence the functionality of the network. Manual node removal and pharmacological simulation plugin present overlapping results (*betweenness*: 95.9%, *stress*: 98.2%, Data Sheet 2, Table S17). The *p*-values, corrected for multiple testing (threshold 0.05), have been calculated after constructing null betweenness centrality distributions by 1000 random deletions of 20 nodes, as many as the MTX targets (Efron and Tibshirani, 1993). Functional clustering analysis has been then performed (Data Sheet 2, Table S18).

COMPARATIVE ANALYSIS

We further run a comparative analysis between our newly constructed multi-*omic* map, EI, and TR, that represent an earlier transcriptional-only version (Wu et al., 2010), to highlight the biological mechanisms that have been better emphasized from the usage of multilayer *omic* data.

Degree was evaluated as the number of edges attached to a node for the undirected networks as EI (and CI) are (i.e., connections among nodes do not indicate *directional* cause-effect nor temporal relationship). For TR (directed network) proteins and their modified instances (such as MAPKs and phosphorylated-MAPKs) were first considered as one (complex) node, then in-degrees (edges *to* the node) and out-degrees (edges *from* the node) of the components (MAPK and phosphorylated-MAPK) were summed up to obtain the undirected degree, after subtracting the number of edges connecting the members of the complex node. To complete the compatibility of the degree defined for undirected maps (and namely EI), given the different sizes of EI and TR, the percentrank of the degree was also computed. The nodes which degree rank was modified by more than 10% between the two networks, were considered as nodes undergoing a *transition*. A node was defined as *accomplished* when its % rank degree was preserved, *loser* when the ranking reduced from TR to EI, *climber* when it increased from TR to EI (Data Sheet 2, Table S19). From a strictly topological point of view, the threshold that defines a node as *accomplished* can be set to zero, and hence this definition identifies only the nodes with the same exact degree. From a biological standpoint, and for an informative biological interpretation of the results, it is not necessary to impose the exact matching of the ranking. For this reason we relaxed the threshold and defined as accomplished the nodes that present the same, higher or lower % rank of the degree with ±10% tolerance, as a reasonable compromise.

Biological meaning for *climbers* and *accomplished* nodes in the transition TR to EI was assessed by enrichment analysis Enrichr (Chen et al., 2013) see Data Sheet 2, Table S20.

RESULTS AND DISCUSSION

After curating all molecular information (**Table 1**) we inferred the network from the reconstructed lists with the PPI approach, which consists of connecting nodes (molecules) based on their interactions at the protein level, a broadly assessed approach in computational biology, and already used for RA in both already cited (Okada et al., 2014; You et al., 2014). All following results pertain to the analysis on the extended interactome (EI), more informative for its larger size.

To validate the ability of our network to model the biomolecular aspects of RA, we first simulated a therapeutic approach with MTX (see methods) and compared the results with the major known effects reported in literature (**Figure 1A**). As a result of the control on 20 MTX targets removal, the network changes its topology (**Figure 1B**; Data Sheet 2, Table S17), and the functional analysis indicates that 32 molecules which BC significantly altered (Data Sheet 2, Table S17, col. 2) pertain to two main functions [Data Sheet 2, Table S18, DAVID (Huang

Table 2 | RA-associated proteins significantly modified upon MTX therapy release and functional annotation clustering in DAVID.

	GO:0042981: reg. of apoptosis GO:0043067: reg. of progr. cell death GO:0010941: reg. of cell death		**GO:0031328: positive reg. cellular biosynth. process GO:0009891: positive reg. biosynth. process**		**GO:0051173: positive reg. nitrogen compound metabolic process**		**GO:0010557: positive regulation macromolec. biosynth. process**	
	BC	***S***	***BC***	***S***	***BC***	***S***	***BC***	***S***
ABL1	↑	↑						
BRCA1	↑	↓	↑	↓	↑	↓	↑	↓
CREBBP			↑	↑	↑	↑	↑	↑
CTNNB1			↑	↓	↑	↓	↑	↓
EGFR	↑	↓	↑	↓	↑	↓		
EP300			↑	↑	↑	↑	↑	↑
ESR1	↑	↓						
HSP90AA1, 2			↑	↓	↑	↓		
LCK	↑	↓						
MAPK1	↑	↑	↑	↑	↑	↑	↑	↑
MYC	↑	↑	↑	↑	↑	↑	↑	↑
PRKCA	↑	↑						
SMAD3	↑	↑	↑	↑	↑	↑	↑	↑
SRC	↑	↑						
STAT3			↑	↓	↑	↓	↑	↓
TP53	↑	↑	↑	↑	↑	↑	↑	↑
TRAF6	↑	↑	↑	↑			↑	↑
VHL	↑	↓	↑	↓	↑	↓	↑	↓
YWHAZ	↑	↓						

Thirty-two proteins were identified to be significantly changed by the 20 MTX target proteins' deletion (1000 permutations, adjusted p-value = 0.01). The topological measures of betweenness and stress centrality were shown to be significantly altered increased (black arrow ↑) or decreased (red arrow ↓) after knocking out the MTX target proteins. Among the listed proteins, enriched for the shown GO categories, STAT3 was found to belong to the host-microbiome interface as defined in Methods. The top 2 functional annotation clusters run on the changed proteins identified enrichment for cell death and biosynthetic process as well as nitrogen compound metabolic process (Functional Annotation Clustering Classification stringency: high, see Supplementary Data Sheet 2, Table S18; BC, betweenness centrality; S, stress centrality).

et al., 2009a)]: regulation of programmed cell death, a known effect of MTX (Spurlock et al., 2011); and metabolic and biosynthetic processes, an alteration known to constitute a side effect of the treatment (Phillips et al., 2003), as well as an area of synergy between host and microbiome (Tremaroli and Backhed, 2012; Devaraj et al., 2013; Winter et al., 2013). Moving down to the gene level, as illustrated in **Table 2**, Signal Transducers and Activators of Transcription 3 (STAT3) deserves particular attention, as it is a crucial player in the JAK/STAT signaling cascade, at the basis of the signal transduction mechanism for many cytokine receptors, highly activated in RA (Paunovic et al., 2008), and an important member of the host-microbiome interface (Zhou and Amar, 2006), being involved in the host susceptibility/defense against intestinal infections at the mucosal level (Miettinen et al., 2000).

From a topological point of view, STAT3 presents enhanced *betweenness* and reduced *stress centralities* after virtual MTX treatment. This is an unusual topological condition—since there is commonly correlation between stress and betweenness—where, upon perturbation (MTX) a higher fraction of shortest paths converges on STAT3 (gain in *betweenness centrality*) despite a decrease in their absolute number (loss of *stress centrality*). This indicates that the networks shrinks and STAT3 becomes more important, a fact that can be translated in biological terms as the compensatory mechanisms induced by the loss of some molecules' presence/activity (MTX targets), which globally force STAT3 to become the molecule through which more numerous (higher *betweenness*) but less efficient molecular reactions (longer paths, lower *stress*) occur.

Overall, STAT3, which is already considered a crucial target in RA for its critical role in the T regulatory/helper 17 lymphoid cells [T_{reg}/Th_{17} balance overabundant in RA (Leipe et al., 2010)] is coherently shown as an indirectly controlled target by MTX explaining the ability of the therapy to rebalance Th17/IL17 ratio (Li et al., 2012).

In conclusion, our map is able to recollect known and yet complex information about the effects of MTX, this represents an important validation of our frame for further simulations. Additionally, our map indicates a clear link between MTX and dysbiosis, which to date has not been explicitly unrevealed, although enterocolitis is a known toxic effect of MTX, linked to the induced nitroxidative stress (Kolli et al., 2008, 2013). This is a critical fact as the known adverse effects of MTX, generally described as immunodepressive, appear to be composed not only by the known oxidative organ stress, but also by an added dysbiosis, possibly mediated by an overload on STAT3.

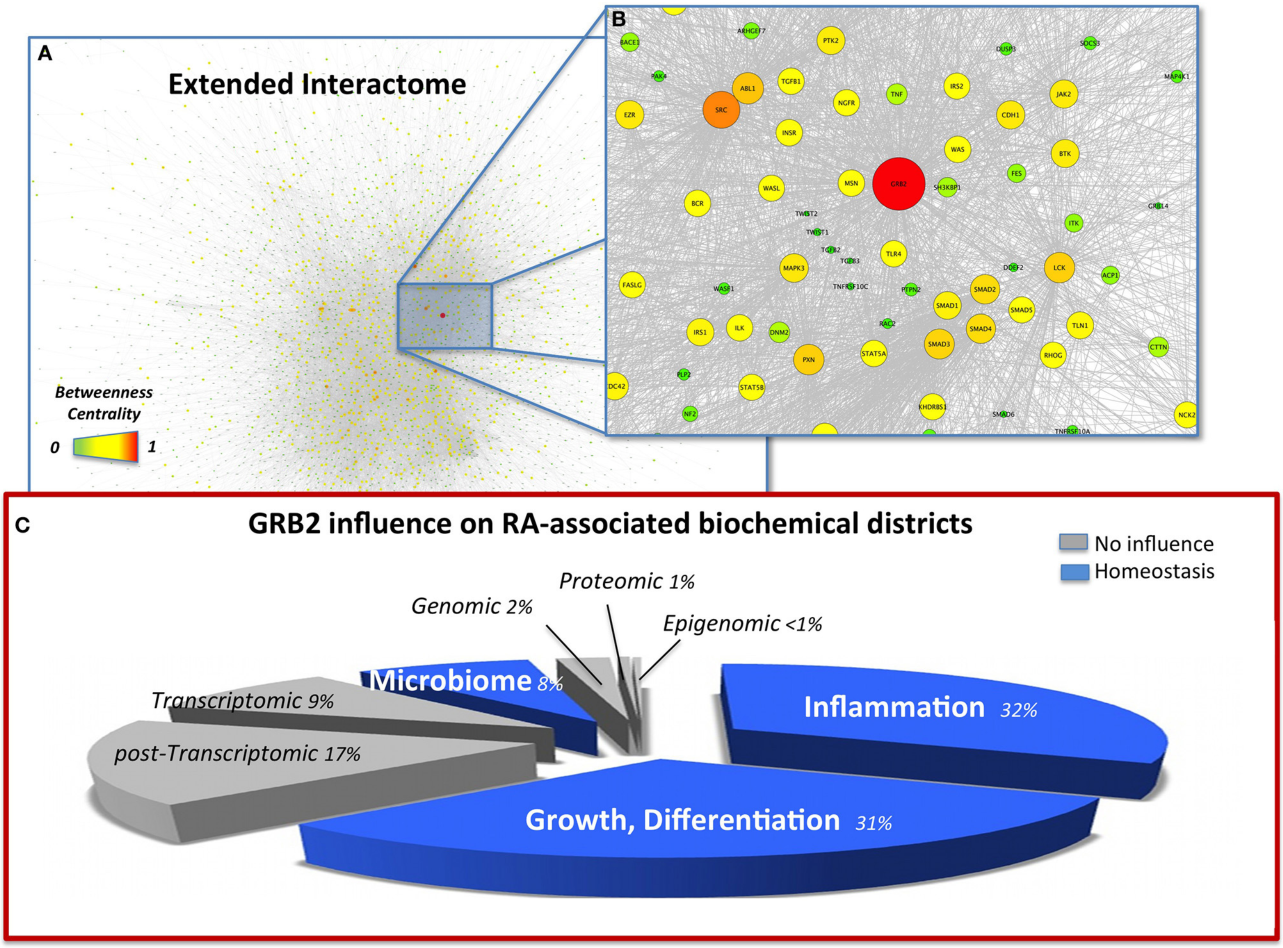


FIGURE 1 | (A) Snapshot of the extended interactome (EI) with nodes highlighted by betweenness centrality (BC), high resolution browsable figure provided in Supplementary Files (Image S2). **(B)** Zoom on the top ranking BC node (GRB2) and its closer interactome. Pathways relevant in the indication of GRB2 as an RA target, able to control inflammation TGF-β (TGFB1-3), TNF-α (TNF, TNFRS10C), MAPK (MAP4K1, MAPK3), degeneracy EMT (TWIST1-2, CDH1), and dysbiosis (TRL4) are also highlighted. **(C)** Visual summary of the influence of GRB2 on the RA-affected districts highlight a homeostatic (blue) influence on inflammation, GI microbiome, growth, differentiation. The pie-chart slices' size is proportional to the number of molecules considered in each district. Districts were merged from the total 13 datasets according to biochemical homogeneity in the following 8 categories: Genomic (DNA, Dataset 1); Epigenomic (mDNA, Dataset 4); Transcriptomic (mRNA, Datasets 7, 9A, 9B); Post-transcriptomic (miRNA, Dataset 8); Proteomic (proteins, Dataset 2); Microbiome (Host-microbiome proteins interface, Oral microbiome Datasets 5, 10, 12, 13); Inflammation (6, 3A, 3B, 3C); Others, i.e., Growth, Differentiation (Datasets 3, 11, 3D, 3E).

The topological analysis highlights the striking relevance of Growth factor receptor-bound protein 2 (GRB2) with values of *BC* more than two-fold (Data Sheet 2, Table S16) compared to the second in rank, the Epidermal growth factor receptor (EGFR). Based on literature, GRB2 is an effective target (Phase I clinical trial, http://www.biopathholdings.com/) for Acute Myeloid Leukemia (AML), Chronic myelogenous leukemia (CML) and Myelodysplastic syndromes (MDS); an important mediator of the oncogenic activities of TGF-β, via epithelial mesenchymal transition (EMT) (Galliher-Beckley and Schiemann, 2008); a crucial player in the host-microbiome interaction of *Helicobacter pylori*, able to induce host cell scattering and proliferation via the activation of the Ras/MEK/ERK pathway (Mimuro et al., 2002); a marker of RA in synoviocytes (Huh et al., 2003). GRB2 is additionally activated by leptin (Pai et al., 2005), abundant in RA (Bokarewa et al., 2003) and able to increase *Prevotella intermedia* LPS-induced TNF-α production (Kim, 2010). Moreover, another member of the *Prevotella* genus (*P. copri*) has recently been liaised to RA (Scher et al., 2013), as a specific marker of GI microbiome dysbiosis associated to the disease. When observed from the network perspective this apparently scattered information fits in a connected map (**Figure 1B**) and hence builds a robust rationale for considering GRB2 as a target for RA. The activation of proliferative and inflammatory pathways as well as EMT, are hallmarks of RA (You et al., 2014) suggesting that the control on GRB2 as a regulator of such mechanisms is appropriate. Additionally, the control on GRB2 exerted by *H. pylori* [already proposed in relation to RA (Melby et al., 1999)] and by *P. intermedia* in the presence of leptin indicate that targeting of GRB2 is not only of relevance to control the phenotypic symptoms of

RA (joints degeneracy) but also the recently highlighted dysbiosis that accompany the disease, via the control of the disruptive mechanisms by which pathogens can exert their action on the host (**Figure 1C**).

Given the relevance of RA as a paradigmatic autoimmune disease, a variety of *in silico* modeling approaches have been devised (Okada et al., 2014; You et al., 2014), and, among those, an early transcriptional only map (hereinafter TR, 302 nodes; Wu et al., 2010). The previous compilation of this simplified version put us in the relatively unique position to be able to quantify the benefit, in terms of information content, of expanding from transcriptional to multi-omic the network modeling of RA. The molecules that gain importance (i.e., have a higher degree) in the multi-*omic* map versus the TR (*climbers*, see Methods and **Figure 2A**) pertain mostly to the MAPK Signaling Pathway (**Figure 2B** and Data Sheet 2, Table S19). This category is also highly enriched for *accomplished* nodes, thus validating the importance of this pathway in the disease. However, *climbers*, all representing genes shared between TR and EI, include molecules known to belong also to the GI interface (SFR, MAP2K4, MAP3K8), absent in the *accomplished*, implying the importance of the involvement of the host-microbiome interface, not taken into account in the TR map. In particular, Interleukin-1 Receptor Associated Kinase-4 (IRAK4, *climber*) is known to play a critical role in initiating response to foreign pathogens (Hofman and Vouret-Craviari, 2012) and was recently presented to the American College of Rheumatology (ACR), based on promising results on the control of B-cell-like diffuse large B-cell lymphoma (DLBCL), as a potential treatment for RA (Chaudahry and Al, 2012). In the network perspective, this choice calls for words of cautions. Indeed, while correlating with regression of some aspects of the disease, the control on IRAK4 affects the response to pathogens, and in particular IRAK4 inhibitors impacts on pDCs in RA patients (Chiang et al., 2011), therefore limiting the appropriate and

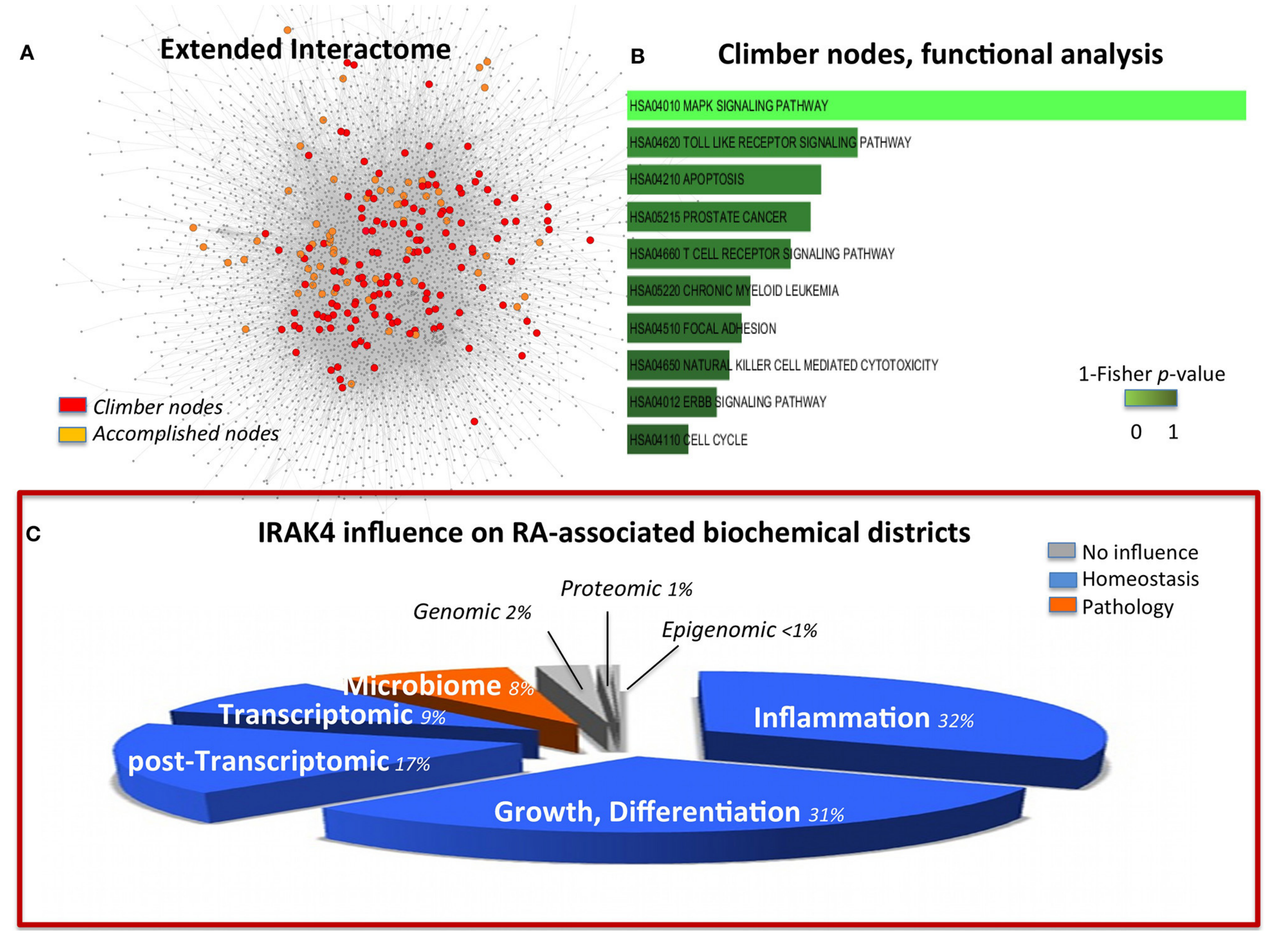


FIGURE 2 | (A) Multi-*omic* map (EI) nodes highlighted according to their role in comparison with a transcriptional-only map (TR). In orange, nodes that maintain their role and importance in both EI and TR (*accomplished*); in red, nodes that gain importance in the multi-*omic* context, (*climbers*). **(B)** Functional analysis of the climber hubs, which highlight the striking significance of MAPK signals. Panel **(C)** is built in the same way of **Figure 1C** to permit easy comparison of the two targets. It represents the summary of the influence of IRAK4 on the RA-affected districts, and highlights a homeostatic (blue) influence on inflammation, growth, differentiation as well as transcriptomic and post-transcriptomic districts. However, the microbiome interface response is impaired by IRAK4 inhibition of the innate immune response to pathogens. The pie-chart slices' size is proportional to the number of molecules considered in each district (as in **Figure 1**).

immediate innate host response in case of bacterial infections (**Figure 2C**).

CONCLUSION

The aim of the designed framework is to draw hypotheses that can support basic research and further clinical practice. In particular, we here highlight two major areas of application: support in the identification of novel drug targets (exemplified by GRB2); support in the identification of potential contraindication to novel therapies, i.e., support in the design of robust clinical trials (exemplified by IRAK4-inhibitors). While the former application joins other efforts in different clinical areas [such as on diabetes (Liu et al., 2007; Santiago and Potashkin, 2013), in cancer (Hwang et al., 2013), and on glioblastoma (Junhua et al., 2012)], the latter descends from the inclusion of numerous data types, including for the first time to our knowledge, the GI microbiome interface. The results discussed in this article are the output of the knowledge distilled from ∼4000 selected molecules and ∼15 public databases, a humongous amount of information carefully and often redundantly peer-reviewed by the scientific community. Future and ongoing research and the resulting discoveries will impact on the breadth and possibly on the topology of our map. To take into account these expected (and desirable) events, our map was drawn using open source programs and pathway molecules' standards to allow full map usability, editing and updating by the whole scientific community.

AUTHOR CONTRIBUTION

Paolo Tieri built and analyzed the map; XiaoYuan Zhou performed pharmacological simulation; XiaoYuan Zhou and Lisha Zhu run functional and comparative analyses; Christine Nardini analyzed the connection to the GI microbiome; Paolo Tieri and Christine Nardini designed the study, analyzed the results and wrote the manuscript; XiaoYuan Zhou and Lisha Zhu contributed to write and revise the manuscript.

DATA SHARING STATEMENT

All data are available publicly, our map is publicly available here: http://www.picb.ac.cn/ClinicalGenomicNTW/RAmultiomic.html

ACKNOWLEDGMENTS

This project is funded by MoST international cooperation program n. 2013DFA30790, and NSFC n. 31070748, PT collaborated in the frame of the European Commission FP7-PEOPLE-2011-IRSES program 31028760, project ID 294935 "KEPAMOD," and as CAS fellow grant n. 2011Y1SA04.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fcell.2014.00059/abstract

REFERENCES

Antonov, A. V., Dietmann, S., Rodchenkov, I., and Mewes, H. W. (2009). PPI spider: a tool for the interpretation of proteomics data in the context of protein-protein interaction networks. *Proteomics* 9, 2740–2749. doi: 10.1002/pmic.200800612

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

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2011). NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.* 39, D1005–D1010. doi: 10.1093/nar/gkq1184

Berger, S. I., Posner, J. M., and Ma'ayan, A. (2007). Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics* 8:372. doi: 10.1186/1471-2105-8-372

Bingham, C. O. 3rd., and Moni, M. (2013). Periodontal disease and rheumatoid arthritis: the evidence accumulates for complex pathobiologic interactions. *Curr. Opin. Rheumatol.* 25, 345–353. doi: 10.1097/BOR.0b013e32835fb8ec

Bokarewa, M., Bokarew, D., Hultgren, O., and Tarkowski, A. (2003). Leptin consumption in the inflamed joints of patients with rheumatoid arthritis. *Ann. Rheum. Dis.* 62, 952–956. doi: 10.1136/ard.62.10.952

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, D1228–D1233. doi: 10.1093/nar/gks1147

Chan, E. S., and Cronstein, B. N. (2002). Molecular action of methotrexate in inflammatory diseases. *Arthritis Res.* 4, 266–273. doi: 10.1186/ar419

Chaudahry, D., and Al, E. (2012). "Identification of highly potent and selective Interleukin-1 receptor-associated kinase-4 inhibitor for the treatmetn of rheumatic diseases," in *American College of Rheumatology (ACR) Annual Scientific Meeting* (Washington, DC).

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

Chiang, E. Y., Yu, X., and Grogan, J. L. (2011). Immune complex-mediated cell activation from systemic lupus erythematosus and rheumatoid arthritis patients elaborate different requirements for IRAK1/4 kinase activity across human cell types. *J. Immunol.* 186, 1279–1288. doi: 10.4049/jimmunol.1002821

Consortium, U. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148. doi: 10.1093/nar/gkp846

Cutolo, M., Sulli, A., Pizzorni, C., Seriolo, B., and Straub, R. H. (2001). Anti-inflammatory mechanisms of methotrexate in rheumatoid arthritis. *Ann. Rheum. Dis.* 60, 729–735. doi: 10.1136/ard.60.8.729

Devaraj, S., Hemarajata, P., and Versalovic, J. (2013). The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clin. Chem.* 59, 617–628. doi: 10.1373/clinchem.2012.187617

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24, i223–i231. doi: 10.1093/bioinformatics/btn161

Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J. Biomed. Inform.* 44, 839–847. doi: 10.1016/j.jbi.2011.05.002

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall. doi: 10.1007/978-1-4899-4541-9

Farquharson, D., Butcher, J. P., and Culshaw, S. (2012). Periodontitis, Porphyromonas, and the pathogenesis of rheumatoid arthritis. *Mucosal Immunol.* 5, 112–120. doi: 10.1038/mi.2011.66

Feldmann, M., Andreakos, E., Smith, C., Bondeson, J., Yoshimura, S., Kiriakidis, S., et al. (2002). Is NF-kappaB a useful therapeutic target in rheumatoid arthritis? *Ann. Rheum. Dis.* 61(Suppl. 2), ii13–ii18. doi: 10.1136/ard.61.suppl_2.ii13

Galliher-Beckley, A. J., and Schiemann, W. P. (2008). Grb2 binding to Tyr284 in TbetaR-II is essential for mammary tumor growth and metastasis stimulated by TGF-beta. *Carcinogenesis* 29, 244–251. doi: 10.1093/carcin/bgm245

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Hayden, M. S., and Ghosh, S. (2012). NF-kappaB, the first quarter-century: remarkable progress and outstanding questions. *Genes Dev.* 26, 203–234. doi: 10.1101/gad.183434.111

Hernandez-Toro, J., Prieto, C., and De Las Rivas, J. (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics* 23, 2495–2497. doi: 10.1093/bioinformatics/btm373

Hodgman, C. (2007). Integrative biology–the way forward. *Brief. Bioinform.* 8, 208–209. doi: 10.1093/bib/bbm036

Hofman, P., and Vouret-Craviari, V. (2012). Microbes-induced EMT at the crossroad of inflammation and cancer. *Gut Microbes* 3, 176–185. doi: 10.4161/gmic.20288

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

Huang, W., Wang, P., Liu, Z., and Zhang, L. (2009b). Identifying disease associations via genome-wide association studies. *BMC Bioinformatics* 10(Suppl. 1):S68. doi: 10.1186/1471-2105-10-S1-S68

Huh, S. J., Paik, D. J., Chung, H. S., and Youn, J. (2003). Regulation of GRB2 and FLICE2 expression by TNF-alpha in rheumatoid synovium. *Immunol. Lett.* 90, 93–96. doi: 10.1016/j.imlet.2003.07.002

Hutchings, C. J., Koglin, M., and Marshall, F. H. (2010). Therapeutic antibodies directed at G protein-coupled receptors. *MAbs* 2, 594–606. doi: 10.4161/mabs.2.6.13420

Hwang, T. H., Atluri, G., Kuang, R., Kumar, V., Starr, T., Silverstein, K. A., et al. (2013). Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers. *BMC Genomics* 14:440. doi: 10.1186/1471-2164-14-440

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.* 31, e15. doi: 10.1093/nar/gng015

Iskar, M., Zeller, G., Zhao, X. M., Van Noort, V., and Bork, P. (2012). Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr. Opin. Biotechnol.* 23, 609–616. doi: 10.1016/j.copbio.2011.11.010

Jin, G., Zhou, X., Wang, H., Zhao, H., Cui, K., Zhang, X. S., et al. (2008). The knowledge-integrated network biomarkers discovery for major adverse cardiac events. *J. Proteome Res.* 7, 4013–4021. doi: 10.1021/pr8002886

Junhua, Z., Shihua, Z., Yong, W., Junfei, Z., and Xiang-Sun, Z. (2012). "Identifying mutated core modules in glioblastoma by integrative network analysis," in *Systems Biology (ISB), 2012 IEEE 6th International Conference* (Xi'an), 304–309.

Karouzakis, E., Gay, R. E., Gay, S., and Neidhart, M. (2011). Epigenetic deregulation in rheumatoid arthritis. *Adv. Exp. Med. Biol.* 711, 137–149. doi: 10.1007/978-1-4419-8216-2_10

Kim, S. J. (2010). Leptin potentiates prevotella intermedia lipopolysaccharide-induced production of TNF-alpha in monocyte-derived macrophages. *J. Periodontal Implant Sci.* 40, 119–124. doi: 10.5051/jpis.2010.40.3.119

Kim, T. Y., Kim, H. U., and Lee, S. Y. (2010). Data integration and analysis of biological networks. *Curr. Opin. Biotechnol.* 21, 78–84. doi: 10.1016/j.copbio.2010.01.003

Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L., Fergerson, R., Hewett, M., et al. (2001). Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge base. *Pharmacogenomics J.* 1, 167–170. doi: 10.1038/sj.tpj.6500035

Kolli, V. K., Abraham, P., and Rabi, S. (2008). Methotrexate-induced nitrosative stress may play a critical role in small intestinal damage in the rat. *Arch. Toxicol.* 82, 763–770. doi: 10.1007/s00204-008-0287-9

Kolli, V. K., Kanakasabapathy, I., Faith, M., Ramamoorthy, H., Isaac, B., Natarajan, K., et al. (2013). A preclinical study on the protective effect of melatonin against methotrexate-induced small intestinal damage: effect mediated by attenuation of nitrosative stress, protein tyrosine nitration, and PARP activation. *Cancer Chemother. Pharmacol.* 71, 1209–1218. doi: 10.1007/s00280-013-2115-z

Kolly, L., Busso, N., Palmer, G., Talabot-Ayer, D., Chobaz, V., and So, A. (2010). Expression and function of the NALP3 inflammasome in rheumatoid synovium. *Immunology* 129, 178–185. doi: 10.1111/j.1365-2567.2009.03174.x

Lee, Y., Li, H., Li, J., Rebman, E., Achour, I., Regan, K. E., et al. (2013). Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *J. Am. Med. Inform. Assoc.* 20, 619–629. doi: 10.1136/amiajnl-2012-001519

Lehner, B., and Lee, I. (2008). Network-guided genetic screening: building, testing and using gene networks to predict gene function. *Brief. Funct. Genomic Proteomic* 7, 217–227. doi: 10.1093/bfgp/eln020

Leipe, J., Grunke, M., Dechant, C., Reindl, C., Kerzendorf, U., Schulze-Koops, H., et al. (2010). Role of Th17 cells in human autoimmune arthritis. *Arthritis Rheum.* 62, 2876–2885. doi: 10.1002/art.27622

Li, Y., Jiang, L., Zhang, S., Yin, L., Ma, L., He, D., et al. (2012). Methotrexate attenuates the Th17/IL-17 levels in peripheral blood mononuclear cells from healthy individuals and RA patients. *Rheumatol. Int.* 32, 2415–2422. doi: 10.1007/s00296-011-1867-1

Liu, M., Liberzon, A., Kong, S. W., Lai, W. R., Park, P. J., Kohane, I. S., et al. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* 3:e96. doi: 10.1371/journal.pgen.0030096

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489, 220–230. doi: 10.1038/nature11550

Lynn, D. J., Winsor, G. L., Chan, C., Richard, N., Laird, M. R., Barsky, A., et al. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* 4, 218. doi: 10.1038/msb.2008.55

Makarov, S. S. (2001). NF-kappa B in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Res.* 3, 200–206. doi: 10.1186/ar300

Mathews, R. J., Robinson, J. I., Battellino, M., Wong, C., Taylor, J. C., Eyre, S., et al. (2013). Evidence of NLRP3-inflammasome activation in rheumatoid arthritis (RA); genetic variants within the NLRP3-inflammasome complex in relation to susceptibility to RA and response to anti-TNF treatment. *Ann. Rheum. Dis.* 73, 1202–1210. doi: 10.1136/annrheumdis-2013-203276

Maynard, C. L., Elson, C. O., Hatton, R. D., and Weaver, C. T. (2012). Reciprocal interactions of the intestinal microbiota and immune system. *Nature* 489, 231–241. doi: 10.1038/nature11551

Mcinnes, I. B., and Schett, G. (2011). The pathogenesis of rheumatoid arthritis. *N. Engl. J. Med.* 365, 2205–2219. doi: 10.1056/NEJMra1004965

Melby, K. K., Kvien, T. K., and Glennas, A. (1999). *Helicobacter pylori*–a trigger of reactive arthritis? *Infection* 27, 252–255. doi: 10.1007/s150100050022

Miagkov, A. V., Kovalenko, D. V., Brown, C. E., Didsbury, J. R., Cogswell, J. P., Stimpson, S. A., et al. (1998). NF-kappaB activation provides the potential link between inflammation and hyperplasia in the arthritic joint. *Proc. Natl. Acad. Sci. U.S.A.* 95, 13859–13864. doi: 10.1073/pnas.95.23.13859

Miettinen, M., Lehtonen, A., Julkunen, I., and Matikainen, S. (2000). Lactobacilli and streptococci activate NF-kappa B and STAT signaling pathways in human macrophages. *J. Immunol.* 164, 3733–3740. doi: 10.4049/jimmunol.164.7.3733

Mikuls, T. R., Thiele, G. M., Deane, K. D., Payne, J. B., O'dell, J. R., Yu, F., et al. (2012). *Porphyromonas gingivalis* and disease-related autoantibodies in individuals at increased risk of rheumatoid arthritis. *Arthritis Rheum.* 64, 3522–3530. doi: 10.1002/art.34595

Mimuro, H., Suzuki, T., Tanaka, J., Asahi, M., Haas, R., and Sasakawa, C. (2002). Grb2 is a key mediator of *Helicobacter pylori* CagA protein activities. *Mol. Cell* 10, 745–755. doi: 10.1016/S1097-2765(02)00681-0

Moulos, P., Valavanis, I., Klein, J., Maglogiannis, I., Schanstra, J., and Chatziioannou, A. (2011). Unifying the integration, analysis and interpretation of multi-omic datasets: exploration of the disease networks of obstructive nephropathy in children. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2011, 3716–3719. doi: 10.1109/IEMBS.2011.6090631

Murie, C., Woody, O., Lee, A. Y., and Nadon, R. (2009). Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics* 10:45. doi: 10.1186/1471-2105-10-45

Nakano, K., Whitaker, J. W., Boyle, D. L., Wang, W., and Firestein, G. S. (2012). DNA methylome signature in rheumatoid arthritis. *Ann. Rheum. Dis.* 72, 110–117. doi: 10.1136/annrheumdis-2012-201526

Oeckinghaus, A., Hayden, M. S., and Ghosh, S. (2011). Crosstalk in NF-κB signaling pathways. *Nat. Immunol.* 12, 695–708. doi: 10.1038/ni.2065

Ogrendik, M. (2013). Rheumatoid arthritis is an autoimmune disease caused by periodontal pathogens. *Int. J. Gen. Med.* 6, 383–386. doi: 10.2147/IJGM.S45929

Okada, M., Kobayashi, T., Ito, S., Yokoyama, T., Abe, A., Murasawa, A., et al. (2013). Periodontal treatment decreases levels of antibodies to *Porphyromonas gingivalis* and citrulline in patients with rheumatoid arthritis and periodontitis. *J. Periodontol.* 84, 74–84. doi: 10.1902/jop.2013.130079

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381. doi: 10.1038/nature12873

Okamoto, T. (2006). NF-kappaB and rheumatic diseases. *Endocr. Metab. Immune Disord. Drug Targets* 6, 359–372. doi: 10.2174/187153006779025685

Pai, R., Lin, C., Tran, T., and Tarnawski, A. (2005). Leptin activates STAT and ERK2 pathways and induces gastric cancer cell proliferation. *Biochem. Biophys. Res. Commun.* 331, 984–992. doi: 10.1016/j.bbrc.2005.03.236

Paunovic, V., Carroll, H. P., Vandenbroeck, K., and Gadina, M. (2008). Signalling, inflammation and arthritis: crossed signals: the role of interleukin (IL)-12,

-17, -23 and -27 in autoimmunity. *Rheumatology (Oxford)* 47, 771–776. doi: 10.1093/rheumatology/kem352

Phillips, D. C., Woollard, K. J., and Griffiths, H. R. (2003). The anti-inflammatory actions of methotrexate are critically dependent upon the production of reactive oxygen species. *Br. J. Pharmacol.* 138, 501–511. doi: 10.1038/sj.bjp.0705054

Platzer, A., Perco, P., Lukas, A., and Mayer, B. (2007). Characterization of protein-interaction networks in tumors. *BMC Bioinformatics* 8:224. doi: 10.1186/1471-2105-8-224

Prieto, C., and De Las Rivas, J. (2006). APID: agile protein interaction dataanalyzer. *Nucleic Acids Res.* 34, W298–W302. doi: 10.1093/nar/gkl128

Roman-Blas, J. A., and Jimenez, S. A. (2006). NF-kappaB as a potential therapeutic target in osteoarthritis and rheumatoid arthritis. *Osteoarthritis Cartilage* 14, 839–848. doi: 10.1016/j.joca.2006.04.008

Roman-Blas, J. A., and Jimenez, S. A. (2008). Targeting NF-kappaB: a promising molecular therapy in inflammatory arthritis. *Int. Rev. Immunol.* 27, 351–374. doi: 10.1080/08830180802295740

Santiago, J. A., and Potashkin, J. A. (2013). Integrative network analysis unveils convergent molecular pathways in parkinson's disease and diabetes. *PLoS ONE* 8:e83940. doi: 10.1371/journal.pone.0083940

Scardoni, G., Montresor, A., Tosadori, G., and Laudanna, C. (2014). Node interference and robustness: performing virtual knock-out experiments on biological networks: the case of leukocyte integrin activation network. *PLoS ONE* 9:e88938. doi: 10.1371/journal.pone.0088938

Scher, J. U., and Abramson, S. B. (2011). The microbiome and rheumatoid arthritis. *Nat. Rev. Rheumatol.* 7, 569–578. doi: 10.1038/nrrheum.2011.121

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., et al. (2013). Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *Elife* 2:e01202. doi: 10.7554/eLife.01202

Scher, J. U., Ubeda, C., Equinda, M., Khanin, R., Buischi, Y., Viale, A., et al. (2012). Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis. *Arthritis Rheum.* 64, 3083–3094. doi: 10.1002/art.34539

Sharman, J. L., Benson, H. E., Pawson, A. J., Lukito, V., Mpamhanga, C. P., Bombail, V., et al. (2013). IUPHAR-DB: updated database content and new features. *Nucleic Acids Res.* 41, D1083–D1088. doi: 10.1093/nar/gks960

Sidiropoulos, P. I., Goulielmos, G., Voloudakis, G. K., Petraki, E., and Boumpas, D. T. (2008). Inflammasomes and rheumatic diseases: evolving concepts. *Ann. Rheum. Dis.* 67, 1382–1389. doi: 10.1136/ard.2007.078014

Simmonds, R. E., and Foxwell, B. M. (2008). Signalling, inflammation and arthritis: NF-kappaB and its relevance to arthritis and inflammation. *Rheumatology (Oxford)* 47, 584–590. doi: 10.1093/rheumatology/kem298

Smale, S. T. (2011). Hierarchies of NF-κB target-gene regulation. *Nat. Immunol.* 12, 689–694. doi: 10.1038/ni.2070

Smit, M. D., Westra, J., Vissink, A., Doornbos-Van Der Meer, B., Brouwer, E., and Van Winkelhoff, A. J. (2012). Periodontitis in established rheumatoid arthritis patients: a cross-sectional clinical, microbiological and serological study. *Arthritis Res. Ther.* 14, R222. doi: 10.1186/ar4061

Spurlock, C. F. 3rd., Aune, Z. T., Tossberg, J. T., Collins, P. L., Aune, J. P., Huston, J. W. 3rd., et al. (2011). Increased sensitivity to apoptosis induced by methotrexate is mediated by JNK. *Arthritis Rheum.* 63, 2606–2616. doi: 10.1002/art.30457

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., et al. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* 42, 508–514. doi: 10.1038/ng.582

Stamp, L. K., Hazlett, J., Roberts, R. L., Frampton, C., Highton, J., and Hessian, P. A. (2012). Adenosine receptor expression in rheumatoid synovium: a basis for methotrexate action. *Arthritis Res. Ther.* 14, R138. doi: 10.1186/ar3871

Tieri, P., and Nardini, C. (2013). Signalling pathway database usability: lessons learned. *Mol. Biosyst.* 9, 2401–2407. doi: 10.1039/c3mb70242a

Tieri, P., Termanini, A., Bellavista, E., Salvioli, S., Capri, M., and Franceschi, C. (2012). Charting the NF-κB pathway interactome map. *PLoS ONE* 7:e32678. doi: 10.1371/journal.pone.0032678

Tremaroli, V., and Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489, 242–249. doi: 10.1038/nature11552

Trenkmann, M., Brock, M., Ospelt, C., and Gay, S. (2010). Epigenetics in rheumatoid arthritis. *Clin. Rev Allergy Immunol.* 39, 10–19. doi: 10.1007/s12016-009-8166-6

Van Loo, G., and Beyaert, R. (2011). Negative regulation of NF-kappaB and its involvement in rheumatoid arthritis. *Arthritis Res. Ther.* 13, 221. doi: 10.1186/ar3324

Varani, K., Padovan, M., Govoni, M., Vincenzi, F., Trotta, F., and Borea, P. A. (2010). The role of adenosine receptors in rheumatoid arthritis. *Autoimmun. Rev.* 10, 61–64. doi: 10.1016/j.autrev.2010.07.019

Varani, K., Padovan, M., Vincenzi, F., Targa, M., Trotta, F., Govoni, M., et al. (2011). A2A and A3 adenosine receptor expression in rheumatoid arthritis: upregulation, inverse correlation with disease activity score and suppression of inflammatory cytokine and metalloproteinase release. *Arthritis Res. Ther.* 13, R197. doi: 10.1186/ar3527

Vincenzi, F., Padovan, M., Targa, M., Corciulo, C., Giacuzzo, S., Merighi, S., et al. (2013). A(2A) adenosine receptors are differentially modulated by pharmacological treatments in rheumatoid arthritis patients and their stimulation ameliorates adjuvant-induced arthritis in rats. *PLoS ONE* 8:e54195. doi: 10.1371/journal.pone.0054195

Winter, S. E., Lopez, C. A., and Baumler, A. J. (2013). The dynamics of gut-associated microbial communities during inflammation. *EMBO Rep.* 14, 319–327. doi: 10.1038/embor.2013.27

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., et al. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* 10, R130. doi: 10.1186/gb-2009-10-11-r130

Wu, C., Zhu, J., and Zhang, X. (2012). Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics* 13:182. doi: 10.1186/1471-2105-13-182

Wu, G., Zhu, L., Dent, J. E., and Nardini, C. (2010). A comprehensive molecular interaction map for rheumatoid arthritis. *PLoS ONE* 5:e10137. doi: 10.1371/journal.pone.0010137

You, S., Yoo, S. A., Choi, S., Kim, J. Y., Park, S. J., Ji, J. D., et al. (2014). Identification of key regulators for the migration and invasion of rheumatoid synoviocytes through a systems approach. *Proc. Natl. Acad. Sci. U.S.A.* 111, 550–555. doi: 10.1073/pnas.1311239111

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M. J. (2008). A navigator for human genome epidemiology. *Nat. Genet.* 40, 124–125. doi: 10.1038/ng0208-124

Zhou, Q., and Amar, S. (2006). Identification of proteins differentially expressed in human monocytes exposed to *Porphyromonas gingivalis* and its purified components by high-throughput immunoblotting. *Infect. Immun* 74, 1204–1214. doi: 10.1128/IAI.74.2.1204-1214.2006

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

Received: 01 July 2014; accepted: 26 September 2014; published online: 04 November 2014.

Citation: Tieri P, Zhou XY, Zhu L and Nardini C (2014) Multi-omic landscape of rheumatoid arthritis: re-evaluation of drug adverse effects. Front. Cell Dev. Biol. ***2****:59. doi: 10.3389/fcell.2014.00059*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 22 August 2014
doi: 10.3389/fcell.2014.00037

From molecular signatures to predictive biomarkers: modeling disease pathophysiology and drug mechanism of action

Andreas Heinzel[1], Paul Perco[1], Gert Mayer[2], Rainer Oberbauer[3], Arno Lukas[1] and Bernd Mayer[1]*

[1] emergentec biodevelopment GmbH, Vienna, Austria
[2] Department of Internal Medicine IV, Medical University of Innsbruck, Innsbruck, Austria
[3] Department of Internal Medicine III, KH Elisabethinen Linz and Medical University of Vienna, Vienna, Austria

Edited by:
Christine Nardini, Partner Institute for Computational Biology, China

Reviewed by:
Zahraa Naji Sabra, American University of Beirut, Lebanon
Satyaprakash Nayak, Pfizer Inc., USA

***Correspondence:**
Bernd Mayer, emergentec biodevelopment GmbH, Gersthoferstrasse 29-31, Vienna, 1180, Austria
e-mail: bernd.mayer@emergentec.com

Omics profiling significantly expanded the molecular landscape describing clinical phenotypes. Association analysis resulted in first diagnostic and prognostic biomarker signatures entering clinical utility. However, utilizing Omics for deepening our understanding of disease pathophysiology, and further including specific interference with drug mechanism of action on a molecular process level still sees limited added value in the clinical setting. We exemplify a computational workflow for expanding from statistics-based association analysis toward deriving molecular pathway and process models for characterizing phenotypes and drug mechanism of action. Interference analysis on the molecular model level allows identification of predictive biomarker candidates for testing drug response. We discuss this strategy on diabetic nephropathy (DN), a complex clinical phenotype triggered by diabetes and presenting with renal as well as cardiovascular endpoints. A molecular pathway map indicates involvement of multiple molecular mechanisms, and selected biomarker candidates reported as associated with disease progression are identified for specific molecular processes. Selective interference of drug mechanism of action and disease-associated processes is identified for drug classes in clinical use, in turn providing precision medicine hypotheses utilizing predictive biomarkers.

Keywords: omics, integration, molecular model, biomarker, target, systems biology, systems pharmacology, precision medicine

INTRODUCTION

Despite a continuously rising number of clinical trials the rate of bringing novel medication to the clinic is stalling (Pammolli et al., 2011). Here, Omics profiling and high throughput drug screening technologies at the interface of large scale clinical data have triggered novel conceptual strategies aimed at improved patient stratification for enabling precision medicine (Trusheim et al., 2011; Hollebecque et al., 2014). For implementing such approaches a number of issues need to be addressed including: (i) mirroring the clinical categorization of a phenotype on a molecular level description, (ii) spotting molecular factors mechanistically driving disease progression, (iii) drug-based intervention specifically addressing such progression mechanisms, and (iv) predictive biomarkers allowing fit-for-purpose analysis regarding a match of relevant pathophysiology and drug mechanism of action on the individual patient level (Heinzel et al., 2012).

A clinically well-established example is HER2 positive breast cancer characterized by overexpression of a member of the epidermal growth factor receptor family (ERBB2) playing a mechanistic role in progressive disease. In case the factor is proving positive for a patient the specific presentation is amenable for treatment tackling growth signaling (Hicks and Kulkarni, 2008). Still, the clinical presentation of breast cancer shows heterogeneous pathophysiologies apart HER2 positive subtypes. In consequence, when aiming at a comprehensive assessment of progressive breast cancer phenotypes multimarker panels are needed, e.g., implemented by a multiplexed assay holding 70 individual molecular features (Buyse et al., 2006). Such multimarker panels have generally become a promising strategy for characterizing complex clinical presentations, e.g., utilizing a serum marker panel for predicting coronary artery disease in symptomatic patients, or a urinary proteomics profile for early diagnosis of diabetic kidney disease (LaFramboise et al., 2012; Zürbig et al., 2012).

Failure for identifying a single causative factor as proxy for determining progression of a complex clinical phenotype becomes apparent when comparing the performance of marker panels with single markers, with the latter e.g., reviewed by Hellemons et al. for onset and progression of diabetic kidney disease (Hellemons et al., 2012). In clinical practice a different type of biomarker may be utilized, providing a phenotypic readout primarily reflecting the functional status of an organ in contrast to the pathophysiological characteristics. In kidney disease such functional markers are used in patient management as well as clinical trial design, including the estimated glomerular filtration rate (eGFR) and proteinuria (reflecting glomerular filtration and permeation of macromolecules across the glomerular capillary wall, respectively).

Association of these parameters with worsening of diabetic kidney disease, together with increasing incidence of endpoints as cardiovascular events is undisputed (Adler et al., 2003). However, these markers do not provide information on the specific molecular characteristics of the disease. Functional markers render stratification for tailored therapy in the concept of precision medicine essentially impossible.

The molecular pathway of primary interest in the present clinical setting of diabetic kidney disease is the renin-angiotensin system (RAS), in its activity at foremost controlling blood pressure and fluid balance. Blockade of the RAS has been able to reduce the incidence of renal events in patients with and without diabetes mellitus (Ruggenenti et al., 1998; Brenner et al., 2001). In a study by Lewis et al. angiotensin receptor blockade by Irbesartan reduced the risk of a primary composite endpoint (doubling of baseline serum creatinine concentration, development of end-stage renal disease or death from any cause) during a follow up period of 2.6 years by 20% when compared to the placebo (Lewis et al., 2001). Nevertheless, 50% of patients in the Irbesartan group reached the primary endpoint after 54 months. In an effort to increase the efficacy of RAS antagonistic therapy an angiotensin receptor blocker was combined with placebo or the angiotensin converting enzyme (ACE) inhibitor Lisinopril (Fried et al., 2013). The combination therapy did not reduce the incidence of a combined renal endpoint. On the contrary an increased risk of hyperkalemia and acute kidney injury was observed confirming other reports questioning the safety of this approach (Mann et al., 2008; Parving et al., 2008).

Next to addressing RAS, organ-specific molecular processes involving inflammation and oxidative stress have been implicated in progressive tubulointerstitial fibrosis, the best histological, hence molecular mechanistic predictor of an adverse renal disease prognosis (Rodríguez-Iturbe and García García, 2010). Bardoxolone, a nuclear factor-erythroid-2-related factor 2 activator with anti-oxidative capacity increased eGFR in patients with advanced diabetic renal disease (Pergola et al., 2011). However, a large prospective controlled randomized trial with hard endpoints had to be stopped because of severe side effects (De Zeeuw et al., 2013).

As given with these examples for chronic kidney disease (but in its conceptual fundament holding true for a multitude of highly prevalent chronic diseases), many of the recent interventional studies failed to achieve their goals. Here biomarkers promise to take a key role in selecting patients for studies and/or to predict the long term effects of a drug on hard endpoints. Upfront stratification in randomized controlled trials by separating patients by drug response as measured by biomarkers serving as endpoint surrogate and then randomizing the groups separately is an approach which is, at least from a statistical point of view, preferable to *post-hoc* analysis (De Leon, 2012). Such an enrichment strategy is currently e.g., tested in the SONAR study (clinicaltrials.gov reference NCT01858532) addressing diabetic nephropathy (DN).

However, with respect to fit of specific drugs biomarkers need to carry predictive value, i.e., a biomarker shall on a patient-specific level identify responders benefitting from drug effect. In this setting various levels need to be considered involving genetic and environmental components defining disease presentation and progression. The drug target may see genetic polymorphism impacting drug binding, but polymorphism may further involve drug transport and drug metabolism (Johnson, 2001). A significant number of genetic polymorphisms have in the meantime become drug label-relevant regarding drug efficacy, but also toxicity and side effects (U.S. Food and Drug Administration, 2014). Pharmacogenomics has clearly demonstrated that the genetic background of an individual introduces heterogeneity in drug response.

Still, this setting assumes a homogeneous patient population with respect to the molecular mechanistic factors determining disease progression, only exhibiting differences in genetic peculiarities of one and the same molecular mechanistic context. In such setting functional biomarkers appear sufficient for identifying progressive disease, and drug variance is fully explained by the genetic background in regard to the mechanism of action of a specific drug.

A complementary perspective may be that the molecular mechanistic background and progression-relevant molecular factors are *per se* diverse and patient-specific, naturally determining drug response (Mayer et al., 2012). In such scenario a biomarker needs to serve as proxy of key mechanistic factors characterizing and driving a disease on a patient-specific level, combined with educating on the specific interference of disease mechanism with drug mechanism of action. For capturing these constraints a detailed molecular map of a clinical phenotype and its interference with a drug mechanism of action is needed, and here integration of Omics profiling adds to identifying such mechanisms (Fechete et al., 2011; Mühlberger et al., 2012).

An a priori stratification of patients based on an appropriately chosen biomarker panel reflecting the pathophysiology of a given patient (group) allowing to determine a match with a specific drug's mechanism of action appears as promising approach. As recently discussed by Himmelfarb et al. fresh approaches are critical in finding therapies to kidney disease benefiting patients, outlining the importance of improving the translational aspect in clinical research (Himmelfarb and Tuttle, 2013). Here, omics technologies have added significantly to the data landscape characterizing chronic kidney disease, however, in a first instance mainly expanding the candidate set of apparently relevant processes and pathways, going in hand with a large number of biomarker candidates, which individually hamper clinically relevant assessment on disease progression (Fechete et al., 2011; Hellemons et al., 2012).

Integrative approaches in the realm of Systems Biology have been proposed for reaching a consensus description of chronic kidney disease pathophysiology, including molecular models of DN as well as of the reno-cardial axis (He et al., 2012; Komorowsky et al., 2012; Mayer et al., 2012; Heinzel et al., 2013). Still, a translation process needs to be followed, joining disease pathophysiology, stratification markers allowing enrichment strategies, combined with on a molecular mechanistic level matching drugs for allowing precision medicine (Mirnezami et al., 2012). In this work we exemplify such procedure on DN being the major clinical presentation leading to end stage renal disease.

MATERIALS AND METHODS

GENERAL DATA SOURCES

Protein coding genes identified as associated with DN were collected from public domain transcriptomics data sources, complemented with molecular features reporting such association in scientific literature. Molecular signatures educating on ACE inhibitor mechanism of action were extracted from public domain transcriptomics sources. Proteins discussed as biomarkers or drug target candidates in the context of DN were extracted from scientific literature, with the set of targets further extended with known drug targets of drugs currently utilized in clinical trials including renal endpoints. Protein-protein interaction information and molecular pathway maps were retrieved from public domain databases.

Clinical phenotype molecular data

A literature search in NCBI Pubmed utilizing the query string *diabetic nephropathies[majr] AND (microarray analysis[mh] OR gene expression profiling[mh]) AND humans[mh] NOT review* resulted in 37 transcriptomics studies. Explicitly restricting to explorative, array-based mRNA expression studies on human kidney tissue yielded four studies as suitable for inclusion in further analysis. For Berthier et al. and Cohen et al. expression signatures could be retrieved directly from the publications (Cohen et al., 2008; Berthier et al., 2009). For Woroniecka et al. and Baelde et al. the raw expression profiles were retrieved from Gene Expression Omnibus (GSE30122, GSE1009) (Baelde et al., 2004; Woroniecka et al., 2011). Robust Multi-array Average (RMA) normalization for the data set of Woroniecka et al. and MAS5 normalization for the data set of Baelde et al., followed by Significance Analysis of Microarrays (SAM) was employed for identifying features showing differential regulation comparing diabetic kidney disease and healthy control samples. In case of microdissected sample material separate analysis was done for the glomerular and tubulointerstitial compartment.

To further complement the set of DN-associated features a literature mining approach based on Pubmed Medical Subject Headings (MeSH) annotation and publication to gene links provided in gene2pubmed (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz) was executed. A Pubmed search using *diabetic nephropathies[majr] AND human[mh]* as query string was performed for identifying publications of relevance in the context of DN, resulting in 10,766 publications. Protein coding genes explicitly discussed in these publications were extracted from gene2pubmed by filtering based on Pubmed ID and Taxonomy ID (9606 for human).

Finally, the sets of differentially regulated features identified in the individual transcriptomics studies as well as the set of genes from literature extraction were consolidated on the Ensembl gene namespace (**Table 1**).

Biomarker and target annotation from scientific literature

A NCBI Pubmed search for publications holding *Diabetic Nephropathies* further qualified by one of the following qualifiers *pathology, physiopathology, enzymology, metabolism, complications, blood, diagnosis, urine,* and *epidemiology* as major MeSH concept, further demanding one of the MeSH concepts *Biological Markers* or *Tumor Markers, Biological* was performed for identifying publications discussing biomarker candidates. For

Table 1 | Diabetic nephropathy molecular data space.

Data type	Study setup	# Protein coding genes	References
Transcriptomics, tissue biopsies	Comparison of healthy references (GFR > 60) and established DN (GFR 30-59);		Berthier et al., 2009
	Glomerular compartment:	5	
	Tubulointerstitial compartment:	7	
Transcriptomics, tissue biopsies	Comparison of healthy references (GFR > 60) and established DN (GFR 30-59);		Woroniecka et al., 2011
	Glomerular compartment:	164	
	Tubulointerstitial compartment:	183	
Transcriptomics, tissue biopsies	Comparison of healthy references (GFR > 60) and patients with type 2 diabetes > 5 years;		Baelde et al., 2004
	Glomerular compartment:	167	
Transcriptomics, tissue biopsies	Comparison of healthy references and established DN (no further details provided)		Cohen et al., 2008
	Tubulointerstitial compartment:	69	
Literature extraction	PubMed MeSH query as defined in main text	415	–
Total number of unique protein coding genes		881	

Provided is the data type, study setup details, number of protein coding genes identified as DN-associated, and literature reference for a study.

retrieving drug target candidates the term *Diabetic Nephropathies* with the qualifiers *drug therapy* and *therapy* was used, respectively. The search revealed 615 publications for biomarkers and 2,692 for drug targets. Their respective Pubmed IDs were subsequently used for extracting human genes from the gene2pubmed file, resulting in 54 biomarker candidates and 19 drug target candidates.

Target annotation via drugs under investigation

Clinical trial data for completed and currently ongoing clinical trials were retrieved from ClinicalTrials.gov (http://clinicaltrials.gov/). The advanced search as provided on the ClinicalTrials.gov webpage was used for identifying studies that fulfilled the following two criteria: *Study Type* equals *Interventional Studies* and *Condition* contains *Diabetic Nephropathy,* revealing 206 clinical studies. Title and trial description were manually reviewed for focus on renal disease, resulting in 124 studies further considered. Respective drug interventions were mapped to their DrugBank entries (Law et al., 2014), extracting human drug targets as listed, being further mapped on the Ensembl gene namespace. In total 86 drug targets were identified using this approach, of which one was also part of the 19 target candidates retrieved from mining of scientific literature essentially covering basic and translational research activities.

Drug mechanism of action molecular data

A set of ACE inhibitors was retrieved from the Anatomical Therapeutic Chemical (ATC) classification system maintained by the World Health Organization (WHO). 16 compounds classified under *ACE inhibitors, plain* (ATC code: C09AA) were identified and used for subsequent data extraction from DrugMatrix (https://ntp.niehs.nih.gov/drugmatrix/index.html). For six out of the 16 drugs sets of genes being affected by drug presence in rat kidney tissue after drug administration were available within DrugMatrix. Obtained rat gene sets were subsequently mapped from Unigene IDs (Sayers et al., 2009) to Ensembl rat IDs and from there further to human ortholog genes according to Ensembl (**Table 2**).

Table 2 | Drug mechanism of action data space.

Drug name	# Protein coding genes	Database references
Benazepril	442	ICX5600735
Captopril	535	ICX5602791
Enalapril	526	ICX5601254
Lisinopril	558	ICX5601689
Quinapril	572	ICX5602295
Ramipril	519	ICX5602317
Total number of unique protein coding genes	2058	

Given is the drug name, number of associated human protein coding genes identified as significantly affected by drug presence in transcriptomics profiling, and DrugMatrix reference identifier.

MOLECULAR PATHWAY AND PROTEIN INTERACTION DATA

KEGG and Panther pathway membership information for protein coding genes was obtained via KEGG's REST service and from the plain-text database file available on the Panther web site, respectively (Thomas et al., 2003; Kanehisa et al., 2014). Human protein-protein interaction data from BioGRID, INTACT and Reactome were extracted from the respective plain-text files provided by the individual data sources (Stark et al., 2006; Kerrien et al., 2012; Croft et al., 2014). Gene and protein identifiers provided in the original sources were mapped to their respective Ensembl gene IDs. Protein-protein interaction data were further merged into a protein-protein interaction network using Ensembl gene IDs as common denominator of the individual networks.

MOLECULAR PATHWAY AND PROCESS IDENTIFICATION

Molecular pathways and processes were analyzed on the one hand on the basis of a literature review of KEGG and Panther pathways already discussed as relevant in the context of DN. In a second approach de-novo identification of DN molecular processes was performed utilizing the DN pathophysiology feature set. A segmentation algorithm for the identification of processes in the DN protein-protein interaction network was pursued for assembling a molecular process model for DN. Utilizing an analogous procedure a molecular mechanism of action model for ACE inhibitors was constructed utilizing expression signatures obtained from DrugMatrix.

DN pathways from literature

A NCBI Pubmed search for publications utilizing the query string *"diabetic nephropathy"[ti] OR "diabetic nephropathies"[ti]) AND (pathway[ti] OR pathways[ti])* was performed resulting in 53 publications holding the keywords in the title. Subsequently, named entity recognition was performed to annotate occurrence of pathway names according to KEGG and Panther entries in the title and abstract of these publications. Finally, abstracts holding a pathway name were manually reviewed to ensure an association of the identified pathway in the context of DN, leading to 27 individual pathways discussed in literature as being afflicted with DN. Relations between pathways were inferred based on shared genes and the number of protein-protein interactions spanning across pathway boundaries.

Molecular process models

Computing molecular process models followed the procedure described in Mayer et al. (2012); Heinzel et al. (2014). In essence, three main steps are performed: (i) mapping of a feature signature being either the DN pathophysiology association (**Table 1**) or the ACE mechanism of action set (**Table 2**) on the consolidated protein interaction network, followed by induced subgraph extraction. Nodes with a degree of zero are removed from the subgraph. (ii) molecular process identification via utilizing a segmentation algorithm (MCODE with default settings, Bader and Hogue, 2003), and (iii) determining inter-process relations defined by the number of protein-protein interactions observed between any actual two molecular processes contrasted against the number of interactions between two random sets of nodes with matching node set size.

Enrichment analysis

For identifying significance of enrichment of molecular feature sets in molecular processes and pathways a Fisher's exact test with a significance level set to 0.05 was used. Benjamini Hochberg correction was employed to adjust for multiple testing.

RESULTS

DN MOLECULAR PATHWAYS

Screening scientific literature resulted in 27 molecular pathways being observed in the context of DN according to KEGG and Panther pathway annotation (**Figure 1**). The pathway map is dominated by linked signaling components, with major elements being MAPK-VEGF, and Jak-STAT-cytokine-cytokine receptor interaction further interacting with TGF-beta signaling, covering among others mechanisms of hypoxia response and fibrosis, respectively (Rudnicki et al., 2009; Loeffler and Wolf, 2014). Additional mechanistic aspects include stress response and involvement of extracellular matrix (McLennan et al., 2013; Tan and de Haan, 2014). Further, a number of specific pathways in the context of metabolism are included, as well as the RAS, with the latter however showing no direct links to other pathways on the molecular feature overlap or direct protein interaction level.

Screening for biomarker candidates in scientific literature resulted in 54 protein coding genes, extraction of drug target candidates from literature as well as clinical trials brought forward 104 such genes. Of the 54 biomarker candidates 23 are assigned to the DN pathway map, for the 104 target candidates 52 are involved (**Table 3**).

Significant coverage regarding biomarker as well as target candidates is again seen for central signaling components including chemokine signaling, cytokine-cytokine receptor interaction, complemented by MAPK and PI3K-Akt signaling. Also mechanisms are addressed including key features as VEGFA and TGFB1. No specific targeting is seen for counteracting structural changes in ECM, and minor efforts appear to be assigned to adapting stress response. For seven out of 20 pathways discussed no biomarker or target annotation is identified, and complementary a large number of such features are assigned also outside the pathway landscape presented in **Figure 1**. Prominent examples for void biomarker assignment include connective tissue growth factor (CTGF) as factor in fibrosis not being assigned in KEGG, the same being true for uromodulin (UMOD) shown to be associated with progressive disease including genetic polymorphisms (Deshmukh et al., 2013; James et al., 2013). CTGF is also discussed in the therapeutic context via utilizing a monoclonal antibody-based approach (Adler et al., 2010).

Testing the DN pathophysiology feature set retrieved from consolidation of transcriptomics profiles regarding enrichment in the given DN pathway landscape identified seven such pathways as significant, however, missing central mechanisms as hypoxia response or TGFB signaling. In contrast other pathways beyond the map given in **Figure 1** appeared significantly enriched, including focal adhesion, cell adhesion molecules and adherence junctions, linking to the signaling aspects involved in the disease.

DN MOLECULAR MODEL

Complementary to analysis on molecular pathways as defined in KEGG and Panther we performed a network segmentation

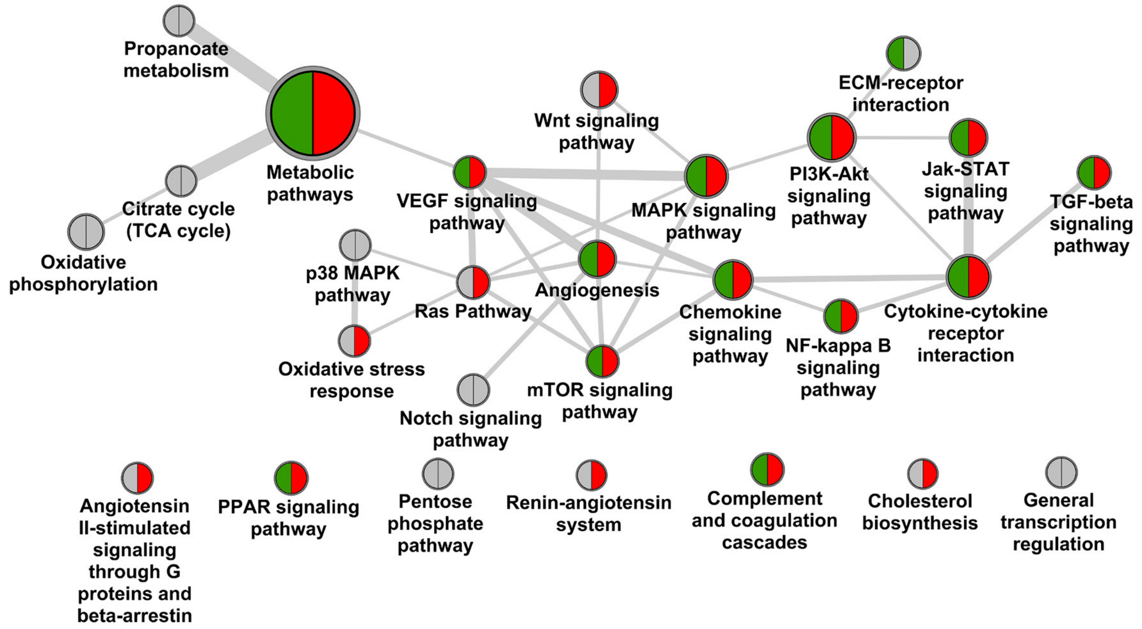

Propanoate metabolism
Metabolic pathways
Citrate cycle (TCA cycle)
Oxidative phosphorylation
Wnt signaling pathway
VEGF signaling pathway
p38 MAPK pathway
Ras Pathway
Angiogenesis
Oxidative stress response
Notch signaling pathway
mTOR signaling pathway
MAPK signaling pathway
ECM-receptor interaction
PI3K-Akt signaling pathway
Jak-STAT signaling pathway
TGF-beta signaling pathway
Chemokine signaling pathway
NF-kappa B signaling pathway
Cytokine-cytokine receptor interaction
Angiotensin II-stimulated signaling through G proteins and beta-arrestin
PPAR signaling pathway
Pentose phosphate pathway
Renin-angiotensin system
Complement and coagulation cascades
Cholesterol biosynthesis
General transcription regulation

FIGURE 1 | Pathway landscape of diabetic nephropathy. Nodes of the graph represent KEGG and Panther pathways (node diameter scales with number of protein coding genes assigned), edges between nodes scale with the number of genes overlapping as well as interactions of genes across pathways according to the protein interaction network. Pathways are marked for holding biomarker candidates (green) and drug target candidates (red).

Table 3 | Molecular pathway annotation, diabetic nephropathy.

Pathway name	# Genes	Biomarker	Drug target	Enrichment
Angiogenesis	148	HSPB2, VEGFA, HSPB2-C11orf52	JUN, VEGFA	No
Angiotensin II-stimulated signaling through G proteins and beta-arrestin	35	–	AGTR1	No
Chemokine signaling	190	CCL2, NFKB1, CCL5	CCL2	Yes
Cholesterol biosynthesis	11	–	HMGCR	No
Complement and coagulation cascades	69	F2, FGB, MBL2	SERPIND1, SERPINC1	Yes
Cytokine-cytokine receptor interaction	272	CCL2, LEP, VEGFA, TNFRSF11B, CCL5, PRL, TGFB1	CCL2, TGFB1, VEGFA, TNFSF12, IL18, IL1B, FLT1	Yes
ECM-receptor interaction	87	SPP1, FN1	–	Yes
Jak-STAT signaling	158	LEP, PRL	SOCS1	No
MAPK signaling	256	TGFB1, FGF23, NFKB1	CACNA1H, CACNA1I, CACNB4, CACNA1S, CASP3, CACNA2D3, TGFB1, CACNB3, CACNA1A, CACNA1B, CACNA1C, CACNA1D, CACNA1F, CACNA1G, JUN, CACNB2, CACNG1, IL1B, CACNA2D1, CACNB1	No
Metabolic pathways	1165	XYLT2, PTGDS, KL, PON1, PON2	PTGS2, PDXK, QPRT, ALOX5, NT5E, IMPDH1, ACSL4, XDH, CES1, NNMT, ANPEP, HMGCR, IMPDH2, CYP11B2	No
mTOR signaling	61	VEGFA	PDPK1, VEGFA, INS	No
NF-kappa B signaling	90	NFKB1	PTGS2, IL1B	No
Oxidative stress response	44	–	JUN	No
PI3K-Akt signaling	345	SPP1, VEGFA, FN1, NFKB1, PRL, FGF23	PDPK1, FLT1, VEGFA, INS	Yes
PPAR signaling	71	ADIPOQ	PPARG, ACSL4, FABP1, PDPK1, PPARA, ADIPOQ	No
Ras Pathway	69	–	PDPK1, JUN	No
Renin-angiotensin system	17	–	ACE2, AGTR1, REN, ANPEP, ACE	Yes
TGF-beta signaling	80	TGFB1, SMAD1	TGFB1	No
VEGF signaling	62	VEGFA	PTGS2, VEGFA	No
Wnt signaling	139	–	JUN	Yes
–	–	SPON2, WTAP, UMOD, LCN2, HP, VNN1, AGER, TGFBI, RBP4, NPHS1, HBA1, HBA2, DEFA1B, LPA, CST3, CTGF, ACTA1, PGC, S100A9, DPP4, ALB, CCKAR, GSTP1, DEFA3, S100A8, DEFA1, MMP9, CDH1, S100A4, NPPB, HAVCR1	SOAT1, SLC6A4, ADORA1, MC2R, SIRT1, CYCS, RETN, EDNRA, CRH, EDNRB, KCNA1, ADORA2A, CALM2, CALM3, CALM1, PTX3, PDE3A, KCNMA1, P2RY12, SLC12A1, SLC12A3, GLP1R, DPP4, PDE5A, NR3C2, KCNJ11, ITGB2, KIF6, MMP9, CA12, TUBB1, NAMPT, HCAR3, HCAR2, AR, HBA1, HBA2, CA9, KCNH2, CA2, CA1, CASP1, TUBB, CA4, AHR, CTGF, ABCA1, PDE4A, PDE4B, SCN5A, MMP2, NPC1L1	
Citrate cycle (TCA cycle)	31	–	–	No
General transcription regulation	30	–	–	No
Notch signaling	48	–	–	No
Oxidative phosphorylation	122	–	–	No
p38 MAPK	34	–	–	No
Pentose phosphate	27	–	–	No
Propanoate metabolism	32	–	–	No

Provided is the KEGG pathway name, number of genes assigned to the pathway according to the pathway source, biomarker, and drug target candidates included in the pathway (gene symbols), and indication of significance of enrichment of such pathway on the basis of the consolidated DN kidney tissue transcriptomics data.

procedure aimed at identifying DN molecular process segments defined by topological characteristics of the DN-specific subgraph. From the in total 881 protein coding genes included in the DN molecular pathophysiology gene set (**Table 1**) 880 were also part of the consolidated interaction network, and 634 were identified as member of the induced subgraph (**Figure 2A**). From the total set of 880 features 246 protein coding genes had no interaction to any other feature of the DN consensus set, hence being disregarded in molecular model computation. Apparent is the relatively minor overlap of features extracted from literature when compared to signatures from transcriptomics. From the in total 516 unique features consolidated from four transcriptomics profiling experiments and 414 features derived from scientific literature 49 are shared.

After MCODE segmentation 200 molecular features remained in process segments, forming a molecular model holding 23 process segments (**Figure 2B**). Median number of protein coding genes per process segment is 6, with the largest segment encoding 29 features, the smallest 3. Equivalently to the pathway graph in **Figure 1** a process graph serves as approximation of individual molecular process characteristics together with their dependencies. Six process segments of the process model hold both, biomarker as well as target candidate annotation, with others encoding just one of the two or none. Of the 54 biomarker candidates 22 are included in the molecular model, the respective number for the 104 targets candidates is 16.

DN MOLECULAR MODEL AND DRUG MECHANISM OF ACTION MODEL INTERFERENCE

Consolidating transcriptomics signatures reflecting the impact of ACE inhibitors on the kidney interactome in a rat model utilizing six representative drugs resulted in 2058 molecular features (**Table 2**), with 661 features being identified in a least two of the six drug signatures. Mapping this consensus ACE feature subset on the consolidated interaction network allowed representation of 656 features. The induced subgraph included 332 features, after segmentation resulting in 12 process segments holding in total 92 molecular features (**Figure 3**, left). Median process feature set size was 8, with a maximum of 19 and a minimum of 3.

Interfering the ACE mechanism of action molecular model with the DN molecular model on the level of feature overlap (**Figure 3**) identified specific process segments of the DN molecular model also holding biomarker candidates (**Table 4**).

All four process segments of DN showing interference with the ACE drug mechanism of action model hold biomarker candidates. Two segments provide significant enrichment also on the level of molecular pathways, showing an integration of chemokine and cytokine signaling, RAS and complement and coagulation cascades for one process segment, the second process segment reflects components of PI3K-Akt signaling in the context of TGFB signaling and ECM receptor interaction.

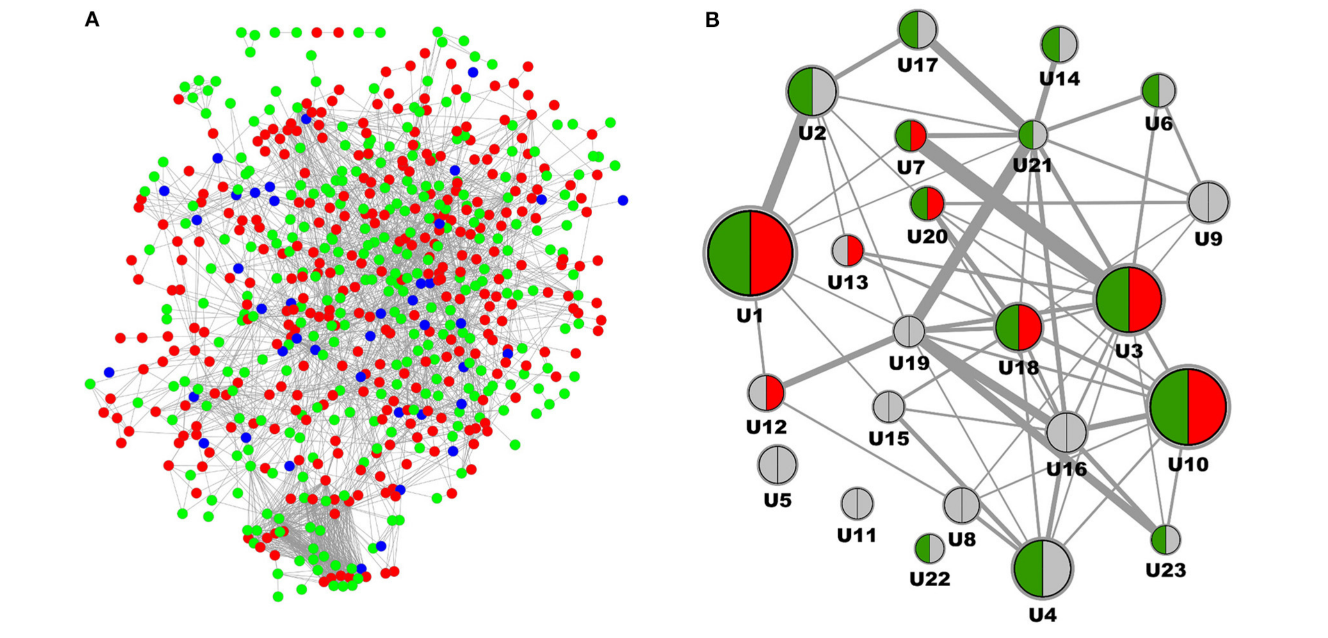


FIGURE 2 | Molecular model representation of diabetic nephropathy. (A) Induced subgraph where each node represents a protein coding gene being reported as associated with DN, edges denote interactions according to the underlying interaction network. Features derived from Omics studies are given in red, features delineated from literature mining are given in green, features identified in both data sources are depicted in blue. **(B)** Molecular model representation of DN where each node represents a process segment with the node diameter scaling with the number of protein coding genes involved, and edges between nodes scaling with the number of interactions of genes across nodes according to the protein interaction network. Segments are indicated for holding biomarker candidates (green) and drug target candidates (red).

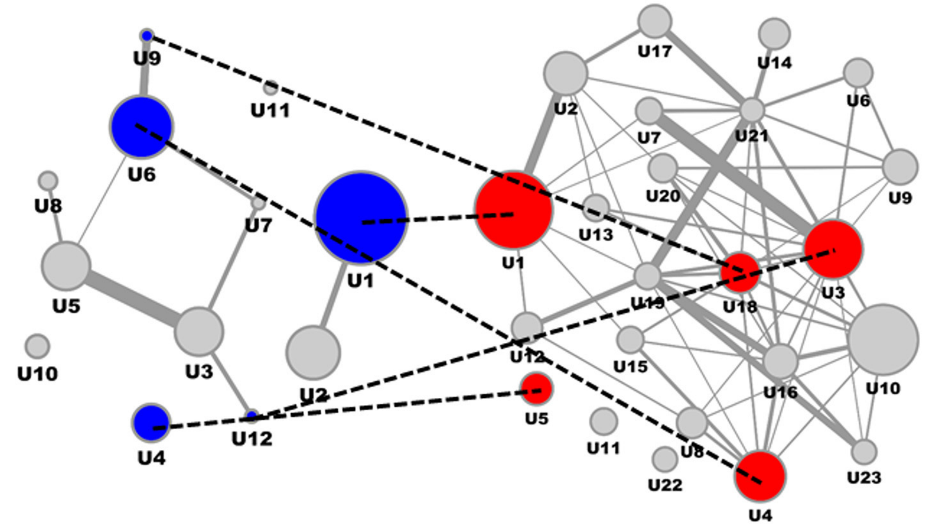


FIGURE 3 | ACE inhibitor mechanism of action molecular model and interference with DN molecular model. ACE Mechanism of Action molecular model (left) and DN molecular model (right), with overlapping process segments of drug and phenotype models indicated by dotted lines. Molecular process segments (U) of the ACE mechanism of action molecular model showing interference with the DN molecular model are given in blue, respective interacting process segments on the DN side are given in red.

Table 4 | Diabetic nephropathy process segment interference.

Segment	# Genes in segment	Interference overlap	Biomarker candidates	Enriched pathways
1	29	7	CCL5	Chemokine signaling; Cytokine-cytokine receptor interaction; Renin-angiotensin system; Complement and coagulation cascades
18	11	2	HBA1, NFKB1, HP, HBA2	–
3	20	3	TGFB1	ECM-receptor interaction; TGF-beta signaling; PI3K-Akt signaling
4	16	2	ACTA1	–

Provided is the process segment number of the DN molecular model, number of genes assigned to the segment, number of features identified as affected according to the drug mechanism of action model, biomarkers involved in the segment (gene symbols), and relevant pathways from the DN pathway map being enriched in such segment.

Biomarker candidates serving as proxy for the interference of ACE and DN molecular models involve the chemokine (C-C motif) ligand 5 involved in immunoregulators and inflammatory processes, hemoglobin alpha 1 and 2 together with haptoglobin, the cytokine transforming growth factor, beta 1, along with the transcription factor NFKB1, finally including actin, alpha 1 involved in cell motility, structure and integrity.

DISCUSSION

For a large spectrum of clinical presentations an impressive number of drug targets have been proposed out of translational and preclinical research, with a significant number further proceeding into clinical trials. Just in the first half of 2014 close to 10,000 new clinical studies were recorded on the platform clinicaltrials.gov. Taking a specific look at diabetic nephropathy as clinical phenotype, 124 interventional trials in any status are identified at clinicaltrials.gov specifically involving the disease term, covering 45 individual drug entities addressing 86 known targets. Via mining scientific literature additional 18 drug targets are identified.

Next to a number of trials utilizing drugs and drug combinations addressing known factors impacting DN progression as the RAS, drug targets are disparately distributed across molecular pathways, hence mechanisms assigned to the disease.

From literature mining 27 different pathways according to KEGG and Panther pathway annotation are discussed as associated with DN, of which 19 hold drug targets. These include well known mechanisms of relevance in DN including hypoxia response or fibrosis, combined with a large set of signaling components. On top, 52 drug targets are embedded in molecular context outside this literature-derived DN pathway landscape.

For biomarker candidates an equivalent situation is found. 54 unique proteins extracted from scientific literature are discussed in any biomarker context, covering 14 of the 27 pathways, with 31 biomarker candidates not assigned to any of the members of the extracted DN pathway map.

Interestingly, predictive performance regarding disease progression of any of the individual biomarker candidates proved limited value. For example, in a review by Hellemons et al. 13 relevant markers were found in the context of nephropathy in diabetes, of which five were found as significantly associated with onset as well as progression of DN again covering various mechanisms including inflammation (e.g., C-reactive protein), cell surface interaction and homeostasis (e.g., E-selectin, ICAM1) and metabolism (triglyceride levels) (Hellemons et al., 2012).

Apparently, individual biomarkers reflecting the status of an individual molecular process, pathway or mechanism cannot capture disease prognosis for the comprehensive DN population. In alternative approaches multimarker panels were included in classifiers on disease diagnosis and prognosis demonstrating improved performance also in blinded validation. In Roscioni et al. a signature of 273 peptides determined in urine were included in a support vector machine-based classifier (Roscioni et al., 2013). The signature held fragments of collagen eventually mirroring alterations in the extracellular matrix turnover and fibrosis together with markers of inflammation as e.g., the proinflammatory protein S100-A9, as well as uromodulin shown to be associated with interstitial fibrosis and tubular atrophy (Nkuipou-Kenfack et al., 2014).

One contributing factor for needing multimarker panels may be individual variance of baseline biomarker levels, where inclusion of multiple markers specifically in non-linear classification methods adds to robustness. However, a second factor may be generic heterogeneity of the patient population. Specific disease presentation may significantly vary not only across stages of disease progression eventually seeing a transition from protective to damaging mechanisms, but even within a specific chronic kidney disease category as defined by present clinical classification provided by KDIGO guidelines (KDIGO Board Members, 2013).

Improved prognostic performance of multimarker panels on top of strict functional classification of stage transitions in DN utilizing albuminuria but also eGFR as clinically used progression parameters clearly support the case of pathophysiological heterogeneity of a, in present clinical terms homogeneous, patient population. However, specifically for albuminuria the role of functional marker vs. factor in disease is discussed (Roscioni et al., 2014).

Deriving robust diagnostic or prognostic classifiers from e.g., proteomics or metabolomics profiling may add to clinical patient management regarding onset as well as intensity of therapeutic measures (Roscioni et al., 2013; Pena et al., 2014). Also in clinical trial design such enrichment strategies may be utilized by e.g., identifying individuals prone to fast disease progression, and randomizing in this high risk cohort into medication and placebo arm (e.g., Priority trial, clinicaltrials.gov reference NCT02040441).

Prognostic biomarkers in contrast to diagnostic parameters with known assignment to molecular processes and pathways further allow an approximation of what specific mechanisms are associated with disease progression. The DN pathway landscape discussed in this work is solely a cross-sectional representation of the disease, in a first place not allowing deciphering which of the 27 individual pathways drive disease progression, and which other pathways are just bystanders or downstream consequences of mechanistic factors of disease. Hence, evaluating biomarker candidates for their association with progressive disease in turn allows determining mechanisms associated with progressive disease. Such knowledge is vital e.g., for determining novel drug targets, demanding to be embedded in disease mechanisms being factors for progressive disease. Remaining question however is if such mechanisms are relevant to the same extent or at all for a specific patient assigned to a clinical phenotype.

A prognostic biomarker set covering all potentially relevant processes enables specific molecular phenotyping of individual patients, being however not sufficient in terms of predicting drug response as a drug mechanism of action is not factored in. Here Systems Pharmacology aims at identifying drug response also on the level of molecular processes and pathways. Rationale is to not only focus on the specific drug target and its assignment to specific mechanisms, but to include the systemic molecular changes triggered by the drug including off-target effects as well as downstream molecular changes. Having a drug mechanism of action as well as a clinical phenotype represented on a molecular process or pathway level allows intersecting both molecular states. If from prognostic biomarker profiling of a patient specific progression-associated molecular disease mechanisms are identified, and a drug exhibits functional interference in such specific mechanisms such patient may be more prone for showing response to the drug. With such setting including knowledge on molecular phenotype composition, molecular process relevance in progressive disease and knowledge on interference of drug mechanism of action biomarker candidates initially serving a prognostic purpose can be rendered into predictive biomarkers on drug response.

Omics profiling has a major contribution to characterizing both, clinical phenotypes as well as drug mechanism of action. Integrating profiling results from clinical samples frequently sees minor overlap of individual studies, being in part driven by insufficient sample size combined with diverging inclusion criteria and sample material used (Fechete et al., 2011). In the example presented here 1010 features in total are identified as differentially regulated in transcriptomics or are being assigned to DN according to literature mining, with 880 unique features. An equivalent misbalance in feature coherence across studies is also found for the ACE inhibitor transcriptomics data. All these drugs address the same functional context, but from the in total 3152 features identified for six drugs included the total number of unique features are still 2058, with 661 being identified in at least two drug signatures.

Next divergence becoming apparent is the limited overlap of enrichment analysis based on signatures from profiling and feature-based literature mining compared to explicit literature mining for molecular pathways. Of the 27 pathways extracted from scientific references only seven are confirmed, however, seeing other pathways enriched not found via literature mining. On top, a major shortcoming is restricted representation of protein coding genes in such pathway maps, e.g., for KEGG covering 6491 and for Panther 2163 protein coding genes, respectively. This limitation not only affects pathway enrichment but also assignment of biomarker and target candidates. Of the in total 104 drug target and 54 biomarker candidates 29 are neither assigned in any KEGG or Panther pathway.

Here a different approach may be followed, namely segmentation of protein interaction networks exhibiting improved coverage of the protein coding gene set. Consolidation of INTACT, Reactome, and BioGRID allows representation of in total 13,907 protein coding genes, clearly expanding beyond public domain pathway databases. In alternative approaches hybrid interaction networks are utilized for further expanding coverage of protein coding genes, but also for improving false negative rates regarding protein-protein interactions and relations (Fechete et al., 2013).

Computing a DN-specific as well as ACE inhibitor-specific induced subgraph followed by topology-based segmentation allows an alternative representation of a molecular process landscape for the clinical presentation as well as the drug mechanism of action. Interference analysis on the level of overlapping protein coding genes resulted in four process segments holding central aspects of DN pathophysiology. Seven biomarker candidates were identified in these interfering molecular processes. CCL5 (RANTES), involved in recruiting monocytes and macrophages to the renal cortex was shown to be suppressed by ACE inhibition, indicating that RANTES expression is mediated via Angiotensin II type 2 receptor (Kashiwagi et al., 2002). Equivalently, in animal models TGFB1 expression was shown to be reduced by ACE inhibitors. Activation of NFKB1 by angiotensin II was shown in vascular smooth muscle and mesangial cells (Hernández-Presa et al., 1997). In a study by Dong et al. analyzing cost effectiveness of ACE inhibitor treatment for patients with type 1 diabetes mellitus the level of glycosylated HbA1c showed clear impact on cost effectiveness of drug use per quality-adjusted life year (QALY) (Dong et al., 2004). The authors concluded that next to patient age also other factors need to be included in therapy considerations.

Apparently, drug mechanism of action affects numerous molecular processes, as exemplified for ACE inhibitors, many of these also afflicted with DN progression. Analyzing the molecular process interface of disease progression-relevant pathophysiology and drug mechanism of action allows proposing predictive markers. Testing such predictive biomarker candidates may educate on relevance of individual processes on a patient level, directly linking to likelihood of drug response.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme under the grant agreement no. 241544 (SysKid).

REFERENCES

Adler, A. I., Stevens, R. J., Manley, S. E., Bilous, R. W., Cull, C. A., and Holman, R. R. (2003). Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). *Kidney Int.* 63, 225–232. doi: 10.1046/j.1523-1755.2003.00712.x

Adler, S. G., Schwartz, S., Williams, M. E., Arauz-Pacheco, C., Bolton, W. K., Lee, T., et al. (2010). Phase 1 study of anti-CTGF monoclonal antibody in patients with diabetes and microalbuminuria. *Clin. J. Am. Soc. Nephrol.* 5, 1420–1428. doi: 10.2215/CJN.09321209

Bader, G. D., and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Baelde, H. J., Eikmans, M., Doran, P. P., Lappin, D. W. P., de Heer, E., and Bruijn, J. A. (2004). Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. *Am. J. Kidney Dis.* 43, 636–650. doi: 10.1053/j.ajkd.2003.12.028

Berthier, C. C., Zhang, H., Schin, M., Henger, A., Nelson, R. G., Yee, B., et al. (2009). Enhanced expression of Janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. *Diabetes* 58, 469–477. doi: 10.2337/db08-1328

Brenner, B. M., Cooper, M. E., de Zeeuw, D., Keane, W. F., Mitch, W. E., Parving, H. H., et al. (2001). Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N. Engl. J. Med.* 345, 861–869. doi: 10.1056/NEJMoa011161

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., et al. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* 98, 1183–1192. doi: 10.1093/jnci/djj329

Cohen, C. D., Lindenmeyer, M. T., Eichinger, F., Hahn, A., Seifert, M., Moll, A. G., et al. (2008). Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis. *PLoS ONE* 3:e2937. doi: 10.1371/journal.pone.0002937

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

De Leon, J. (2012). Evidence-based medicine versus personalized medicine: are they enemies? *J. Clin. Psychopharmacol.* 32, 153–164. doi: 10.1097/JCP.0b013e3182491383

Deshmukh, H. A., Palmer, C. N. A., Morris, A. D., and Colhoun, H. M. (2013). Investigation of known estimated glomerular filtration rate loci in patients with type 2 diabetes. *Diabet. Med.* 30, 1230–1235. doi: 10.1111/dme.12211

De Zeeuw, D., Akizawa, T., Audhya, P., Bakris, G. L., Chin, M., Christ-Schmidt, H., et al. (2013). Bardoxolone methyl in type 2 diabetes and stage 4 chronic kidney disease. *N. Engl. J. Med.* 369, 2492–2503. doi: 10.1056/NEJMoa1306033

Dong, F. B., Sorensen, S. W., Manninen, D. L., Thompson, T. J., Narayan, V., Orians, C. E., et al. (2004). Cost effectiveness of ACE inhibitor treatment for patients with type 1 diabetes mellitus. *Pharmacoeconomics* 22, 1015–1027. doi: 10.2165/00019053-200422150-00005

Fechete, R., Heinzel, A., Perco, P., Mönks, K., Söllner, J., Stelzer, G., et al. (2011). Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics Clin. Appl.* 5, 354–366. doi: 10.1002/prca.201000136

Fechete, R., Heinzel, A., Soellner, J., Perco, P., Lukas, A., and Mayer, B. (2013). Using information content for expanding human protein coding gene interaction networks. *J. Comput. Sci. Syst. Biol.* 6, 73–82. doi: 10.4172/jcsb.1000102

Fried, L. F., Emanuele, N., Zhang, J. H., Brophy, M., Conner, T. A., Duckworth, W., et al. (2013). Combined angiotensin inhibition for the treatment of diabetic nephropathy. *N. Engl. J. Med.* 369, 1892–1903. doi: 10.1056/NEJMoa1303154

He, J. C., Chuang, P. Y., Ma'ayan, A., and Iyengar, R., (2012). Systems biology of kidney diseases. *Kidney Int.* 81, 22–39. doi: 10.1038/ki.2011.314

Heinzel, A., Fechete, R., Mühlberger, I., Perco, P., Mayer, B., and Lukas, A. (2013). Molecular models of the cardiorenal syndrome. *Electrophoresis* 34, 1649–1656. doi: 10.1002/elps.201200642

Heinzel, A., Fechete, R., Söllner, J., Perco, P., Heinze, G., Oberbauer, R., et al. (2012). Data graphs for linking clinical phenotype and molecular feature space. *Int. J. Syst. Biol. Biomed. Technol.* 1, 11–25. doi: 10.4018/ijsbbt.2012010102

Heinzel, A., Mühlberger, I., Fechete, R., Mayer, B., and Perco, P. (2014). Functional molecular units for guiding biomarker panel design. *Methods Mol. Biol.* 1159, 109–133. doi: 10.1007/978-1-4939-0709-0_7

Hellemons, M. E., Kerschbaum, J., Bakker, S. J. L., Neuwirt, H., Mayer, B., Mayer, G., et al. (2012). Validity of biomarkers predicting onset or progression of nephropathy in patients with Type 2 diabetes: a systematic review. *Diabet. Med.* 29, 567–577. doi: 10.1111/j.1464-5491.2011.03437.x

Hernández-Presa, M., Bustos, C., Ortego, M., Tuñon, J., Renedo, G., Ruiz-Ortega, M., et al. (1997). Angiotensin-converting enzyme inhibition prevents arterial nuclear factor-kappa B activation, monocyte chemoattractant protein-1 expression, and macrophage infiltration in a rabbit model of early accelerated atherosclerosis. *Circulation* 95, 1532–1541.

Hicks, D. G., and Kulkarni, S. (2008). HER2+ breast cancer: review of biologic relevance and optimal use of diagnostic tools. *Am. J. Clin. Pathol.* 129, 263–273. doi: 10.1309/99AE032R9FM8WND1

Himmelfarb, J., and Tuttle, K. R. (2013). New therapies for diabetic kidney disease. *N. Engl. J. Med.* 369, 2549–2550. doi: 10.1056/NEJMe1313104

Hollebecque, A., Massard, C., and Soria, J.-C. (2014). Implementing precision medicine initiatives in the clinic: a new paradigm in drug development. *Curr. Opin. Oncol.* 26, 340–346. doi: 10.1097/CCO.0000000000000077

James, L. R., Le, C., Doherty, H., Kim, H.-S., and Maeda, N. (2013). Connective tissue growth factor (CTGF) expression modulates response to high glucose. *PLoS ONE* 8:e70441. doi: 10.1371/journal.pone.0070441

Johnson, J. A. (2001). Drug target pharmacogenomics: an overview. *Am. J. Pharmacogenomics* 1, 271–281. doi: 10.2165/00129785-200101040-00004

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kashiwagi, M., Masutani, K., Shinozaki, M., and Hirakata, H. (2002). MCP-1 and RANTES are expressed in renal cortex of rats chronically treated with nitric oxide synthase inhibitor. Involvement in macrophage and monocyte recruitment. *Nephron* 92, 165–173. doi: 10.1159/000064454

KDIGO Board Members. (2013). KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int. Suppl.* 3, 1–150. doi: 10.1038/kisup.2012.73

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.

Komorowsky, C. V., Brosius, F. C., Pennathur, S., and Kretzler, M. (2012). Perspectives on systems biology applications in diabetic kidney disease. *J. Cardiovasc. Transl. Res.* 5, 491–508. doi: 10.1007/s12265-012-9382-7

LaFramboise, W. A., Dhir, R., Kelly, L. A., Petrosko, P., Krill-Burger, J. M., Sciulli, C. M., et al. (2012). Serum protein profiles predict coronary artery disease in symptomatic patients referred for coronary angiography. *BMC Med.* 10:157. doi: 10.1186/1741-7015-10-157

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Lewis, E. J., Hunsicker, L. G., Clarke, W. R., Berl, T., Pohl, M. A., Lewis, J. B., et al. (2001). Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N. Engl. J. Med.* 345, 851–860. doi: 10.1056/NEJMoa011303

Loeffler, I., and Wolf, G. (2014). Transforming growth factor-β and the progression of renal disease. *Nephrol. Dial. Transplant.* 29(Suppl. 1), i37–i45. doi: 10.1093/ndt/gft267

Mann, J. F. E., Schmieder, R. E., McQueen, M., Dyal, L., Schumacher, H., Pogue, J., et al. (2008). Renal outcomes with telmisartan, ramipril, or both, in people at high vascular risk (the ONTARGET study): a multicentre, randomised, double-blind, controlled trial. *Lancet* 372, 547–553. doi: 10.1016/S0140-6736(08)61236-2

Mayer, P., Mayer, B., and Mayer, G. (2012). Systems biology: building a useful model from multiple markers and profiles. *Nephrol. Dial. Transplant.* 27, 3995–4002. doi: 10.1093/ndt/gfs489

McLennan, S. V., Abdollahi, M., and Twigg, S. M. (2013). Connective tissue growth factor, matrix regulation, and diabetic kidney disease. *Curr. Opin. Nephrol. Hypertens.* 22, 85–92. doi: 10.1097/MNH.0b013e32835b4889

Mirnezami, R., Nicholson, J., and Darzi, A. (2012). Preparing for precision medicine. *N. Engl. J. Med.* 366, 489–491. doi: 10.1056/NEJMp1114866

Mühlberger, I., Mönks, K., Fechete, R., Mayer, G., Oberbauer, R., Mayer, B., et al. (2012). Molecular pathways and crosstalk characterizing the cardiorenal syndrome. *OMICS* 16, 105–112. doi: 10.1089/omi.2011.0121

Nkuipou-Kenfack, E., Duranton, F., Gayrard, N., Argilés, A., Lundin, U., Weinberger, K. M., et al. (2014). Assessment of metabolomic and proteomic biomarkers in detection and prognosis of progression of renal function in chronic kidney disease. *PLoS ONE* 9:e96955. doi: 10.1371/journal.pone.0096955

Pammolli, F., Magazzini, L., and Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* 10, 428–438. doi: 10.1038/nrd3405

Parving, H.-H., Persson, F., Lewis, J. B., Lewis, E. J., and Hollenberg, N. K. (2008). Aliskiren combined with losartan in type 2 diabetes and nephropathy. *N. Engl. J. Med.* 358, 2433–2446. doi: 10.1056/NEJMoa0708379

Pena, M. J., Lambers Heerspink, H. J., Hellemons, M. E., Friedrich, T., Dallmann, G., Lajer, M., et al. (2014). Urine and plasma metabolites predict the development of diabetic nephropathy in individuals with Type 2 diabetes mellitus. *Diabet. Med*. 31, 1138–1147. doi: 10.1111/dme.12447

Pergola, P. E., Raskin, P., Toto, R. D., Meyer, C. J., Huff, J. W., Grossman, E. B., et al. (2011). Bardoxolone methyl and kidney function in CKD with type 2 diabetes. *N. Engl. J. Med.* 365, 327–336. doi: 10.1056/NEJMoa1105351

Rodríguez-Iturbe, B., and García García, G. (2010). The role of tubulointerstitial inflammation in the progression of chronic renal failure. *Nephron Clin. Pract.* 116, c81–c88. doi: 10.1159/000314656

Roscioni, S. S., de Zeeuw, D., Hellemons, M. E., Mischak, H., Zürbig, P., Bakker, S. J. L., et al. (2013). A urinary peptide biomarker set predicts worsening of albuminuria in type 2 diabetes mellitus. *Diabetologia* 56, 259–267. doi: 10.1007/s00125-012-2755-2

Roscioni, S. S., Lambers Heerspink, H. J., and de Zeeuw, D. (2014). Microalbuminuria: target for renoprotective therapy PRO. *Kidney Int.* 86, 40–49. doi: 10.1038/ki.2013.490

Rudnicki, M., Perco, P., Enrich, J., Eder, S., Heininger, D., Bernthaler, A., et al. (2009). Hypoxia response and VEGF-A expression in human proximal tubular epithelial cells in stable and progressive renal disease. *Lab. Invest.* 89, 337–346. doi: 10.1038/labinvest.2008.158

Ruggenenti, P., Perna, A., Gherardi, G., Gaspari, F., Benini, R., and Remuzzi, G. (1998). Renal function and requirement for dialysis in chronic nephropathy patients on long-term ramipril: REIN follow-up trial. Gruppo Italiano di Studi Epidemiologici in Nefrologia (GISEN). Ramipril Efficacy in Nephropathy. *Lancet* 352, 1252–1256.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2009). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 37, D5–D15. doi: 10.1093/nar/gkn741

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

Tan, S. M., and de Haan, J. B. (2014). Combating oxidative stress in diabetic complications with Nrf2 activators: how much is too much? *Redox Rep.* 19, 107–117. doi: 10.1179/1351000214Y.0000000087

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi: 10.1101/gr.772403

Trusheim, M. R., Burgess, B., Hu, S. X., Long, T., Averbuch, S. D., Flynn, A. A., et al. (2011). Quantifying factors for the success of stratified medicine. *Nat. Rev. Drug Discov.* 10, 817–833. doi: 10.1038/nrd3557

U.S. Food and Drug Administration. (2014). *Table of Pharmacogenomic Biomarkers in Drug Labeling* [Internet].

Woroniecka, K. I., Park, A. S. D., Mohtat, D., Thomas, D. B., Pullman, J. M., and Susztak, K. (2011). Transcriptome analysis of human diabetic kidney disease. *Diabetes* 60, 2354–2369. doi: 10.2337/db10-1181

Zürbig, P., Jerums, G., Hovind, P., Macisaac, R. J., Mischak, H., Nielsen, S. E., et al. (2012). Urinary proteomics for early diagnosis in diabetic nephropathy. *Diabetes* 61, 3304–3313. doi: 10.2337/db12-0348

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

Received: 24 June 2014; accepted: 29 July 2014; published online: 22 August 2014.

Citation: Heinzel A, Perco P, Mayer G, Oberbauer R, Lukas A and Mayer B (2014) From molecular signatures to predictive biomarkers: modeling disease pathophysiology and drug mechanism of action. Front. Cell Dev. Biol. ***2****:37. doi: 10.3389/fcell.2014.00037*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in Cell and Developmental Biology

METHODS
published: 24 June 2015
doi: 10.3389/fcell.2015.00028

How to build personalized multi-omics comorbidity profiles

Mohammad Ali Moni [1, 2, 3*†] *and Pietro Liò* [1]

[1] Computer Laboratory, University of Cambridge, Cambridge, UK, [2] Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh, [3] Bone Biology, Garvan Institute of Medical Research, The University of New South Wales, Sydney, NSW, Australia

OPEN ACCESS

Edited by:
Christine Nardini,
Partner Institute for Computational Biology, China

Reviewed by:
Anshu Bhardwaj,
Council of Scientific and Industrial Research, India
Maria Secrier,
European Molecular Biology Laboratory, Germany
Chenggang Yu,
Henry M. Jackson Foundation for the Advancement of Military Medicine, USA

***Correspondence:**
Mohammad Ali Moni,
Bone Biology, Garvan Institute of Medical Research, The University of New South Wales, 384 Victoria street, Durlinghurst, Sydney, NSW 2010, Australia
m.moni@garvan.org.au

†Present Address:
Mohammad Ali Moni,
Garvan Institute of Medical Research, University of New South Wales, Sydney, Australia

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

Received: 30 September 2014
Accepted: 27 April 2015
Published: 24 June 2015

Citation:
Moni MA and Liò P (2015) How to build personalized multi-omics comorbidity profiles. Front. Cell Dev. Biol. 3:28. doi: 10.3389/fcell.2015.00028

Multiple diseases (acute or chronic events) occur together in a patient, which refers to the disease comorbidities, because of the multi ways associations among diseases. Due to shared genetic, molecular, environmental, and lifestyle-based risk factors, many diseases are comorbid in the same patient. Methods for integrating multiple types of omics data play an important role to identify integrative biomarkers for stratification of patients into groups with different clinical outcomes. Moreover, integrated omics and clinical information may potentially improve prediction accuracy of disease comorbidities. However, there is a lack of effective and efficient bioinformatics and statistical software for true integrative data analysis. With the availability of the wide spread huge omics, phenotype and ontology information, it is becoming more and more practical to help doctors in clinical diagnostics and comorbidity prediction by providing appropriate software tool. We developed an R software `POGO` to compute novel estimators of the disease comorbidity risks and patient stratification. Starting from an initial diagnosis, omics and clinical data of a patient the software identifies the association risk of disease comorbidities. The input of this software is the initial diagnosis of a patient and the output provides evidence of disease comorbidities. The functions of `POGO` offer flexibility for diagnostic applications to predict disease comorbidities, and can be easily integrated to high–throughput and clinical data analysis pipelines. `POGO` is compliant with the Bioconductor standard and it is freely available at www.cl.cam.ac.uk/∼mam211/POGO/.

Keywords: comorbidity, multi-omics, ontology, multiplex network, data integration

Introduction

Exploring disease-disease associations by using multi-omics and clinical information is expected to improve our current knowledge of disease relationships, which may lead to further improvements in disease diagnosis, prognosis and treatment (Park et al., 2009). Recent research has increasingly demonstrated that many seemingly dissimilar diseases have common molecular mechanisms and strong associations among them (Yu and Wang, 2015). Because of the associations among diseases, multiple diseases (acute or chronic events) occur together in a patient, which is called disease comorbidities. Comorbidities relationships exist among diseases whenever they impact the same patients significantly more than expected by chance (Žitnik et al., 2013). It represents the co–occurrence of diseases or presence of different illness or medical conditions simultaneously or one after another in the same patient (Hidalgo et al., 2009; Park et al., 2009). The set of sequential disease associations, which refers to disease trajectories, uncovers time based disease comorbidity associations. They can also form the basis for understanding mathematical properties of co-morbidity networks (Hidalgo et al., 2009; Jensen et al., 2014). Comorbidity associations can be due to direct or indirect causal relationships and the shared risk factors among them (Tong and Stevenson, 2007). If two diseases have comorbidity association, the incidence of one of them in an individual may increase the likelihood of another disease occurring. Certain diseases, such as diabetes and obesity often co-occur in the same patient, sometimes one being considered a significant risk factor for the other (Lee et al., 2008). Disease comorbidities are increasingly placing a greater burden on individuals, societies and health care services. It is an important factor for better risk stratification of patients and treatment planning.

Diseases with similar molecular, environmental, and lifestyle risk factors may be comorbid in individuals or may be risk factors for another disorder (Davis et al., 2010). Shared genetic, environmental and lifestyle factors have similar consequences, increasing the co-occurrence of associated diseases in the same individual. So, a person diagnosed for a combination of disorders and exposed to particular environmental, lifestyle and genetic risk factors may be at a increased risk of developing several other genetically and environmentally associated diseases (Barabási et al., 2011). It is now well accepted that phenotypes are determined by genetic material under environmental influences. For instance, many well-known and influential lifestyle factors such as smoking, diet, and alcohol intake are actively related to diabetes type 1 and type 2, and obesity (Astrup, 2001). Moreover, many complex diseases, such as cancer and diabetes, are affected by an integrated effect of environment and epistasis among many genes (Davis et al., 2010).

Recent evidence has exhibited that microRNAs play key roles in the evolution and progression of human diseases. Functionally related microRNAs tend to be associated with phenotypically similar diseases (Lu et al., 2008). Recently, genome-wide association studies (gwas) proved to be useful as a method for exploring phenotypic associations with diseases (Lewis et al., 2011). Single-nucleotide polymorphisms (SNPs), a variation of a single nucleotide, are assumed to play a major role in causing phenotypic differences between individuals. It has become possible to assess systematically the contribution of common SNPs to complex diseases. Copy number variations (CNVs; which involve loss, duplication or rearrangement of long stretches of DNA in individual's genome) can cause various phenotypic abnormalities (Zhang et al., 2009). CNVs are significantly associated with the risk of complex human diseases including inflammatory autoimmune disorders, diabetes etc. (Bae et al., 2011). The development of type 2 diabetes has also been known to be influenced by molecular, lifestyle and environmental factors (Kahn et al., 2006).

Most of the research works focussed on a particular data type, for example gene expression, to find profiles that are associated with particular disease, prognosis and drug response. The integrative analysis of various omics data has become increasingly widespread because each approach has intrinsic caveats. For instance, important information may be missing because of false negatives or may be misleading because of false positives. In addition, by analyzing different types of data in isolation we may miss important information that results from the coordinated activity of biological components at various levels. Some studies indicated that these limitations can be mitigated by integrating two or more omics datasets. Several studies (Goh et al., 2007; Lee et al., 2008; Lu et al., 2008; Hu and Agarwal, 2009; Liu et al., 2009; Park et al., 2009; Schadt, 2009; Jiang et al., 2010; Suthram et al., 2010) reported on the role of a single omic or phenotypic measure to represent disease-disease associations (such as shared pathways or gene ontology). But, one needs to study diverse sources of evidence including miRNA-based relationships, shared environmental factors, ontology, SNPs, CNVs and phenotypic manifestations for better understanding.

Since, diseases may share many different types of associations with varying levels of risk for disease comorbidities, a singular view of associations between diseases is not enough to predict comorbidities. As more and more ontology, phenotype, omics and environmental data sets become publicly available, it is beneficial to improve our understanding of human diseases and diseases comorbidities based on these new system-level biological data. Combination of multiple types of omics, phenotype and ontology data identifies integrative biomarkers for the stratification of patients with clinical outcome. Further, behavioral and environmental aspects should also be considered in order to realize disease-disease associations. Therefore, it is clear that method and tool for stratifying patients and prediction of disease comorbidities in order to reliably predict prognosis or success of treatments are of critical importance in the field of medicine. We propose a computational framework that integrates all available, heterogeneous and relevant data including miRNA-target interactions, miRNA-disease association, phenotype similarities of diseases, GO (gene ontology), SNPs, CNVs and known disease-environmental associations to capture the complex relationships between phenotypes, genotypes and clinical comordibidity. Therefore, the underlying goal of this chapter is to integrate diverse sets of omics, environmental and phenotypic data, and to develop the comprehensive models of interaction between the disease associated factors for the prediction of the patient specific disease comorbidity, and to develop comorbidity map.

In the case of a complex or even in an unknown case of diseases, physicians may get assistance to take decision quickly and efficiently by using effective software tool. We developed an R software tool POGO to compute statistically significant associations among diseases, to predict disease comorbidity risk and to develop comorbidity maps, which are useful for the physicians and informative for the patients. To perform the computation of the comorbidity risk, this software uses clinical, gene expression, miRNA, SNPs, CNVs, ontology, phenotypic, and environmental data. The inputs of this software is the initial diagnostic result of the patient. The goal of this software is to construct comorbidity maps that incorporate disease interactions, omics, phenotypic and ontology information, and environmental influences. It is a user-friendly and interactive personalised disease and disease comorbidity prediction software. It provides different comorbidity assessment and stratification; integration of omics information with

POGO output data could be used to predict more accurate survival probability of patients. The functions included in POGO offer flexibility for applications, and can be easily integrated into highthroughput analysis pipelines for translation medicine.

Implementation

POGO provides a number of processing options to find comorbidity maps of a patient. R bioconductor annotation data packages "org.Hs.eg.db," "HPO.db," and "GO.db" are used for the annotation and mapping between gene symbol, Entrez id, HPO term, OMIM id and GO term (Gentleman et al., 2004). POGO is dependent on "DOSE" and "GOSemSim" bioconductor packages for the mapping with different annotation (Yu and Wang, 2015). We used the mapping manually constructed by Goh et al. (2007) and Park et al. (2009) to convert OMIM IDs to ICD-9 codes. A set of differential expressed gene symbols/Entrez ids/OMIM id/miRNA ids/HPO terms/GO terms/3 or 5 digit ICD-9-CM code of any disease can be used as input of POGO functions. Flow diagram of POGO software is shown in **Figure 1**.

GO–disease Association

GO enables us to analyse disease association by adopting semantic similarity measures to expand our knowledge of the relationships among different diseases. We downloaded the ontology file and annotations of Homo sapiens from the Gene Ontology database[1] in April 2014. In total, we collected 171,888 annotations between 13,166 genes and 10,787 GO terms. We developed a function comorbidityGO for the computation of GO based disease comorbidity in an ontology sense. It is a GO-based enrichment analysis function to measure association among diseases and to explore their functional associations from gene sets. We implemented a semantic similarity measurement to quantify the association between gene ontology and their associated diseases. The semantics of GO terms are encoded into a numeric format and the different semantic contributions of the distinct relations are considered. Moreover, hypergeometric test is applied to a gene set to calculate the significance of a GO term, and the significant GO term sets are selected according to their *p*-values. Gene set enrichment analysis are used for predicting

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

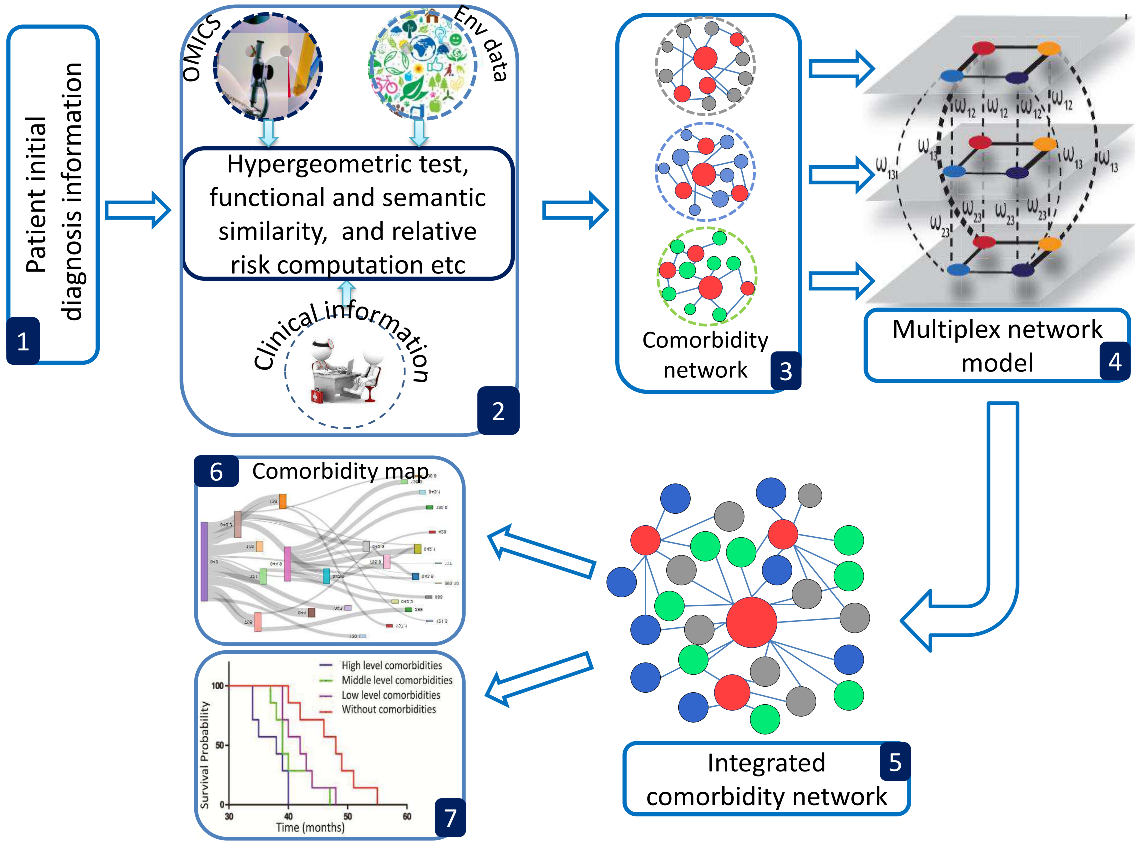


FIGURE 1 | Overview framework of POGO software. (1) POGO takes as input preliminary diagnosis data of a patient and check the validation of the input. (2) It preprocesses and updates required databases, performs statistical computation (hypergeometric and semantic similarity tests), and calculates relative risk between diseases. (3) Comorbidity scores and disease network are provided as a result to the user. (4) Multiplex model is applied for data integration to produce integrated comorbidity network as (5). (6, 7) Visualization of the comorbidity map and survival probability of patient considering comorbidity. Env is used to indicate environment.

the significance of gene–disease and disease–disease associations. `comorbidityGO` function operates by using either of the following input: GO id, disease OMIM id, a list of gene symbols, Entrez gene ids or ICD-9 code of the patient disease. This function provides disease comorbidity associations and network based on the GO. `comorbidityGO` requires two parameters: id list and id type. An example and its output is given in **Figure 2**.

```
> comorbidityGO( "189907" , "OMIM" )

OMIM            GO EVIDENCE ONTOLOGY  PATH   SYMBOL ENTREZID ICD9CM
189907 GO:0000122       IEA       BP 04950     TCF2     6928    250
189907 GO:0001714       IEA       BP 04950   BCKDHB      594  270.3
189907 GO:0005634       IDA       CC 04950     TCF2     6928    189
189907 GO:0044212       IEA       MF 04950     TCF2     6928  593.9
... ...
```

Phenotype–disease Association

`POGO` integrated HPO database that has integrated HPO terms to represent patients phenotypic abnormalities (Robinson et al., 2008). The OMIM (McKusick, 2007) is also incorporated with `POGO`, and associated to HPO by annotations from http://www.human-phenotype-ontology.org. The associations are generated using the information about the phenotypes of a particular syndrome and the corresponding genes that are known to cause this syndrome when mutated. With the development of omics techniques, the number of uncovered gene-phenotype associations has increased notably over the last few years. In our approach, phenotypes are linked with diseases through associating phenotype-gene with gene-disease bipartite graphs by applying neighborhood-based methods. All the paths from a phenotype to a disease are explored by considering causative genes to assign a weight based on frequency and linked the phenotype to the disease in a new phenotype-disease bipartite graph. Then, we introduced a Bidirectionally-induced Importance Weight prediction method to phenotype-disease bipartite graph in order to approximate the weights of the edges of diseases with phenotypes, by considering link information from both sides of the phenotype-disease bipartite graph. The construction of the phenotype network is based on the phenotypic similarity score among different disease phenotypes. In the phenotype network, the association between any two different disease phenotypes was fixed when their phenotypic similarity score exceeded the significance threshold. For visualization, `POGO` includes links between disease pairs for which the co-occurrence is notably greater than the random expectation based on phenotype prevalence of the diseases. The function `comorbidityHPO` of `POGO` package is able to take input an OMIM id/3 or 5 digit ICD-9-CM code of a disease or a list of gene symbols/Entrez ids and provides comorbidity pattern of diseases based on the phenotype disease associations. `comorbidityHPO` requires two parameters: id list and id type. An example and its output is given in **Figure 3**.

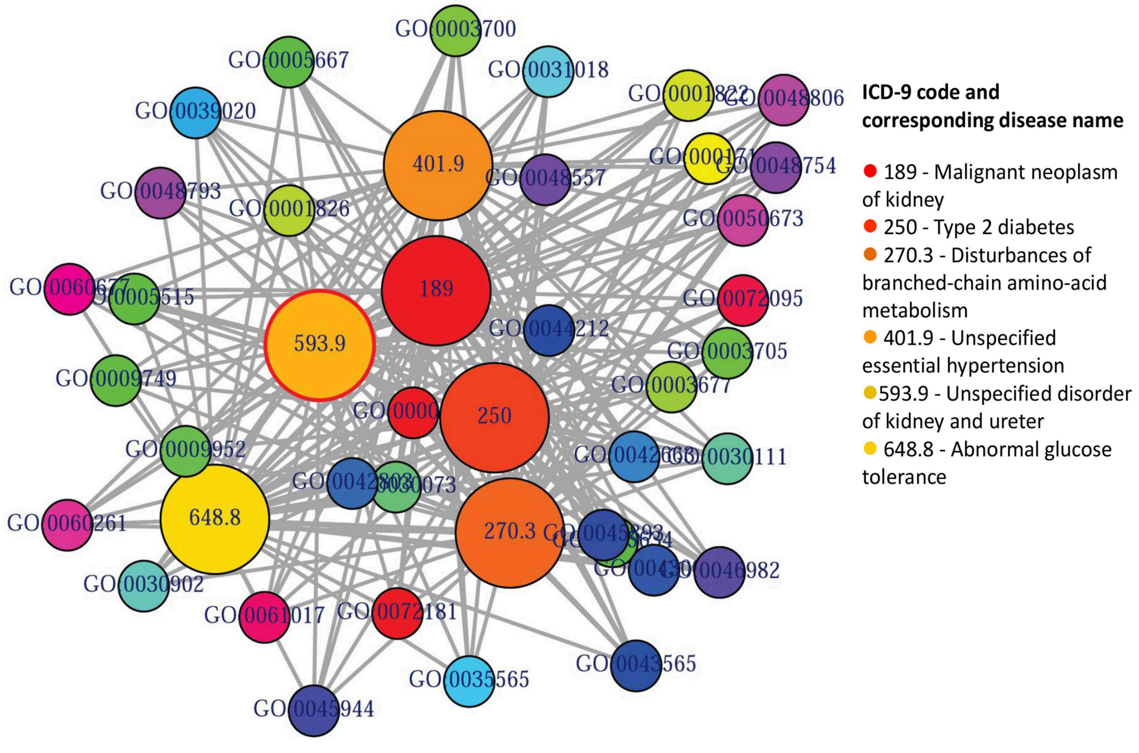
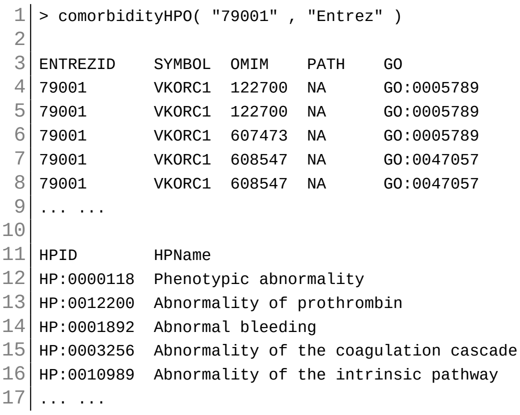


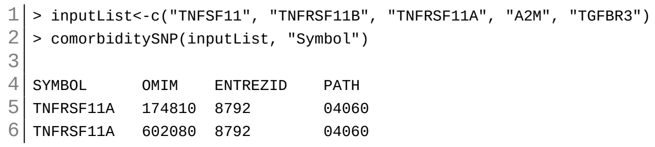
FIGURE 2 | Output figure and statistics of >`comorbidityGO ("189907", "OMIM")`. The OMIM disease id of the “Diabetes mellitus, insulin-dependent” is 189907, which is used as input to the `comorbidityGO`. We show disease comorbidity for the Diabetes mellitus through the GO-disease associations. The size of the nodes represents the degree of associations. ICD-9 codes are used to represent disease categories.



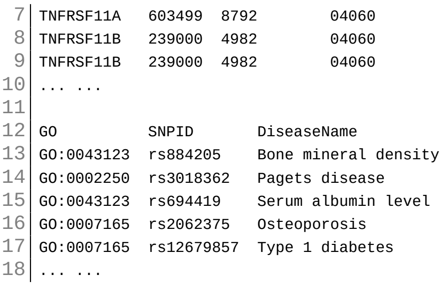
```
> comorbidityHPO( "79001" , "Entrez" )

ENTREZID    SYMBOL  OMIM    PATH    GO
79001       VKORC1  122700  NA      GO:0005789
79001       VKORC1  122700  NA      GO:0005789
79001       VKORC1  607473  NA      GO:0005789
79001       VKORC1  608547  NA      GO:0047057
79001       VKORC1  608547  NA      GO:0047057
... ...

HPID        HPName
HP:0000118  Phenotypic abnormality
HP:0012200  Abnormality of prothrombin
HP:0001892  Abnormal bleeding
HP:0003256  Abnormality of the coagulation cascade
HP:0010989  Abnormality of the intrinsic pathway
... ...
```

Disease–SNPs Association

At present there are only a few databases of genetic variations associated with diseases. Despite the needs for analyzing SNP and disease association, most of the existing databases are based only on functional variants at specific locations on the genome, or deal with only a few genes correlated with disease. There is no integrated resource to widely support genes, SNPs, and disease associated information. Therefore, we integrated data from different databases (dbSNP Sherry et al., 2001, HGVbase Fredman et al., 2002, JSNP Hirakawa et al., 2002, GAD Becker et al., 2004 and OMIM McKusick, 2007) and literature Yang et al., 2008 for studying SNPs-diseases associations. We integrated the information to present the interrelationships among SNPs located in genes, genes associated with diseases, and SNPs associated with diseases. It can aid the understanding of the genes which cause diseases and the impact of SNPs on diseases. For associated information among genetic variation and diseases, we built a database, SNP, which is a combined database of genes, genetic variation and diseases for the utilization in POGO. Two diseases are connected if they share at least one SNP that is statistically significant dysregulated to the disease related gene. Our software is designed to capture the relationships between SNPs associated with disease and disease-causing genes. POGO computes disease-disease association by adopting semantic similarity measures and hypergeometric test. Neighborhood based benchmark method is used to identify the comorbidity pattern among diseases (Goh et al., 2007). We built the associated network as a bipartite graph; each common neighbor node is selected based on the Jaccard coefficient method (Goh et al., 2007). comorbiditySNP function of POGO takes as input any of these three options: a list of gene symbols, a list of Entrez gene ids, SNPs ids or an OMIM id. This function provides disease comorbidity associations and network based on the SNPs-gene-disease associations. comorbiditySNP requires two parameters: id list and id type. An example and its output is given in **Figure 4**.



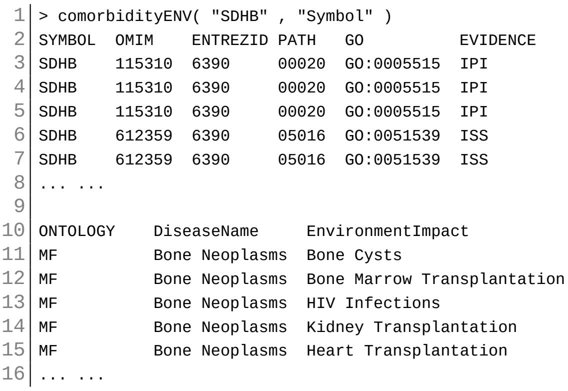
```
> inputList<-c("TNFSF11", "TNFRSF11B", "TNFRSF11A", "A2M", "TGFBR3")
> comorbiditySNP(inputList, "Symbol")

SYMBOL      OMIM    ENTREZID    PATH
TNFRSF11A   174810  8792        04060
TNFRSF11A   602080  8792        04060
```



```
TNFRSF11A   603499  8792        04060
TNFRSF11B   239000  4982        04060
TNFRSF11B   239000  4982        04060
... ...

GO          SNPID       DiseaseName
GO:0043123  rs884205    Bone mineral density
GO:0002250  rs3018362   Pagets disease
GO:0043123  rs694419    Serum albumin level
GO:0007165  rs2062375   Osteoporosis
GO:0007165  rs12679857  Type 1 diabetes
... ...
```

Disease–environment Association

The analysis of environment-disease associations is important to investigate the molecular mechanism of a disease. POGO integrated "etiome," human disease etiological factors database (Liu et al., 2009), and developed a function comorbidityENV to predict the comorbidity risk based on disease environment association (Liu et al., 2009). Integrating genetic, nutritional, behavioral and environmental factors results in the "etiome," which they defined as the comprehensive compendium of disease etiology (Liu et al., 2009). They used natural language processing to look for annotations in articles, and thus creating associations between diseases and environmental information. "etiome" has been developed with the identified 3342 environment related factors that are associated with 3159 complex diseases (Liu et al., 2009). They also identified 1100 genes associated with 1034 diseases from the genetic association studies database GAD (Becker et al., 2004). GAD has 863 diseases information with both genetic and environmental etiological factors. By using all these information, POGO is able to develop comorbidity map by incorporating relations between the diseases themselves as well as relations to environmental factors. This software identifies the disease–disease associations using the associations among environment and their associated diseases. Hypergeometric test is used for extracting associations among environment and diseases; graph topological structure is used to measure the similarity between diseases (Wang et al., 2007). comorbidityENV function takes as input any of the following options: a list of gene symbols, a list of Entrez gene ids or an OMIM id. This function provides disease comorbidity associations and network based on the gene-environment-disease associations. comorbidityENV requires two parameters: id list and id type. An example and its output is given in **Figure 5**.



```
> comorbidityENV( "SDHB" , "Symbol" )
SYMBOL  OMIM    ENTREZID PATH   GO          EVIDENCE
SDHB    115310  6390     00020  GO:0005515  IPI
SDHB    115310  6390     00020  GO:0005515  IPI
SDHB    115310  6390     00020  GO:0005515  IPI
SDHB    612359  6390     05016  GO:0051539  ISS
SDHB    612359  6390     05016  GO:0051539  ISS
... ...

ONTOLOGY    DiseaseName     EnvironmentImpact
MF          Bone Neoplasms  Bone Cysts
MF          Bone Neoplasms  Bone Marrow Transplantation
MF          Bone Neoplasms  HIV Infections
MF          Bone Neoplasms  Kidney Transplantation
MF          Bone Neoplasms  Heart Transplantation
... ...
```

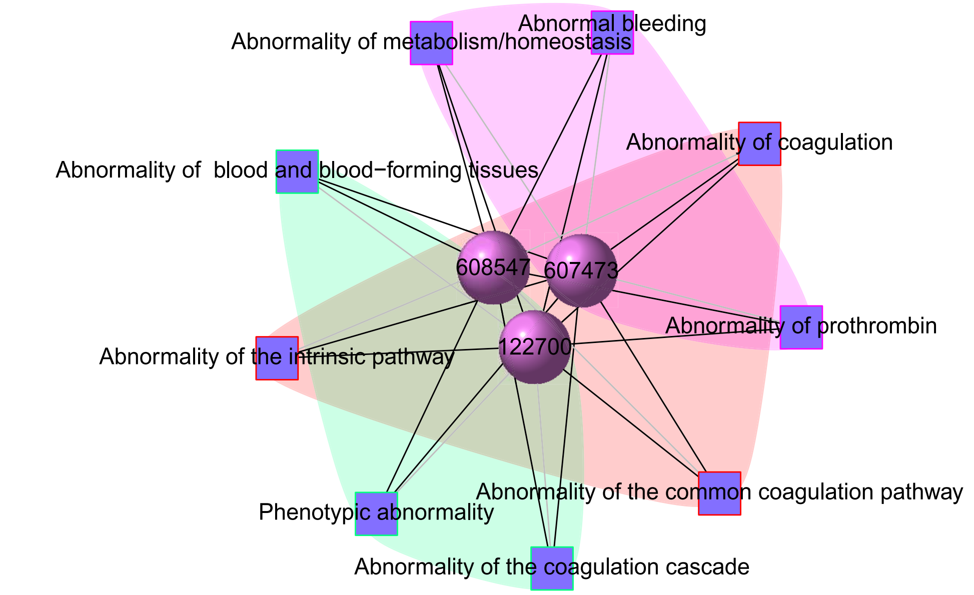

Abnormal bleeding
Abnormality of metabolism/homeostasis
Abnormality of coagulation
Abnormality of blood and blood−forming tissues
608547
607473
Abnormality of prothrombin
Abnormality of the intrinsic pathway
122700
Abnormality of the common coagulation pathway
Phenotypic abnormality
Abnormality of the coagulation cascade

FIGURE 3 | Output figure and statistics of > `comorbidityHPO ("79001", "Entrez")`. The Entrez disease id "79001" is used as input to the `comorbidityHPO`. We show an example of disease comorbidity map for this gene through the phenotype-disease associations. Here the square nodes represent the phenotypes and spheres represent OMIM disease ids.

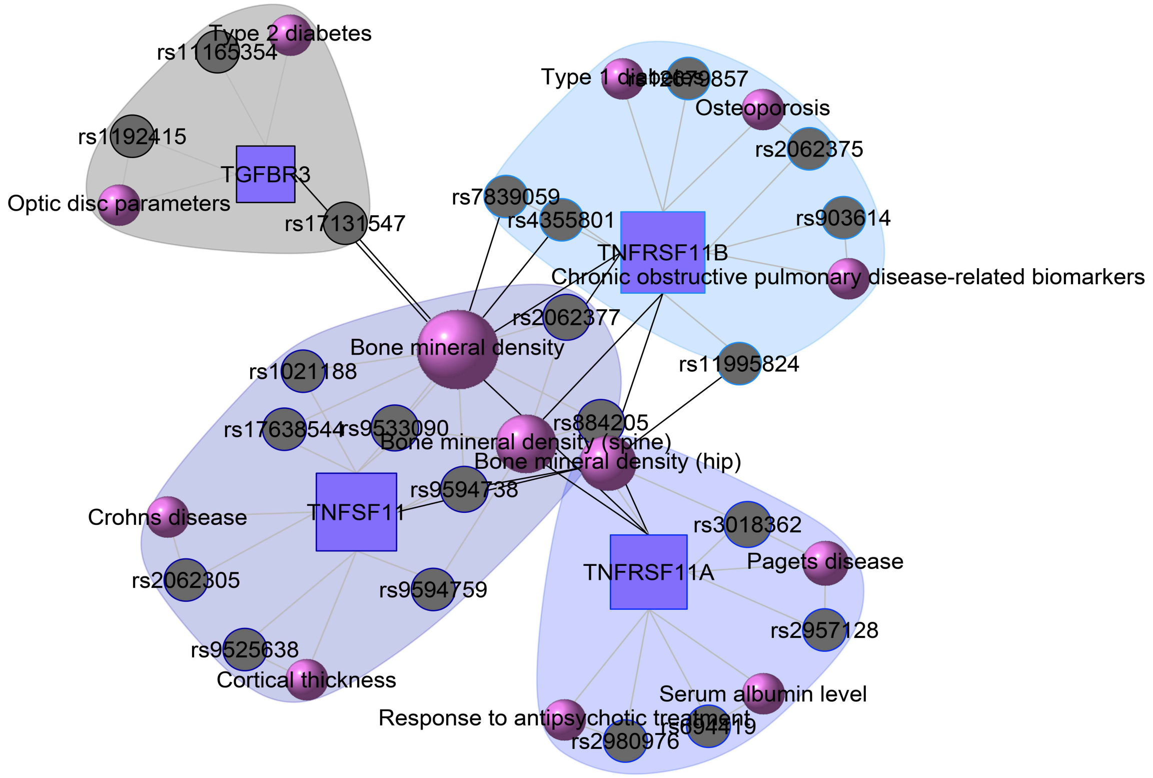


FIGURE 4 | Output figure and statistics of >`comorbiditySNP (c("TNFSF11", "TNFRSF11B", "TNFRSF11A", "A2M", "TGFBR3"), "Symbol")`. We show an example of disease comorbidity through the SNPs-gene-disease associations. Here the square nodes represent the genes symbols, circles represent SNPs ids, and spheres represent diseases names. The size of the nodes represents the degree of associations.

miRNA–disease Association

MicroRNA (miRNA) performs its regulatory function through its target genes. Two diseases are connected if they share at least one gene and/or one miRNA that is statistically significant dysregulated (Goh et al., 2007). miRNAs with similar functions tend to be associated with diseases with similar phenotypes, and vice versa (Lu et al., 2008). Based on these hypothesis, we used a framework to identify miRNA-disease associations through the direct identified association from the miRNA-disease association database and indirect association from the combined database of miRNA-target and gene-disease associations. `POGO` makes use of microRNA-target databases, miR2Disease (Jiang et al., 2009), HMDD (Li et al., 2014), and gene-disease association databases, OMIM (McKusick, 2007), to explore the mRNA and miRNA association between diseases. We filtered out invalid miRNA-disease associations with incorrect disease names or miRNA names. We used National Library of Medicine[2] to obtain the correct disease names. We used miRBase to get the correct miRNA names (Kozomara and Griffiths-Jones, 2011). For a miRNA-disease pair, firstly, `POGO` maps the causal genes of the disease. It uses a *p*-value to measure the significance of the association between the miRNA and the disease. OMIM diseases ids are mapped with ICD-9-CM codes based on the literature (Park et al., 2009). Neighborhood based benchmark method is used to identify the comorbidity pattern among diseases. We build the associated network as a bipartite graph; each common neighbor node is selected based on the Jaccard coefficient method (Goh et al., 2007). `comorbiditymiRNA` function of `POGO` takes as input any of the following options: a list of gene/miRNA symbols, a list of Entrez gene ids, an ICD-9 code, an GO id or an OMIM id. This function provides disease comorbidity associations and network based on the disease-miRNA associations. `comorbiditymiRNA` requires two parameters: id list and id type. An example and its output is given in **Figure 6**.

```
> comorbiditymiRNA( "TNFRSF11A" , "Symbol" )

ENTREZID miRNAID        DiseaseName                          SYMBOL
8792     hsa-miR-432    Duchenne muscular dystrophy (DMD)    TNFRSF11A
8792     hsa-miR-324-3p primary biliary cirrhosis (PBC)      TNFRSF11A
8792     hsa-miR-324-3p lupus nephritis                      TNFRSF11A
8792     hsa-miR-432    miyoshi myopathy (MM)                TNFRSF11A
8792     hsa-miR-664    multiple sclerosis                   TNFRSF11A
8792     hsa-miR-432    nemaline myopathy (NM)               TNFRSF11A
... ...

GO          EVIDENCE    ONTOLOGY    OMIM    PATH
GO:0002250  IMP         BP          174810  5323
GO:0002250  IMP         BP          174810  5323
GO:0009897  IDA         CC          174810  4060
GO:0009897  IDA         CC          602080  5323
GO:0002250  IMP         BP          602080  4380
GO:0002250  IMP         BP          612301  5323
... ...
```

CNV–disease Association

Copy number variants are hypothesized to cause diseases through several mechanisms. Sometimes, the combination of two or more copy number variants can produce a complex disease. Additionally, complex diseases might occur when copy number variants are combined with other genetic and environmental factors (McCarroll and Altshuler, 2007). Diseases might be caused by copy number variants due to both additional copies of sequence (duplications) and losses of genetic material (deletions). We used Database Genomic Variants (DGV[3]) database and developed a function `comorbidityCNV` to predict the comorbidity risk based on CNVs-disease association (MacDonald et al., 2014). `POGO` makes use of DGV and OMIM (McKusick, 2007) to explore the genetic association between diseases. Two diseases are connected if they share similar copy number variations. OMIM diseases ids are mapped with ICD-9-CM codes based on the literature (Park et al., 2009). Neighborhood based benchmark method is used to identify the comorbidity pattern among diseases (Goh et al., 2007). We build the associated network as a bipartite graph; each common neighbor node is selected based on the Jaccard coefficient method (Goh et al., 2007). `comorbidityCNV` function of `POGO` takes as input any of the following options: a list of gene symbols, a list of Entrez gene ids or an OMIM id. This function provides disease comorbidity associations and network based on the disease-CNV associations. `comorbidityCNV` requires two parameters: id list and id type. An example and its output is given in **Figure 7**.

```
> comorbidityCNV("602228", "OMIM")

SYMBOL    OMIM ENTREZID  PATH   GO          EVIDENCE
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
... ...

ONTOLOGY  CNV.ID   Chr  Start       End         VarSubtype
MF       nsv7211    10  108617417   118351740   Inversion
MF       nsv7553    10  114845707   114890646   Loss
MF    esv2074123    10  114876971   114877374   Deletion
MF      nsv24033    10  114877162   114877217   Loss
MF     nsv527837    10  114888608   114911079   Loss
... ...
```

Integrated Comorbidity Prediction Using Multiplex

As a single source of genomic data is prone to bias, incompleteness and noise, integration of different genomic data sources is designed to accomplish reliable disease comorbidities prediction. Systematic integration and comparison of multiple layers of information is required to provide deeper insights into biological systems. We incorporated a multiplex network model into `POGO` to integrate multiple omics, environmental and phenotypic information. To leverage the potential of multi-omics studies, exploratory data analysis methods that provide systematic integration and comparison of multiple layers of omics information are required. We applied our multiplex method of integrating different types of data

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

[3] http://dgv.tcag.ca/.

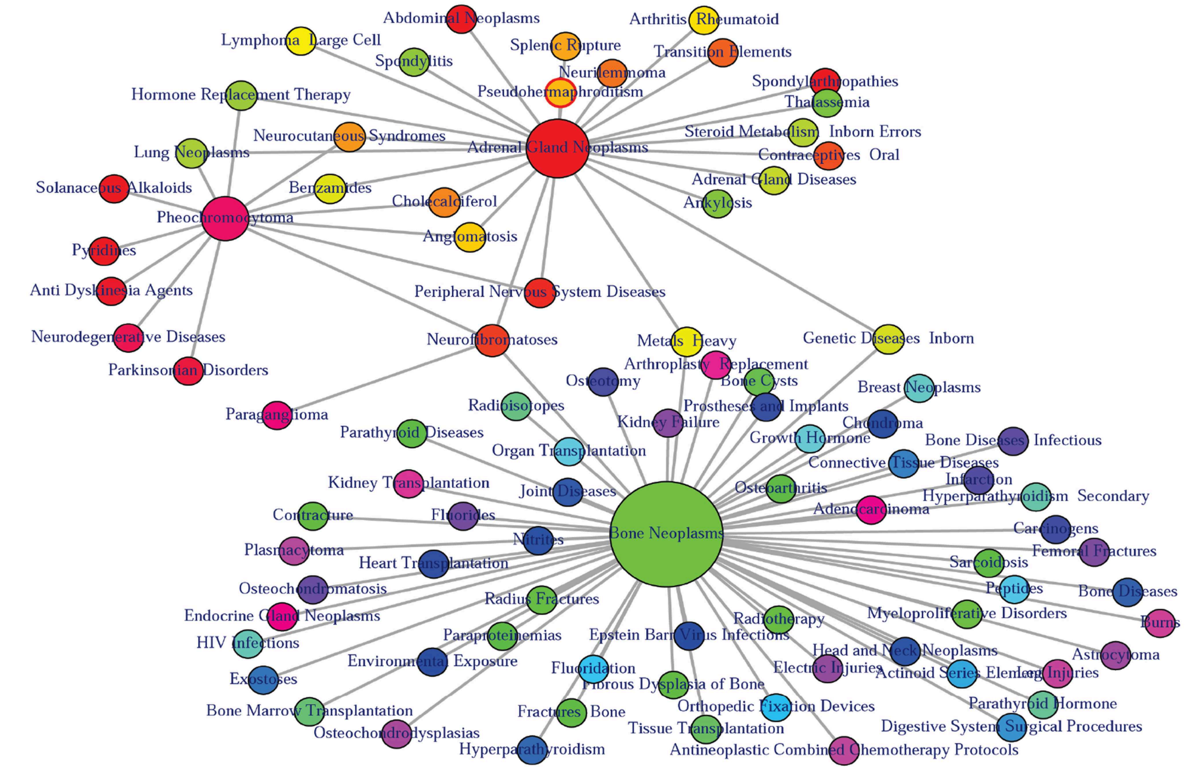


FIGURE 5 | Output figure and statistics of `>comorbidityENV("SDHB", "Symbol")`. The gene symbol “SDHB” is used as input to the `comorbidityENV`. We show disease comorbidity map for this gene input through the disease-environmental associations. The size of the nodes represents the degree of associations.

by modeling similarities between diseases in a multiplex network. The multiplex network allows us to model diseases by representing each data type as a layer in the multiplex. Importantly, this allows us to capture the interactions between the various types of data, such as the interdependence of mRNA expression and signaling pathways with clinical information of the disease comorbidities. We developed a function `comorbidityMultiplex` to predict the integrated comorbidity risk. `comorbidityMultiplex` function takes as input any number of layers information. This function provides integrated comorbidity associations and network. As an example of integrating with this function we considered three different types of data for three layers of our multiplex network: mRNA-disease, pathway-disease and clinical association information. An example and its output is given in **Figure 8**.

In this example, we considered association information of 10 diseases, which are the output of other functions of `POGO`. The ICD-9 code of the 10 diseases are 155, 157, 199, 286, 287, 571, 572, 574, 576, and 782. `POGO` identified disease-disease comorbidity associations network based on the gene-disease association and pathway-disease association, which are shown in **Figures 8A,B** respectively. It is notable that there is no shared pathway for the disease 572 with the 9 other diseases. The comorbidity network based on the clinical information is shown in **Figure 8C**. We used all these three association networks for the input of our multiplex network (see **Supplementary Tables S1–S3**). In this case, the multiplex network is comprised of three layers, each with 10 nodes. In each layer, each node has a weighted undirected edge connecting it to every other node in the same layer. In addition, each disease is connected to itself in every other layer

```
> input = c("S1.txt", "S2.txt", "S3.txt")
> sv<-c(1, 1, .5) #strength value of each layer
> comorbidityMultiplex(input, sv)
... ...
$aggG
        ICD.155 ICD.157 ICD.199 ICD.286 ICD.287 ICD.571 ICD.572 ICD.574
ICD.155     0.0    20.0    14.0    12.0     9.5    12.0    22.5     2.0
ICD.157    20.0     0.0    12.5     5.5    14.0     3.0     5.0     2.5
ICD.199    14.0    12.5     0.0     2.0     7.5     2.0     3.0     1.5
ICD.286    12.0     5.5     2.0     0.0    16.5     4.5     7.5     1.5
ICD.287     9.5    14.0     7.5    16.5     0.0     6.5     7.5     1.5
ICD.571    12.0     3.0     2.0     4.5     6.5     0.0    27.5     2.5
ICD.572    22.5     5.0     3.0     7.5     7.5    27.5     0.0     2.5
ICD.574     2.0     2.5     1.5     1.5     1.5     2.5     2.5     0.0
... ...
```

by the strength of interaction between the data types. So the multiplex network created using POGO is formed of three layers using the mRNA, pathway and clinical data. Each layer provided information on the same diseases. This result is a 30 × 30 multiplex matrix, since a multiplex matrix is formed of $n \times h$ rows and columns where n is the number of patients and h is the number of layers. Our software POGO can find the disease comorbidities by integrating all the descriptive layers, taking into account the properties of the multiplex. All these three categories association data are used as input of our multiplex network and predicted the integrated disease comorbidities network as shown in the **Figure 8D**.

Comorbidity Mapping

Patient medical records contain important clarification regarding the co-occurrences of diseases affecting the same patient. Two diseases are connected if they are co-expressed in a significant number of patients in a population (Hidalgo et al., 2009). To estimate the correlation starting from disease co-occurrence, we need to quantify the strength of the comorbidity risk. We used two comorbidity measures to quantify the strength of comorbidity associations between two diseases: (i) the Relative Risk (fraction between the number of patients diagnosed with both diseases and random expectation based on disease prevalence) as the quantified measures of comorbidity tendency of two disease pairs; and (ii) ϕ-correlation (Pearsons correlation for binary variables) to measure the robustness of the comorbidity association (Moni and Lio, 2014). We used the relative risk RR_{ij} and ϕ-correlation ϕ_{ij} of observing a pair of diseases i and j affecting the same patient. The RR_{ij} allows us to quantify the co-occurrence of disease pairs compared with the random expectation. When two diseases co-occur more frequently than expected by chance, we will get $RR_{ij} > 1$ and $\phi_{ij} > 0$. The two comorbidity measures are not completely independent of each other. We included links between disease pairs for which the co-occurrence is notably greater than the random expectation based on population prevalence of the diseases. Clinical information is from the http://www.icd9data.com in the ICD-9-CM format and collected from Hidalgo et al. (2009). The function comorbidityMap of POGO package is able to take input an OMIM id/3 or 5 digit ICD-9-CM code of a disease or a list of gene symbols/Entrez ids and provides comorbidity map of the patient based on the relative risk and ϕ-correlation. comorbidityMap requires two parameters: id list and id type. An example and its output is given in **Figure 9**.

Methods

Diseases are connected when they share at least one significant dysregulated gene/miRNA/SNP/CNV/GO/phenotype or environmental factor. Let a specific set of associated diseases D and a set of significant biomarker genes G, gene-disease associations attempt to find whether gene $g \in G$ is associated with disease $d \in D$. If G_i and G_j are the sets of significant up and down dysregulated genes associated with diseases i and j respectively then the number of shared dysregulated genes (n_{ij}^{g}) associated with both diseases i and j is as follows:

$$n_{ij}^{g} = N(G_i \cap G_j) \tag{1}$$

We calculated the similarity between a pair of diseases based on the number of entities (gene, SNP, CNV, miRNA, HPO or environmental factor) that shared between them. For an instance, in case of gene-disease association, we generated a list of genes known to be associated with each disease, and the disease similarity (association) was calculated based on how many genes are shared between a pair of diseases. The similarity is defined as

$$Sim(i,j) = \frac{N(G_i \cap G_j)}{\sqrt{N(G_i)} * \sqrt{N(G_j)}}, \tag{2}$$

where $N(G_i)$ and $N(G_j)$ are the number of genes linked to disease i and j respectively, and $N(G_i \cap G_j)$ is the number of genes associated to both disease i and j. SNP-sharing, CNV-sharing, miRNA-sharing, HPO-sharing and environmental factors were also generated with the same approach used for gene-sharing.

Hypergeometric test is implemented for enrichment analysis (Subramanian et al., 2005). It is used to assess whether the number of selected genes or ontology associated with disease is larger than expected. To determine whether any disease annotate a specified list of genes at frequency greater than what would be expected by chance, POGO calculates a p-value using the hypergeometric distribution. Significance of the enrichment analysis is assessed by the hypergeometric test and the p-value is adjusted by false discovery rate (FDR). The hypergeometric p-value is calculated using the following formula:

```
> comorbidityMap("042", "ICD9")
ICD.9.D1 ICD.9.D2 Prevalence.D1 Prevalence.D2 Co.occurrenceD1D2 RRij
"011"    "018"    16646         639           110               134.842507
"011"    "031"    16646         3693          807               171.170619
"011"    "042"    16646         1067          64                 46.984060
"011"    "112"    16646         141325        752                 4.168058
"011"    "117"    16646         9094          179                15.418178
 ... ...

 CI1          CI2         phi        t
131.740584  138.0174686 0.0334998   12.600646
170.628511  171.7144495 0.1024054   38.700702
 45.141791   48.9015140 0.0148728    5.591768
  4.153894    4.1822713 0.0118565    4.457522
 15.199244   15.6402660 0.0136184    5.120042
 ... ...
```

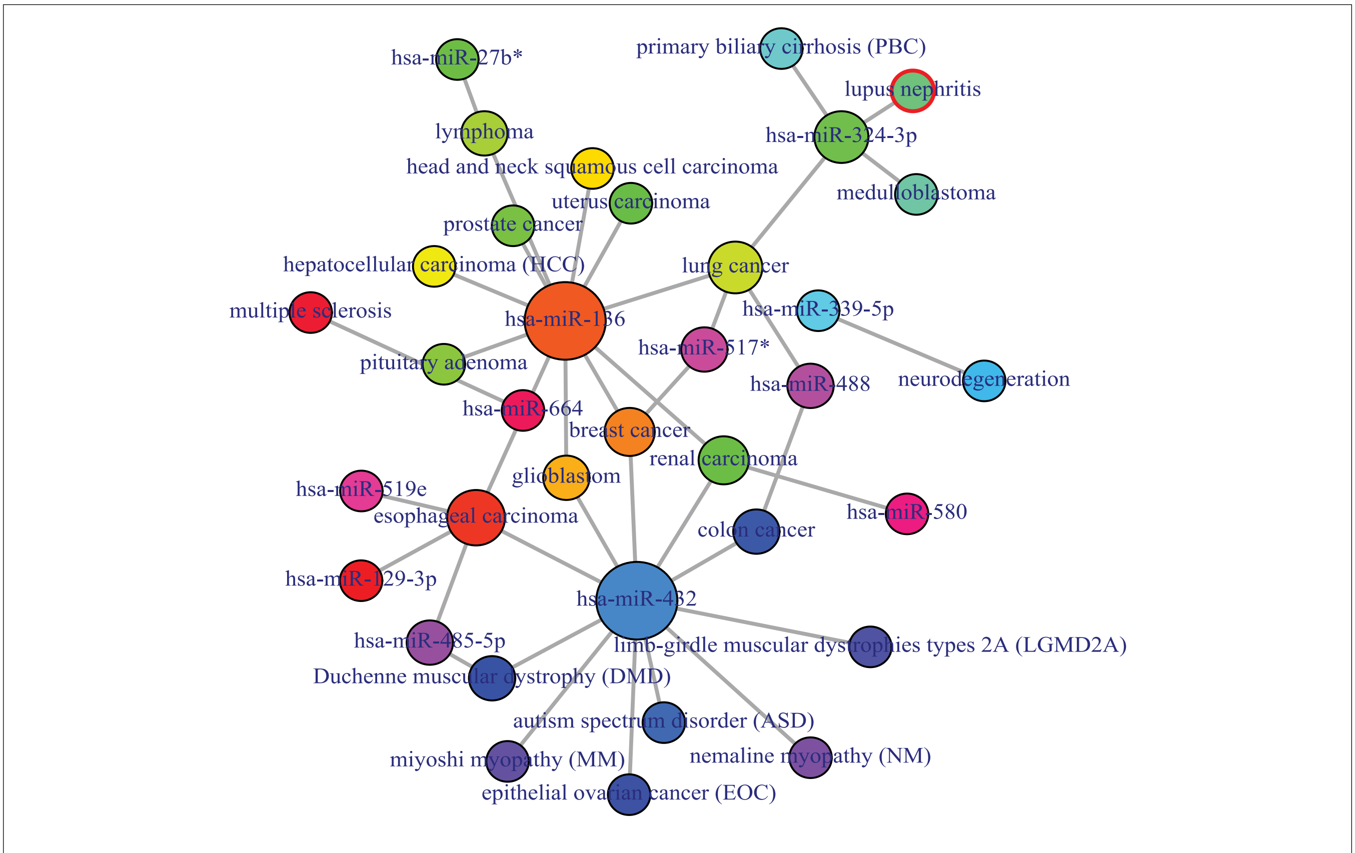


FIGURE 6 | Output figure and statistics of `> comorbiditymiRNA( "TNFRSF11A", "Symbol" )`. The gene Symbol TNFRSF11A is used as input to the `comorbiditymiRNA`. We show the comorbidities originated using the miRNA-disease associations information. The size of the nodes represents the degree of associations.

$$p - value = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \tag{3}$$

where N is the total number of reference genes, M is the number of genes that are associated to the disease of interest, n is the size of the list of genes of interest and k is the number of genes within that list which are associated to the disease. In case of GO term the p-value reports the likelihood of finding n genes annotated with a particular GO term in the set of interest by chance alone, given the number of genes annotated with that GO terms in the reference set. A biological process, molecular function or cellular location which are represented by a GO term is called enriched if the p-value is less than 0.05.

The co-occurrence indicates the number of common miRNAs/genes/ontology/SNPs/CNVs between two diseases. We applied the Jaccard index or Jaccard similarity coefficient, which is known as a standard method for comparing the similarity between two sets of entities. Each common neighbor is calculated based on the Jaccard Index method to calculate the strength of co-occurrence, where association score for a node pair is as:

$$Ass_{i,j} = \frac{N(G_i \cap G_j)}{N(G_i \cup G_j)} \tag{4}$$

We improved the performance of the association scores based on the Adamic and Adar measure (Adamic and Adar, 2003), which weights the impact of neighbor disease nodes inversely with respect to their total number of connections as follows:

$$AssScore(i,j) = \sum_{n \in N(G_i \cap G_j)} \frac{1}{log(degree(n))} \tag{5}$$

This inverse frequency technique is based on the principle that rare relationships are more specific and have more impact on the disease association.

Finally POGO calculates disease-disease interaction score. The score refers to the strength of the interaction between the diseases based on the protein interaction. The interaction score (ϕ_{ij}) is assigned for each disease pair i and j as follows:

$$\phi_{ij} = log(n_{ij}^g * N + Z) - \log(NG_i * NG_j + Z) \tag{6}$$

Here, NG_i and NG_j are the total number of genes for the disease, i and j, respectively. n_{ij}^g is the total number of common genes between the two diseases. N is the size of entire proteins involved in the disease protein network. Z is a constant ($Z = 1$) introduced to avoid out-of bound errors, if $NG_i = NG_j = n_{ij}^g = 0$.

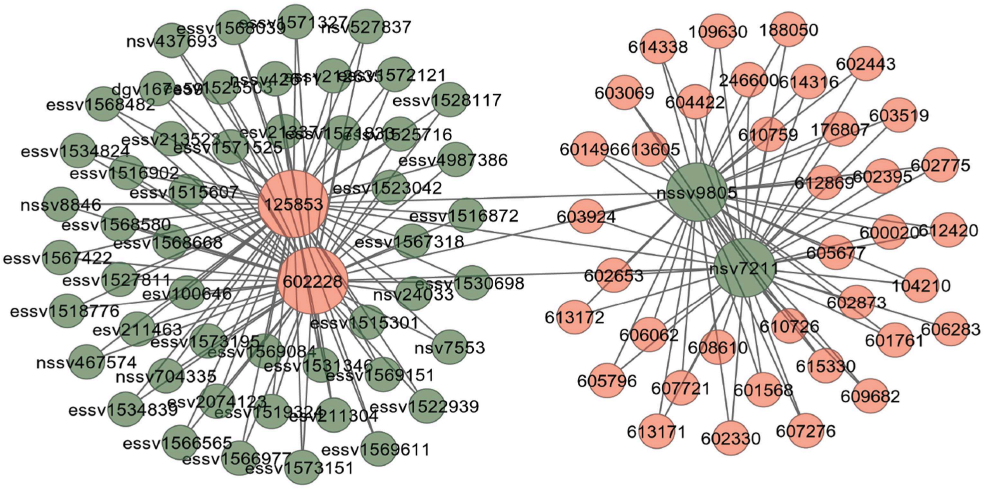
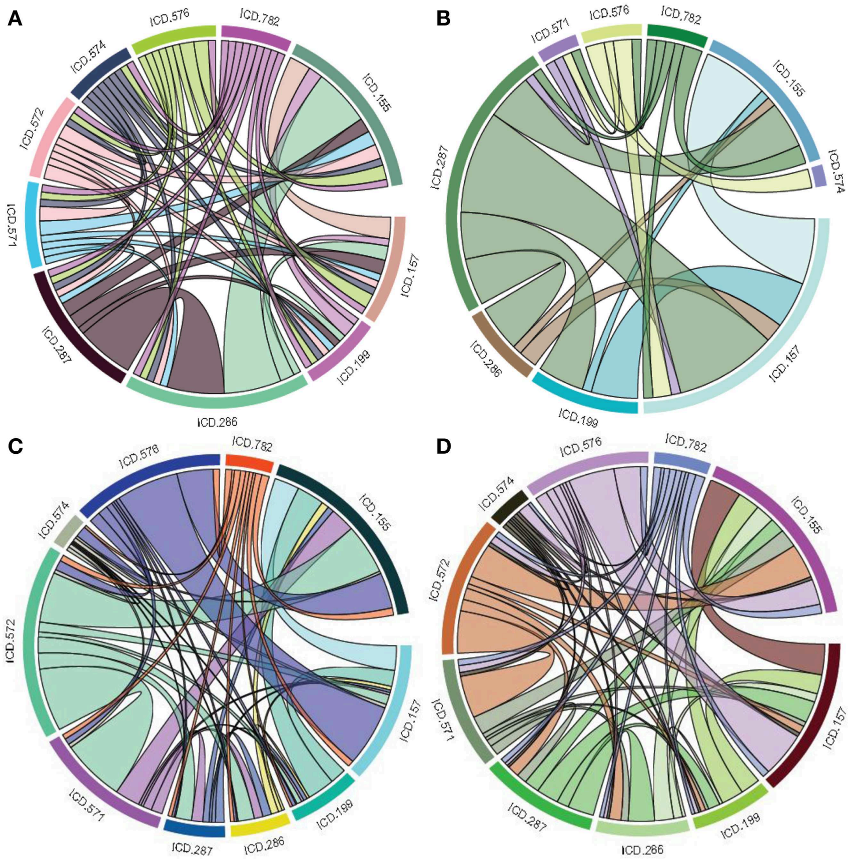


FIGURE 7 | Output figure and statistics of > `comorbidityCNV ("602228", "OMIM")`. The OMIM disease id of the "Type 2 Diabetes mellitus" is 602228, which is used as input to the `comorbidityCNV`. We show disease comorbidity for the "Type 2 Diabetes mellitus" through the CNVs-disease associations. Here the light red color nodes represent the OMIM disease ids and light green color nodes represent the CNVs ids. The size of the nodes represents the degree of associations.



ICD-9 code and corresponding disease name

155 - Malignant neoplasm of liver

157 - Malignant neoplasm of pancreas

199 - Malignant neoplasm without specification of site

286 - Coagulation defects

287 - Purpura and other hemorrhagic conditions

571 - Chronic liver disease and cirrhosis

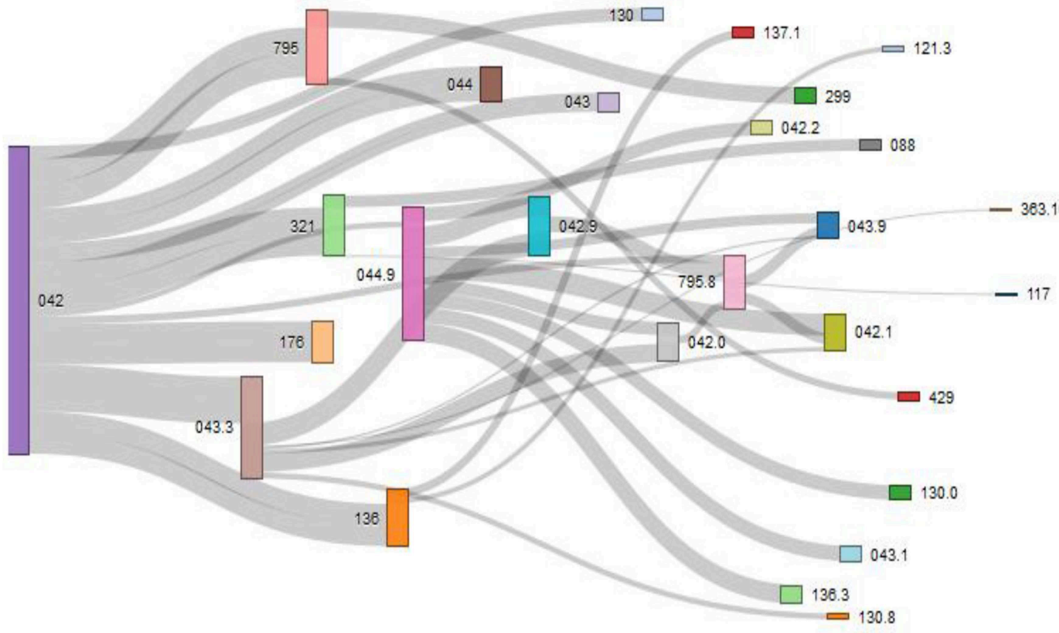
572 - Liver abscess and sequelae of chronic liver disease

574 - Cholelithiasis

576 - Disorders of biliary tract

782 - Symptoms involving skin and other integumentary tissue

FIGURE 8 | Disease comorbidities network are constructed by applying the multiplex network model. Each disease is denoted by the ICD-9-CM code. **(A)** is a comorbidity association network based on the gene disease association data. **(B)** is a comorbidity association network based on the pathway disease association data. **(C)** is a comorbidity association network based on the clinical information. **(D)** is a comorbidity association network based on the integrated multiplex network output of the input of **(A–C)** as layers of the model.



ICD-9 code and corresponding disease name

042 - HIV infection

042.0 - HIV with specified infections

042.2 – HIV with specified malignant neoplasm

042.9 - Acquired immunodeficiency syndrome, unspecified

043 - HTLV-III/LAV infection

043.1 - HTLV-III/LAV infection causing specified diseases of the central nervous system

043.3 - HTLV-III/LAV infection causing other specified conditions,

043.9 Acquired immunodeficiency syndrome-related complex with or without other conditions

044 - Other HTLV-III/LAV conditions

044.9 - HTLV-III/LAV infection, not otherwise specified

088 - Arthropod-borne diseases

117 – Mycoses

121.3 - Fascioliasis

130 – Toxoplasmosis

130.0 - Meningoencephalitis due to toxoplasmosis

130.8 - Multisystemic disseminated toxoplasmosis

136 - Unspecified infectious & parasitic diseases

136.3 – Pneumocystosis

137.1 - Late effects of central nervous system tuberculosis

176 - Kaposi's sarcoma

299 - Pervasive developmental disorders

321 - Type 2 diabetes mellitus

363.10 - Disseminated chorioretinitis

429 - Ill-defined descriptions and complications of heart disease

795 - Nonspecific abnormal cytological, histological, immunological, and dna test findings

795.8 - Abnormal tumour markers

FIGURE 9 | Output figure and statistics of `>comorbidityMap ("042", "ICD9")`. The icd-9-CM code of the HIV is 042, which is used as input to the `comorbidityMap`. We show disease comorbidity for the HIV infection (042) with other diseases, whose ICD-9-CM codes are 042.0 (with specified infections), 042.1 (causing other specified infections), 042.2 (with specified malignant neoplasms), 042.9 (acquired immunodeficiency syndrome, unspecified), 043 (HTLV-III/LAV infection), 043.1 (HTLV-III/LAV infection causing specified diseases of the central nervous system), 043.3 (HTLV-III/LAV infection causing other specified conditions), 043.9 (acquired immunodeficiency syndrome-related complex with or without other conditions), 044 (Other HTLV-III/LAV conditions), 044.9 (HTLV-III/LAV infection, not otherwise specified), 088 (arthropod-borne diseases), 117 (mycoses), 121.3 (fascioliasis), 130 (toxoplasmosis), 130.0 (meningoencephalitis due to toxoplasmosis), 130.8 (multisystemic disseminated toxoplasmosis), 136 (unspecified infectious and parasitic diseases), 136.3 (pneumocystosis), 137.1 (late effects of central nervous system tuberculosis), 176 (Kaposi's sarcoma), 299 (pervasive developmental disorders), 321 (type 2 diabetes mellitus), 363.10 (disseminated chorioretinitis), 429 (ill-defined descriptions and complications of heart disease), 795 (nonspecific abnormal cytological, histological, immunological, and dna test findings), and 795.8 (abnormal tumor markers). POGO uses color rectangle to classify different disease codes and the size of the rectangle is used to represent the severity of that disease.

The expected result of ϕ_{ij} is positive, when the disease pair is over-represented and negative, when the disease pair is under-represented. Co-occurrence also indicates the number of shared patients. So, we used weighting scheme to avoid the bias based on disease prevalence. The mutual information weight $W(d_i, d_j)$ between two diseases d_i and d_j is defined as

$$W(d_i, d_j) = log\left(\frac{p(d_i, d_j)}{p(d_i) * p(d_j)}\right) \tag{7}$$

where the numerator is the observed co-occurrence (joint probability) and the denominator is the random expectation of co-occurrence (product of marginal probabilities).

The use of semantic similarity between biological processes to estimate disease association could enhance the identification and characterization of disease association besides identifying novel biological processes involved in the diseases. Graph-based methods using the topology of GO graph structure is used to compute semantic similarity. We adapted the approach for computing the functional similarity of GO terms from Wang et al. (2007, 2010). Semantic values of GO term are measured according to the DAG of corresponding disorders. Semantic similarity for any pair of GO term is calculated based on disease semantic value. Formally, a GO term a can be represented as a graph $DAG_a = (a, T_a, E_a)$, where T_a is the set of all GO terms in DAG_a, including term a itself and all of its ancestor terms in the GO graph, and E_a is the set of corresponding edges that connect the GO terms in DAG_a. To encode the semantic of a GO term in a measurable format to enable a quantitative comparison, Wang firstly defined the semantic value of term a as the combined contribution of all terms in DAG_a to the semantics of term a (Wang et al., 2007). Terms closer to term a in DAG_a contribute more to its semantics (Wang et al., 2010). Thus, the contribution of a GO term t in DAG_a is defined to the semantics of GO term a as the S value of the term t related to term a, $S_a(t)$, which can be calculated as:

$$S_a(t) = \begin{cases} S_a(a) = 1 \; if \, t = a \\ S_a(t) = max\{w_e * S_a(t') | t' \in \; children \; of \, (t)\} \; if \, t \neq a \end{cases} \tag{8}$$

where w_e is the semantic contribution factor for edge e ($e \in E_a$) linking term t with its child term $t^{'}$. It is assigned between 0 and 1 according to the types of associations. Term a contributes to its own is defined as one. Then the semantic value of GO term a, $SV(a)$ and the semantic value of GO term b, $SV(b)$ are calculated as:

$$SV(a) = \sum_{t \in T_a} S_a(t), \quad SV(b) = \sum_{t \in T_b} S_b(t) \tag{9}$$

Thus, for the given two GO terms a and b, the semantic similarity between these two terms is defined as:

$$S_{sim}(a, b) = \sum_{t \in T_a \cap T_b} \frac{S_a(t) + S_b(t)}{SV(a) + SV(b)} \tag{10}$$

where $S_a(t)$ is the semantic value of term t related to GO term a and $S_b(t)$ is the semantic value of GO term t associated to GO term b. The semantic similarity between two sets of GO terms A and B is calculated as

$$Sim(A, B) = \frac{1}{|A| + |B|}\left(\sum_{a \in A} Sim(a, B) + \sum_{b \in B} Sim(b, A)\right) \tag{11}$$

where $|A|$ and $|B|$ represent the numbers of terms in sets A and B respectively.

To obtain more insight into the shared risk factors mechanism of associated human genetic diseases, mapping was implemented from disease phenotype to gene based on the disease-gene association. With the integration of huge numbers and diverse set of experimental data, prediction of gene-phenotype interactions has emerged as a very productive subfield with great importance for the understanding of human disease. Given a specific set of human phenotype D, a set of human genes G and evidence E, these approach attempt to find whether gene $g \in G$ is associated with phenotype $d \in D$. It is notable that E could be gene-disease associations obtained through genetic studies. To quantitatively explore the phenotypic similarity between different phenotype records P_i and P_j, according to Zhang et al. (2010) we defined the association measure as cosine of the angle between their corresponding phenotype feature vectors using the following formula:

$$Sim(P_i, P_j) = \frac{\sum_{k=1}^{N} w_{k,i} * w_{k,j}}{\sqrt{\sum_{k=1}^{N}(w_{k,i})^2} * \sqrt{\sum_{k=1}^{N}(w_{k,j})^2}} \tag{12}$$

where N is the total mapping concepts, $w_{k,i}$ and $w_{k,j}$ were the k-th term, weight in phenotype record P_i and P_j, respectively.

For each of the phenotype clusters, mapping was implemented from disease phenotypes to their associated disease genes based on the disease-gene association list in the GAD and OMIM databases. Therefore, we can get the corresponding gene subsets mapped to different phenotype clusters. OMIM disease ids were mapped to the hierarchy of HPO to retrieve the matched HPO terms. Then, a new HPO similarity is calculated for each pair of phenotypes by Jaccard similarity Index

$$Sim_{HPO} = \frac{|P1 \cap P2|}{|P1 \cup P2|} \tag{13}$$

where $P1$ and $P2$ are the set of the matched HPO terms of the two phenotypes, respectively.

The way to assign terms to objects is to add annotations. In our case, the entities represent genes and terms corresponding to phenotypes (HPO terms) or biological processes (GO terms). The specificity of the terms associated with genes allows us to calculate the most significant relationships between them, which use to be related to its proximity to the root.

Each disease is generally mapped to multiple phenotypic features. In order to compute associations between two diseases, $d1$ and $d2$, we adapt a method previously developed for estimating protein similarity with GO (Pesquita et al., 2008), where each feature of $d1$ is matched with the most similar feature of $d2$ and the average is taken over all such pairs of features:

$$sim(d1 \rightarrow d2) = avg\left[\sum_{s \in d1} \max_{t \in d2} sim(s, t)\right] \qquad (14)$$

Equation (14) is not symmetric with respect to $d1$ and $d2$, the final similarity metric is defined as the mean of Equation (14) taken in both orientations:

$$sim(d1, d2) = \frac{1}{2} * sim(d1 \rightarrow d2) + \frac{1}{2} * sim(d2 \rightarrow d1) \qquad (15)$$

This metric is used to indicate the similarity between two disorders, each of which is mapped to multiple HPO terms.

Multiplex Network Model for Data Integration

We developed multiplex network model to integrate diverse set of omics and clinical data to predict disease comorbidities. It is a special type of multilayered network which is called the multiplex network, in which the same nodes are present in all layers, i.e., $V_1 = V_2 = \ldots\ldots = V_M = V$ and where nodes can only have interlayer connections to their counterpart nodes, i.e., $E_{\alpha\beta} = (v, v); v \in V$ for all $\alpha, \beta \in 1, \ldots, M, \alpha \neq \beta$ (Boccaletti et al., 2014).

Let's consider that we have a set of associated diseases. Each pair of diseases has different types of associated data describing them in some way. In each data type, diseases have some level of association to each other and each data type has a level of dependency or interaction. Each layer in the multiplex represents a particular type of data with each node representing a disease in each layer of the multiplex. The edges between nodes in each layer represent a measure of association between diseases in corresponding to the level of similarity between diseases for the particular data type which the layer represents. The strength of interaction between each data type can be modeled by a weight connecting each layer in the multiplex. **Figure 10** shows an example with three layers (data types) and four diseases. In this case we can model the association among diseases in a multiplex network that can be represented in a matrix as follows:

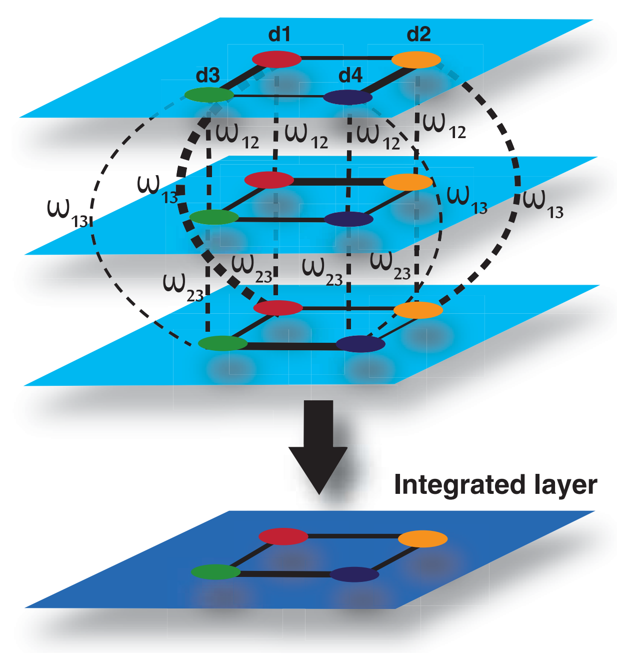


FIGURE 10 | Multiplex formed by three input layers, each representing a data type, and four nodes, each representing a disease. The 4th layer is an output layer, which is an integrated layer of the 3 input layers.

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A}_1 & \omega_{12}\boldsymbol{I} & \ldots & \omega_{1h}\boldsymbol{I} \\ \omega_{21}\boldsymbol{I} & \boldsymbol{A}_2 & \ldots & \omega_{2h}\boldsymbol{I} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{h1}\boldsymbol{I} & \omega_{h2}\boldsymbol{I} & \ldots & \boldsymbol{A}_h \end{pmatrix}, \qquad (16)$$

where h is the number of layers, A_i is the adjacency matrix of layer i, ω_{ij} is the interlayer interaction strength from layer i to j and I is the corresponding identity matrix. The strength between layers in the multiplex, ω, represents a measure of dependency or strength of interaction between the layers. The edge weights between nodes represent a measure of similarity between nodes in the same layer, normalized between zero and one. Therefore, it is natural for the values of ω to represent a measure of dependence between zero and one, where zero and one indicate independence and total dependence between the layers respectively. In our case the strength of interaction is undirected and symmetric, i.e., $\omega_{i,j} = \omega_{j,i}$.

To compute an overall disease similarity between patients given all sets of data, we can find the disease similarity by aggregating the descriptive layers in some way, taking into account the properties of the multiplex. Estrada and Gómez-Gardeñes (2014) defined the aggregate network, $\hat{G}$, of a multiplex network as follows. Let $G_1 = (V_1, E_1)$, $G_2 = (V_1, E_2)$, ..., $G_h = (V_1, E_h)$ be the set of layers in the multiplex. Then $\hat{G} = (\hat{V}, \hat{E})$ where $\hat{V} = V_1$ and $\hat{E} = \cup_{i=1}^{h} E_i$. In other words, the aggregate is defined as the union of all edges across all layers of the multiplex. In the literature, the aggregate of a multiplex is often defined in this way. This method can aggregate layers of a multiplex in which the layers are unweighted graphs. However, it is not sufficient for a weighted graph, particularly a complete weighted graph. In addition, the strengths between layers are not accounted for.

Let's consider that the edge weights between nodes provides a normalized measure of similarity between zero and one. We can define the weight of a path between two nodes in the multiplex to be the product of the edges between each node in each step of the path. Since the weight between nodes is a measure of similarity or information shared between the nodes, it follows that the weight of the path provides a measure of information flowing through the path.

There are a number of ways we can provide a new measure of similarity between two nodes given the properties of the multiplex network. One way would be to take the mean of the direct paths connecting each patient to and from another patient in each and every layer. We defined this mathematically as follows:

$$R_{direct} = \frac{\sum_{i=1}^{h}(\boldsymbol{M}|_{p_iq_i} + \sum_{j=1,j\neq i}^{h}\boldsymbol{M}^2|_{p_iq_j})}{h^2}, \tag{17}$$

Where h is the number of layers in the multiplex, $M|_{p_iq_i}$ is the element in the multiplex matrix representing the weight between node p and q in layer i and $M^2|_{p_iq_j}$ is the element in the square of the multiplex network, representing the weight of the path from node p in layer i to node q in layer j. Another way would be to take the maximum or minimum information shared directly between two nodes.

$$R_{direct_{min}} = \min_{i=1}^{h}\left(\boldsymbol{M}|_{p_iq_i} + \sum_{j=1,j\neq i}^{h}\boldsymbol{M}^2|_{p_iq_j}\right) \tag{18}$$

$$R_{direct_{max}} = \max_{i=1}^{h}\left(\boldsymbol{M}|_{p_iq_i} + \sum_{j=1,j\neq i}^{h}\boldsymbol{M}^2|_{p_iq_j}\right) \tag{19}$$

In many situations, a pair of nodes in a network does not communicate only through the shortest-path routes connecting both nodes, but also through all possible routes connecting both nodes. The number of these possible routes can be enormous. Moreover, the information can also go back and forth before connecting the pair of nodes. Network communicability, which was introduced by Estrada and Gómez-Gardeñes (2014), attempts to quantify such correlation effects in the communication between nodes in complex networks. Estrada and Gomez-Gardenes defined communicability as a measure that "quantifies the number of possible routes that two nodes have to communicate with each other." In multiplex networks, the communicability, C, between two nodes p and q, is a weighted sum of all walks from p to q.

$$C_{pq} = \boldsymbol{I} + \boldsymbol{M} + \frac{\boldsymbol{M}^2}{2\,!} + \ldots = \sum_{k=0}^{k}\frac{\boldsymbol{M}^k}{k\,!}\bigg|_{pq}. \tag{20}$$

Hence, the communicability between nodes p and q is given by:

$$C_{pq} = [e^{(\mathbf{A}_L + \mathbf{V}_{LL})}]_{pq} = [e^{\mathbf{M}}]_{pq}, \tag{21}$$

where the p, q-th entry in the minor, $\boldsymbol{C}$, defines the communicability broadcasted from node p in layer i to node q in layer j. Therefore, the communicability broadcasted and received by the nodes in the multiplex is given by:

$$\boldsymbol{C} = e^{(\mathbf{A}_L + \mathbf{V}_{LL})} = \begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1h} \\ C_{21} & C_{22} & \ldots & C_{2h} \\ \vdots & \vdots & \ddots & \ddots \\ C_{h1} & C_{h2} & \ldots & C_{hh} \end{pmatrix} \tag{22}$$

Since all nodes are present in each layer of the multiplex, we can calculate the integrated communicability from node p and q in all layers in the multiplex by taking the harmonic mean of the communicability between them in each minor in the matrix $\boldsymbol{C}$.

$$\hat{C}_{pq} = \frac{h}{\sum_{i=1}^{h}\frac{1}{[C_{i,i}]_{pq}} + \sum_{j,k=1,j\neq k}^{h}\frac{1}{[C_{jk}]_{pq}}}. \tag{23}$$

Hence, the integrated communicability matrix is formed by:

$$\hat{\mathbf{C}} = \begin{pmatrix} 0 & \hat{C}_{12} & \ldots & \hat{C}_{1h} \\ \hat{C}_{21} & 0 & \ldots & \hat{C}_{2h} \\ \vdots & \vdots & \ddots & \ddots \\ \hat{C}_{h1} & \hat{C}_{h2} & \ldots & 0 \end{pmatrix}, \tag{24}$$

where $\hat{C}_{ij}$ represents the interaction of layer i with layer j. Therefore, this multiplex network model is applicable to integrate omics and clinical information of a number of diseases or patients in an efficient way.

Evaluation

We incorporated verified data from different data source with our software. Data integration reduces noise associated with each experimental limitation, thus increases sensitivity and specificity to detect true association relationships which results in less number of false positives. By integrating different types of omics and clinical data can produce more reliable predictions with increased sensitivity and specificity for detecting true functional disease comorbidity associations. This can help in finding the hidden connections between complex diseases. Such connections between complex diseases reflect common biological pathways and biological functions that may become manifest in the form of comorbidity. For an example, we show a comparative representation of dysregulated genes and lifestyle impact on the disease comorbidity in **Figure 11**. Here, panel A (see **Figure 11A**) represents an example of only dysregulated genetic influences on diseases with good lifestyle. Panel B (see **Figure 11B**) shows an example of only bad lifestyle influences on diseases with no genetic variation. Panel C (see **Figure 11C**) represents the combined impacts of lifestyle and dysregulated genes on diseases. Here, we observed that the combined impact of both lifestyle and dysregulated genes influences more and multiway on the diseases and disease comorbidities. It is conceivable that by integrating the data ranging from genotype to multiple levels of phenotypes, more precise and robust stratification of the patients with clinical outcome difference can be achieved.

Discussion

Development of methods combining omics, ontology and clinical information could assist clinical decision making and

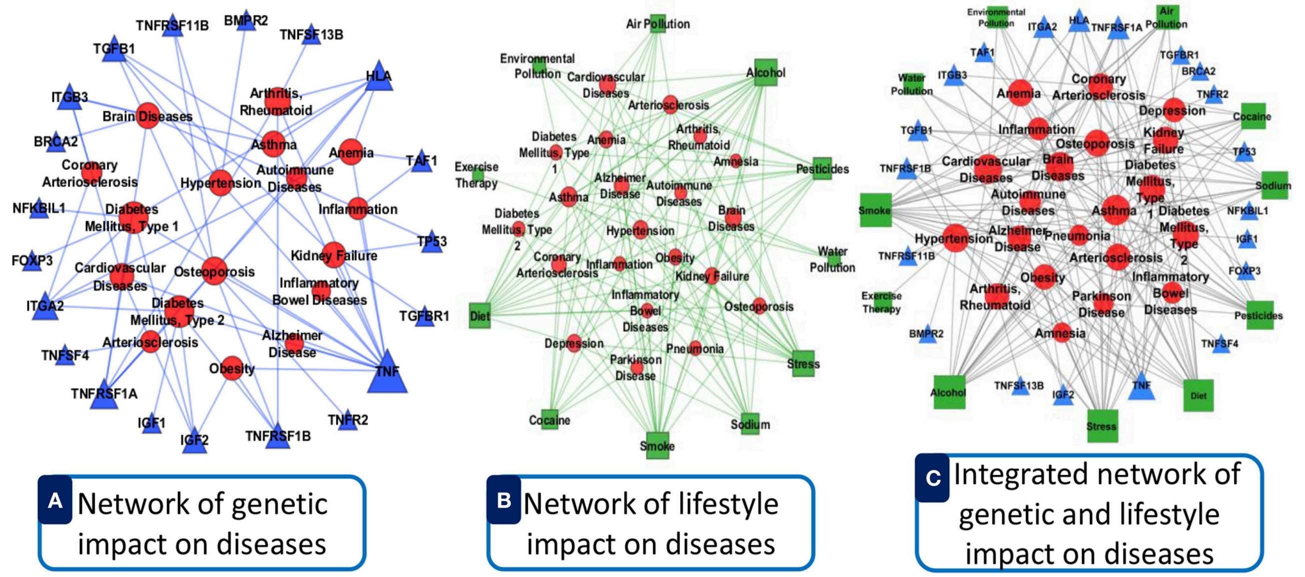


FIGURE 11 | Comparative representation of dysregulated genes and lifestyle impact on the disease comorbidity. (A) Represents an example of only dysregulated genetic influences on diseases with good lifestyle. **(B)** Shows an example of only bad lifestyle influences on diseases with no genetic variation. **(C)** Represents the combined impacts of lifestyle and dysregulated genes on diseases. Here, the circular red nodes nodes represent diseases name, triangular blue nodes represent genes symbols and square green nodes represent lifestyle factors. The size of the nodes represents the degree of associations.

represent a large step toward personalized medicine. Proactive and personalized medicine will bring fundamental changes to health care, taking carefully targeted preventative or therapeutic action at the earliest indications of risk or disease. In order to facilitate the necessary changes, better tool is needed for assessing risk and optimizing treatments, which in turn require better understanding of disease interdependencies, genetic influence, and translation into a patient's future. However, most software is designed to make a prediction about a single disease or a class of some specific diseases based on the single omics or clinical information. Phenomizer is a web-based system that produces a ranked list of hereditary diseases, taking a set of clinical features (Köhler et al., 2009). This system only considers the phenotypic annotation to diseases, and semantic similarity metrics to measure phenotypic similarity between query phenotypes and disease phenotypes with the use of the HPO (Robinson and Mundlos, 2010). Another software DGFinder which is used to assess candidate genes in interested chromosome regions for their possibility relating to a given disease (Yuan et al., 2010). It integrated a dataset containing 1045 genes related to 305 diseases. Hidalgo et al. analyzed comorbidity associations using the medical records (Hidalgo et al., 2009). There are some online information retrieval tools, such as AmiGO[4] and QuickGO[5], to collect gene annotation data from various databases and manually discover the correlations or similarities of gene products by their biological functions (Binns et al., 2009). FindZebra (Dragusin et al., 2013) is a vertical search engine for rare diseases. This system does not consider the genetic effects on disease or phenotypic effects on genes rather it presents a list of disease documents for a given query of symptoms. CARE uses collaborative filtering methods to predict each patient's disease risks based only on their own medical history and that of similar patient's information (Davis et al., 2010). Recently, a tool KnIT has been developed for the complete medical literature knowledge integration (Spangler et al., 2014). DisGeNET is a coherent tool that analyses and interprets human gene network to disease network (Bauer-Mehren et al., 2010). It is able to display gene-disease association networks as bipartite graphs and provides gene centric and disease centric views of the data.

An R package "comorbidities" is able to categorize ICD-9-CM codes based on published 30 comorbidity indices using Deyo adaptation of Charlson index and the Elixhauser index (Deyo et al., 1992; Elixhauser et al., 1998). Our previous R package comoR that provides relative risk, ϕ-correlation, associated genes and pathway between the comorbidity diseases (Moni and Lio, 2014). It is limited to gene expression and pathway molecular data. To our knowledge, there is no available complete software tool for the prediction of disease comorbidities maps based on the multiple omics, gene ontology, phenotype and environmental influences. So, we developed POGO, another R package that implements different statistical approach for the prediction of disease comorbidity maps by integrating diverse set of data. This software could provide comorbidity mapping among all diseases using ontology, miRNA, SNPs, CNVs, phenotypic and environmental information. This software also incorporated a prediction model that explores the past medical patient

[4] http://www.godatabase.org.

[5] http://www.ebi.ac.uk/ego/.

history to determine the risk of patients to develop future diseases.

Patient's omics data is becoming important for clinical decision making, including disease risk assessment, disease diagnosis and subtyping, drug therapy and dose selection (Ullman-Cullere and Mathew, 2011). In the near future, physician will have to consider omics implications to patient care throughout their clinical work flow, including electronic prescribing of medications. In the not-so-distant future, as we move in to an era of personalized and preventive medicine, healthy individuals may be tracked by multiple layers of omic and clinical data in an effort to track potential disease progression. Our software tool incorporated an integrated framework to establish the associations between genetic diseases and ontology information, which may help to uncover the molecular mechanisms of genetic diseases. The identified disease patterns from POGO could be useful for further investigations with regards to their diagnostic utility or help in the prediction of novel therapeutic targets. Therefore, POGO could be helpful for the personalized medicine system. They are able to detect many diseases at the earliest detectable phase, weeks, months, and maybe years before symptoms appear. POGO could easily be integrated into pipelines for high-throughput analysis, such as Galaxy, and other gene expression data mining, protein interactions validation, predicting causal relationships among phenotypes and miRNA-regulated network interpretation. The underlying hypothesis behind this line of research is that once we catalog all disease-disease relations through the omics, ontology, phenotypic and environmental influence, we will be able to predict the susceptibility of each individual to future diseases using various molecular biomarkers, ushering us into an era of predictive medicine.

Thus, a combination of genetic, ontology and population-level data and information could be analyzed by this software tool to establish and study novel hypotheses about unknown disease mechanisms and disease comorbidity. Understanding how different diseases relate to each other will not only provide us with a global view of disease associations, but also provide potentially new insights into the etiology, classification, and design of novel therapeutic interventions. This has led to the advent of stratified medicine, which translates advances in basic research by targeting etiological mechanisms underlying diseases. Method and tool for stratifying (classifying) patients in order to reliably predict prognosis or success of treatments are of critical importance in the field of medicine. However, with the identification of the new omics and clinical information, we need to update the integrated databases of the POGO. Using the temporal data explored by the time dimension approach, POGO could be extended to predict the time of expected disease diagnosis in addition to the likelihood of occurrence. The result is a patient stratification could be based on more complete profiles than the primary diagnosis. Therefore, POGO is useful for the stratified medicine.

Conclusion

Integration of multi-omics, ontology and phenotypic information is important for comorbidity prediction and patient stratification. Therefore, our methodological framework and software for integrating genetic and clinical data could be applicable in clinical decision making for personalized medicine. We expect that this combined approach may increase accuracy and decrease effort for disease comorbidity diagnosis. POGO software tool provides robust approaches to study disease comorbidity mappings by integrating omics, phenotype and ontology information, which can be easily integrated into pipelines for high-throughput and clinical data analysis, and to predict causal inference of a disease. This software tool will help to gain a better understanding of the complex pathogenesis of disease risk phenotypes and the heterogeneity of disease comorbidities. Moreover, the disease comorbidity patterns identified using this software tool could be useful for diagnostic utility or to help in the prediction of novel therapeutic targets. Thus, this software tool could be applicable in personalized medicine and clinical bioinformatics. So our software tool for comorbidity diagnosis and patient stratification could result in effective aids to the health practice. This will not only result in improving health outcomes of the patient, but also in reducing the health care costs.

Availability and Requirements

The software package POGO has been written in the platform independent R programming language. It requires R version 2.16 or newer to run. The software is freely available at www.cl.cam.ac.uk/~mam211/POGO/ and will appear in Comprehensive R Archive Network (CRAN) at (http://cran.r-project.org/).

Funding

This work is supported by the EU Mission T2D project.

Acknowledgment

PL thanks Claudio Franceschi, Paolo Garagnani and Filippo Castiglione for interesting suggestions.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcell.2015.00028

Table S1 | Data of disease-disease associations based on the shared genes.

Table S2 | Data of disease-disease associations based on the shared pathways.

Table S3 | Data of disease-disease associations based on the clinical information.

References

Adamic, L. A., and Adar, E. (2003). Friends and neighbors on the web. *Soc. Netw.* 25, 211–230. doi: 10.1016/S0378-8733(03)00009-1

Astrup, A. (2001). Healthy lifestyles in europe: prevention of obesity and type ii diabetes by diet and physical activity. *Public Health Nutr.* 4, 499–515. doi: 10.1079/PHN2001136

Bae, J. S., Cheong, H. S., Kim, J.-H., Park, B. L., Kim, J.-H., Park, T. J., et al. (2011). The genetic effect of copy number variations on the risk of type 2 diabetes in a korean population. *PLoS ONE* 6:e19091. doi: 10.1371/journal.pone.0019091

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

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). Disgenet: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26, 2924–2926. doi: 10.1093/bioinformatics/btq538

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36, 431–432. doi: 10.1038/ng0504-431

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). Quickgo: a web-based tool for gene ontology searching. *Bioinformatics* 25, 3045–3046. doi: 10.1093/bioinformatics/btp536

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C., Gómez-Gardeñes, J., Romance, M., et al. (2014). The structure and dynamics of multilayer networks. *Phys. Rep.* 544, 1. doi: 10.1016/j.physrep.2014.07.001

Davis, D. A., Chawla, N. V., Christakis, N. A., and Barabási, A.-L. (2010). Time to care: a collaborative engine for practical disease prediction. *Data Mining Knowl. Discov.* 20, 388–415. doi: 10.1007/s10618-009-0156-z

Deyo, R. A., Cherkin, D. C., and Ciol, M. A. (1992). Adapting a clinical comorbidity index for use with icd-9-cm administrative databases. *J. Clin. Epidemiol.* 45, 613–619. doi: 10.1016/0895-4356(92)90133-8

Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jørgensen, H. L., Cox, I. J., et al. (2013). Findzebra: a search engine for rare diseases. *Int. J. Med. Inform.* 82, 528–538. doi: 10.1016/j.ijmedinf.2013.01.005

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Med. Care* 36, 8–27. doi: 10.1097/00005650-199801000-00004

Estrada, E., and Gómez-Gardeñes, J. (2014). Communicability reveals a transition to coordinated behavior in multiplex networks. *Phys. Rev. E* 89:042819. doi: 10.1103/PhysRevE.89.042819

Fredman, D., Siegfried, M., Yuan, Y. P., Bork, P., Lehväslaiho, H., and Brookes, A. J. (2002). Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.* 30, 387–391. doi: 10.1093/nar/30.1.387

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* 5:e1000353. doi: 10.1371/journal.pcbi.1000353

Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., and Nakamura, Y. (2002). Jsnp: a database of common gene variations in the japanese population. *Nucleic Acids Res.* 30, 158–162. doi: 10.1093/nar/30.1.158

Hu, G., and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PLoS ONE* 4:e6536. doi: 10.1371/journal.pone.0006536

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., et al. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* 5:4022. doi: 10.1038/ncomms5022

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease micrornas through a human phenome-micrornaome network. *BMC Syst. Biol.* 4(Suppl. 1):S2. doi: 10.1186/1752-0509-4-S1-S2

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res.* 37(Suppl. 1), D98–D104. doi: 10.1093/nar/gkn714

Kahn, S. E., Hull, R. L., and Utzschneider, K. M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* 444, 840–846. doi: 10.1038/nature05482

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457–464. doi: 10.1016/j.ajhg.2009.09.003

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

Lee, D.-S., Park, J., Kay, K., Christakis, N., Oltvai, Z., and Barabási, A.-L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9880–9885. doi: 10.1073/pnas.0802208105

Lewis, S. N., Nsoesie, E., Weeks, C., Qiao, D., and Zhang, L. (2011). Prediction of disease and phenotype associations from genome-wide association studies. *PLoS ONE* 6:e27175. doi: 10.1371/journal.pone.0027175

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023

Liu, Y. I., Wise, P. H., and Butte, A. J. (2009). The etiome: identification and clustering of human disease etiological factors. *BMC Bioinform.* 10(Suppl. 2):S14. doi: 10.1186/1471-2105-10-S2-S14

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., et al. (2008). An analysis of human microrna and disease associations. *PLoS ONE* 3:e3420. doi: 10.1371/journal.pone.0003420

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., and Scherer, S. W. (2014). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42, D986–D992. doi: 10.1093/nar/gkt958

McCarroll, S. A., and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* 39, S37–S42. doi: 10.1038/ng2080

McKusick, V. A. (2007). Mendelian inheritance in man and its online version, omim. *Am. J. Hum. Genet.* 80, 588. doi: 10.1086/514346

Moni, M. A., and Lio, P. (2014). comor: a software for disease comorbidity risk assessment. *J. Clin. Bioinform.* 4:8. doi: 10.1186/2043-9113-4-8

Park, J., Lee, D.-S., Christakis, N. A., and Barabási, A.-L. (2009). The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* 5, 262. doi: 10.1038/msb.2009.16

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., and Couto, F. M. (2008). Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinform.* 9(Suppl. 5):S4. doi: 10.1186/1471-2105-9-S5-S4

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83, 610–615. doi: 10.1016/j.ajhg.2008.09.017

Robinson, P. N., and Mundlos, S. (2010). The human phenotype ontology. *Clin. Genet.* 77, 525–534.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi: 10.1111/j.1399-0004.2010.01436.x

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., et al. (2014). "Automated hypothesis generation based on mining scientific literature," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM), 1877–1886.

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

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.* 6:e1000662. doi: 10.1371/journal.pcbi.1000662

Tong, B., and Stevenson, C. (2007). *Comorbidity of Cardiovascular Disease, Diabetes and Chronic Kidney Disease in Australia*. Australian Institute of Health and Welfare.

Ullman-Cullere, M. H., and Mathew, J. P. (2011). Emerging landscape of genomics in the electronic health record for personalized medicine. *Hum. Mutat.* 32, 512–516. doi: 10.1002/humu.21456

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, J. Z., Du, Z., Payattakool, R., Philip, S. Y., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Yang, J. O., Hwang, S., Oh, J., Bhak, J., and Sohn, T.-K. (2008). An integrated database-pipeline system for studying single nucleotide polymorphisms and diseases. *BMC Bioinform.* 9(Suppl. 12):S19. doi: 10.1186/1471-2105-9-S12-S19

Yu, G., Wang, L. G., Yan, G. R., and He, Q. Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31, 608–609. doi: 10.1093/bioinformatics/btu684

Yuan, F., Wang, R., Guan, M., and He, G. (2010). "A novel computational method for predicting disease genes based on functional similarity," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence* (Springer), 42–51.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi: 10.1146/annurev.genom.9.081307.164217

Zhang, S.-H., Wu, C., Li, X., Chen, X., Jiang, W., Gong, B.-S., et al. (2010). From phenotype to gene: detecting disease-specific gene functional modules via a text-based human disease phenotype network construction. *FEBS Lett.* 584, 3635–3643. doi: 10.1016/j.febslet.2010.07.038

Žitnik, M., Janjić, V., Larminie, C., Zupan, B., and Pržulj, N. (2013). Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.* 3:3202. doi: 10.1038/srep03202

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
CELL AND DEVELOPMENTAL BIOLOGY

PERSPECTIVE ARTICLE
published: 14 August 2014
doi: 10.3389/fcell.2014.00039

Fingerprints of a message: integrating positional information on the transcriptome

Erik Dassi* and *Alessandro Quattrone*

Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento, Trento, Italy

Edited by:
Paolo Tieri, Consiglio Nazionale delle Ricerche, Italy

Reviewed by:
Amin R. Mazloom, University of Texas at Arlington, USA
Alessandro Giuliani, Istituto Superiore di Sanità (Italian NIH), Italy

***Correspondence:**
Alessandro Quattrone, Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento, via delle Regole 101, Mattarello, Trento 38123, Italy
e-mail: alessandro.quattrone@unitn.it

The recent explosion of high-throughput sequencing methods applied to RNA molecules is allowing us to go beyond the description of sequence variants and their relative abundances, as measured by RNA-seq. We can now probe for RNA engagement in polysomes, for ribosomes, RNA binding proteins and microRNAs binding sites, for RNA secondary structure and for RNA methylation. These descriptors produce a steadily growing multidimensional array of positional information on RNA sequences, whose effective integration only would bring to decipher the regulatory interplay occurring between proteins, RNAs and their modifications on the transcriptome. This interplay ultimately dictates the degree of mRNA availability to translation, and thus the occurrence of cell phenotypes. However, several issues in data presentation are slowing down effective integration. A standardization effort for new dataset types produced should be urgently undertaken to solve these issues. Providing uniformed experimental details along with datasets processed to be directly usable and employing shared formats would greatly simplify integration efforts, strengthening hypotheses stemming from correlative observations and eventually bringing to mechanistic understanding.

Keywords: transcriptome, integration, post-transcriptional control, translation, RNA-seq, mRNA, data format, standards

PROBING THE BIOLOGICAL STATUS OF WHOLE TRANSCRIPTOMES

The last 15 years have witnessed, starting with the advent of microarray-based gene expression probing, an explosion of high-throughput technologies for the characterization of biological molecules. These technologies, affordable and relatively simple to apply, are steadily paving the way for routine multi-omics studies. The latest of such technologies, high-throughput sequencing (HTS) (Metzker, 2010), has quickly gained widespread acceptance and concurrently enabled several different types of measurements. Its sequence-based nature, permitting to pinpoint relevant features on the genome or transcriptome of interest (position-aware data), and its massively parallel data production capabilities are now indeed applied to the study of a wide array of biological questions. Applications focus on DNA (identification of sequence and copy number variants, mapping of chromatin binding sites by transcription factors and other proteins, chromatin topology studies in nuclei, etc.) (Koboldt et al., 2013) and on RNA (sequence variants of mRNAs and non-coding RNAs, expression levels, mapping of binding sites of RNA binding proteins (RBPs), post-transcriptional modifications etc.), (Ascano et al., 2013; Mutz et al., 2013). Translational regulation of gene expression, in particular, has lately been object of increasing interest: its role in profoundly reshaping transcriptome variations and being the determinant of plasticity in the nascent proteome (Vogel et al., 2010; Stevens and Brown, 2013) is increasingly appreciated. Consequently, omic approaches have been developed to investigate which features of an mRNA may influence its translation rate, which trans-factors play a role in such regulatory processes and how these two aspects combine to yield the final protein levels. We will focus on RNA-centered methods to examine the types of biological information they can provide; we will then look at how this information should be integrated to allow us a better understanding of both the global transcriptome dynamics and their effects on phenotype.

As shown in **Figure 1**, such methods can be classified by their descriptive capability, either *molecular* for the entire RNA or *sub-molecular* for specific RNA portions, and the kind of description they provide, *quantitative*, *qualitative* or both. The description of entire transcripts is provided by RNA-seq (Mutz et al., 2013), an HTS-based method which gives the sequence of coding and non-coding transcripts, including mapping of alternative transcription or termination sites, splice variants produced on the same locus and the presence of expressed sequence polymorphisms. Since different transcripts can be quantified in their relative abundance, this type of information is both qualitative and quantitative. The polysome profiling method (Arava, 2003; Gandin et al., 2014) is based on the separation by sucrose gradient centrifugation of cellular fractions containing polysomes and the subsequent quantification of their mRNA relative (to the total lysate or to the fractions not containing polysomes) abundance, which can be performed by RNA-seq or by the more conventional microarray analysis. The resulting information is a quantitative and qualitative description of the degree of polysomal engagement for every transcript (by which the molecular nature of this method), the so called translatome (Tebaldi et al., 2012); a calculation of

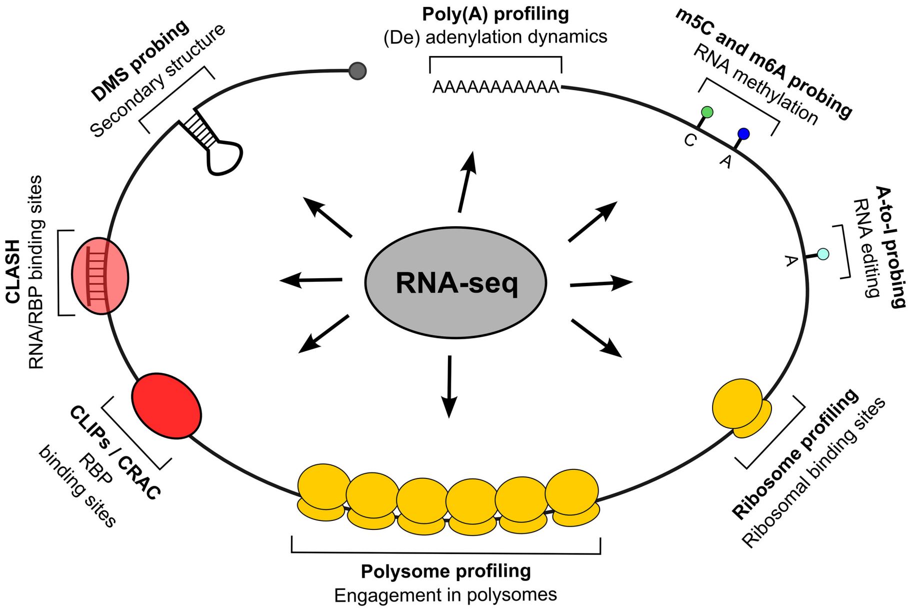


FIGURE 1 | Techniques for positional whole-transcriptome probing. The figure displays techniques allowing to study transcriptomes at various observational levels, with particular regard to positional information; all techniques, indicated by their representative feature on transcripts, are based on RNA-seq.

translational efficiency can be done by this assay. The qualitative component of polysome profiling is given by computational approaches which allow us to investigate the differential association of mRNAs produced by the same gene locus (splice and 5′/3′ variants) with the polysomes (Frac-seq, Sterne-Weiler et al., 2013), or which measure the effect of single-nucleotide polymorphisms on translational efficiency (Li et al., 2013). Ribosome profiling (Ingolia, 2014) aims at providing a snapshot of mRNAs under translation by scoring the transcript regions which are protected from nuclease attack by ribosomes. It is a RNA-seq-based method of the submolecular type: obtainable information can be integrated at the transcript level but has a positional content, so that translation initiation and termination sites, potential translation stalling events, upstream ORF translation, can be derived (Ingolia et al., 2011). Besides engagement in translation, another type of general, qualitative description of transcript status is the secondary structures pattern, recently become available to profiling through nucleotide accessibility probing coupled with RNA-seq (Ding et al., 2014; Rouskin et al., 2014; Talkish et al., 2014; Wan et al., 2014). Eventually, a transcript component which can be investigated is the poly(A) tail: two recent methods, PAL-seq (Subtelny et al., 2014) and TAIL-seq (Chang et al., 2014), exploit RNA-seq to characterize its length and potential modifications (such as uridylation and guanylation). The same principle of nuclease protection exploited in ribosome profiling is then systematically applied in locating RNA-associated "footprints" of RBPs. The CLIP techniques family: HITS-CLIP, PAR-CLIP, and iCLIP (Ule et al., 2003; Hafner et al., 2010; Konig et al., 2010) and the CRAC approach (Granneman et al., 2009) exploit an UV-induced crosslinking of RNA and associated proteins (with the option of using photoactivatable nucleotides, as done in PAR-CLIP) to enable the identification of RNA targets and binding sites for single, immunoprecipitated RBPs. These are therefore submolecular and essentially qualitative approaches. A variant method, CLASH (Helwak et al., 2013), introduces a RNA ligation step to locate sites where other RNAs are associated in trans in a protein complex, allowing to experimentally identify miRNA binding sites. CLIP methods can also be extended to consider many RBPs at once: "global CLIP" approaches such as protein occupancy profiling (Baltz et al., 2012) and PIP-seq (Silverman et al., 2014) thus provide contact sites for all RBPs at once on a transcriptome.

Coming finally to the most submolecular level, that of single nucleotides, mRNA editing events (such as adenosine to inosine conversions) can be revealed either by inosine chemical erasing (ICE), as in Sakurai et al. (2014), or by directly looking for sequence variants in RNA-seq reads (St. Laurent et al., 2013; Bazak et al., 2014). Eventually, RNA 5-methylcytosine and N6-methyladenosine nucleotide methylation can be detected with single-nucleotide precision, respectively by bisulfite conversion (Squires et al., 2012; Edelheit et al., 2013) and immunoprecipitation (Dominissini et al., 2012; Meyer et al., 2012; Khoddami and Cairns, 2013) or by other biochemical methods (Hussain et al., 2013; Liu et al., 2013).

APPROACHES FOR THE INTEGRATION OF TRANSCRIPT-CENTERED OMICS

Currently, several hundred papers employing the described transcriptome-based omics methods have been published, including a considerable number of pure RNA-seq datasets, secondary structure probing, editing and methylation profiles for the most common cell lines and organisms (see **Figure 1**), and at least 40 different CLIP or CLIP-like datasets (Dassi et al., 2014). With such a huge amount of data available, the naturally arising question is how to integrate these different types of information to obtain more insights than if considering single datasets in isolation. Several works have approached this problem so far. As shown in **Table 1**, they can be classified according to the different perspectives adopted in doing so.

A first, post-experimental way of integrating these heterogeneous data sets consists in building a database presenting all the collected data together, thus allowing users to prioritize and validate potential connections. Mining the data, superimposed on a reference genome, can be approached by looking for single genes (as happens in genome browsers) or by studying interesting gene lists (e.g., through functional enrichment or co-regulation analyses). This road was taken by AURA/AURA2 (Dassi et al., 2012, 2014), DORiNA (Anders et al., 2012), and starBase (Li et al., 2014). The first provides RBP and miRNA binding sites, cis-elements sites, RNA editing, and methylated nucleotides; the second offers RBP binding sites and predicted miRNA targets; the last includes RBP binding sites and miRNA interactions with coding and non-coding RNAs. While these databases are of general interest and can be useful for a broad spectrum of preliminary investigations, they still mostly contain data obtained in a limited set of particularly common model systems or cell lines (e.g., HEK293 cells): users will then likely need to trust this information to hold in their system of interest or validate the interaction in their specific conditions (e.g., for an RBP-mRNA interaction, by integrating expression data to check whether it could indeed occur, or by performing a RIP-qPCR assay in their system).

The second, most reliable method is obviously measuring several mRNA features in the system under study, focusing on a specific biological question, and then proceed by intersecting the obtained data to generate hypotheses stemming from the correlation of specific features. An intuitive example of this approach is in profiling the transcriptome and the translatome (the last through polysomal profiling, for instance) in various conditions (e.g., drug treatment vs. control) to identify which genes are subjected to translational control and the impact the treatment may have on translational efficiency (computed as the translatome vs. transcriptome ratio): this has already been done in a number of works (Genolet et al., 2011; Bates et al., 2012; Fu et al., 2012; Tebaldi et al., 2012; Courtes et al., 2013; Dudek et al., 2013; Willimott et al., 2013). A variation on this theme could include, in parallel, a miRNAs profiling in the system to correlate differences in their levels with differences in translational efficiency, generating candidate determinants of the latter changes (Clarke et al., 2012). Another example is the secondary structure and translational efficiency profiling of mRNAs in the system under

Table 1 | Current approaches for positional information integration on the transcriptome.

Name	Description	Scope	Potential issues	References
Integrated databases	Collecting and presenting available datasets of heterogeneous types and biological sources; allowing users to mine the data types in combination	Global over a vast number of different data types	Data quality and processing assessment not always possible; achieving database completeness and constant content update is particularly time-intensive	Anders et al., 2012; Dassi et al., 2012, 2014; Li et al., 2014
Multi-level profiling	Performing various types of measurements (i.e., mRNA levels, RNA secondary structure, RNA methylation) in the same system of interest (e.g., cell line) to derive correlative patterns	Global over a limited number of data types	Need very different experimental and data analysis expertise; results applicability is limited to the studied system	Genolet et al., 2011; Bates et al., 2012; Clarke et al., 2012; Dominissini et al., 2012; Fu et al., 2012; Tebaldi et al., 2012; Courtes et al., 2013; Dudek et al., 2013; Willimott et al., 2013; Zheng et al., 2013; Ding et al., 2014; Mao et al., 2014; Wang et al., 2014a
Measurements & public data exploitation	Performing a small number of measurements (i.e., mRNA levels only) in the system of interest, and exploiting public data to study genes derived from these measurements (i.e., presence of translational regulation) to infer and validate potential regulatory mechanisms and patterns	Over a small number (dozens) of interesting genes	Publicly available data on the system one wants to use may not be available; further validation and/or mechanistic experiments may be needed	Mazza et al., 2013; Avery-Kiejda et al., 2014; Schueler et al., 2014; Wang et al., 2014b

The table describes currently applied approaches to the integration of position-aware RNA datasets. Scope of the various approaches and associated potential issues are outlined along with the references of works employing them.

study, aiming at the identification of structural patterns conferring translational advantages to the mRNAs containing them (Ding et al., 2014; Mao et al., 2014). Along the same line is coupling m6A methylation probing and RNA-seq measurements in the same system: this allows us to understand whether methylation alters mRNA level, stability and splicing patterns in the conditions under investigation (Dominissini et al., 2012; Zheng et al., 2013; Wang et al., 2014a).

The last integration method we describe is based on bridging the previous two approaches: combining a limited number of direct measurements performed in the system of interest with the wealth of data available in public databases such as the ones described above (even though these data may not be produced in the same model). One may thus investigate whether, for instance, an RBP or a miRNA is controlling a group of mRNAs, whether the gene set under analysis is enriched with a particular feature (e.g., a 3′UTR cis-element in the form of a secondary structure, methylated nucleotides, etc.) or match observed patterns for one feature type (e.g., presence of a secondary structure feature) with public data (e.g., presence of trans-factor binding sites) to deduce general rules (e.g., preference of a trans-factor for that given structural feature). While this method leads to hypotheses that need validation as they may not hold in the system of interest, it allows speeding up the investigation and reducing the hypotheses space, consequently lowering experimental uncertainty, time and cost. This approach has been enabled just recently, due to the availability of the databases discussed above. However, in the few published works adopting it, it is usually applied to the integration of data focused on a few specific mRNAs, which have been previously selected for their behavior as observed in the ongoing study (Mazza et al., 2013; Avery-Kiejda et al., 2014; Wang et al., 2014b). One exception is the recent work by Schueler and colleagues, in which protein contact sites obtained by a global PAR-CLIP on two cell lines are integrated with known RBP binding sites to infer differential protein occupancy patterns (Schueler et al., 2014).

Summing up, even though the approaches we have discussed are useful examples of data integration applied to the structure and the behavior of mRNAs, it is evident that these are still early and limited efforts. Indeed, as also testified by the small number of published works, there still is a significant lack of accepted practices and standard procedures which could render these approaches of effective routine usage. Having built a database focused on post-transcriptional regulation (Dassi et al., 2014), we realized that processed data, as submitted by the authors, vary widely in their processing level: if we take CLIPs datasets as an example, some datasets include the definition of sites bound by the studied RBP while others are limited to, for instance, the indication of T > C conversions (for PAR-CLIP); obviously this marked differences put additional burden on whoever wants to use multiple datasets, produced in different experiments, together, in order to generate new hypotheses. Furthermore, methods are often described in many ways, with different levels of detail, representing further obstacles in individuating steps needed to make these datasets truly comparable. A last general issue is the absence of a systematic way to evaluate data quality and robustness, considering for example the presence of replicates, the number of supporting reads and other parameters linked to specific techniques.

THE NEED FOR STANDARDIZATION

Given the outlined issues, we asked which steps could be taken to improve the exploitability and the integration potential of the RNA-centered high throughput data. We propose two simple, preliminary actions. The first is the enforced use of standard file formats with precisely defined fields, a relatively simple goal to achieve. The second is the enforced provision of a minimal set of information—enhancing dataset description, uniformity and allowing quality evaluations—at submission time (similarly to what was established and is currently enforced for microarrays with MIAME and related initiatives; Brazma et al., 2001; Rayner et al., 2006). This could be straightforwardly imposed by repositories commonly used for high-throughput datasets submission such as GEO (Barrett et al., 2013), ArrayExpress (Parkinson et al., 2005), and SRA (Wheeler et al., 2008).

Concerning the first requirement, we need to deal with two types of data: intervals (such as RBP and miRNA binding sites obtained through CLIPs) and per-nucleotide intensities (continuous values such as the ones produced by RNA methylation or secondary structure probing assays). Intervals are most often represented by means of the Browser Extensible Data (BED) format: its main advantage lies in the extreme simplicity of fields definition, which nevertheless allows a certain degree of detail, making it also feasible to represent several datasets in a single file (by for instance using the name field to distinguish the RBP/miRNA and possibly specifying methods and data source publication in the description field). Furthermore, BED files can be converted to bigBed (Kent et al., 2010), the associated binary indexed format that is efficient to process and use with genome browsers even for huge datasets. Concerning continuous values, they are most often stored by means of either a format similar in nature to BED, called bedGraph, or through another common option called Wiggle (Kent et al., 2010). Both formats are stripped down to the essential and are not really intended to allow mixing different datasets in the same file; the file header however leaves room for some description to be added; furthermore, both can be converted to the binary indexed bigWig format (Kent et al., 2010), similarly to what mentioned above for bigBed. Given the versatility and already widespread use of these two formats, coupled with the storage and display efficiency, we propose that they should be deemed as de facto standards and systematically required for new data submissions.

For the second requirement (minimal set of parameters describing a dataset), which information should be considered as essential for the data to be exploited at their full potential? First of all, in the case of CLIP datasets intervals representing binding sites should be provided, rather than including raw per-nucleotide data only. Many scientists would not or cannot go the extra mile to compute intervals out of per-nucleotide data by themselves, and would thus loose the opportunity to use them. Furthermore, methods employed for data analysis should be described, at least briefly, indicating how intervals or per-nucleotide intensities (e.g., in the case of secondary structure data) were computed from raw reads. Eventually, basic quality metrics such as the number of replicates and the read depth supporting a given interval/position, along with call significance *p*-values (where appropriate) should also be provided to let the users judge on the data robustness, eventually allowing the application of homogeneous stringency filters when integrating multiple datasets. We believe that this “information package” could be enough to describe the data under study to an extent that will eventually make going back to the raw data unnecessary: we therefore propose that these information should be required when submitting a dataset of this sort.

Pushing further on this proposal, we may also consider the need for a dedicated repository storing transcriptome-centric positional data. Similarly to what major journals ask for microarrays-containing works, submission to this repository could be a de facto requirement for publication and have an unique ID assigned, to which direct reference could be made in publications further easing data traceability. Using one of the currently available databases as a repository of this sort could also have the advantage of allowing us to display various datasets together, integrated in a transcript-oriented way, thus providing a first glimpse of the data along with the possibility to retrieve them. Of course, this collection of proposals, which goes along the lines of several other “reproducible research” initiatives, can become a reality only if the majority of scientists in the field agree and commit to sustain it by complying with these recommendations.

CONCLUSION

The availability of techniques based on high-throughput sequencing is fostering the investigation of the biological behavior of transcriptomes with an unprecedented level of detail and a continuously increasing amount of available data types: the very nature of this technology effectively allows us to pinpoint the location of features responsible for known and unknown biochemical properties of mRNAs and non-coding RNAs which may ultimately influence mRNA translation. However, the integration of these datasets is still in its infancy, with only a few approaches and applications in the literature and a lot of room for improving and making these efforts much easier and useful. We think that this process could be eased by committing to the introduction of standardization measures involving file formats, minimal information to be provided for dataset description and, possibly, the setup of a dedicated data repository. The choice to advance a proposal limited to transcripts biological features is justified in our opinion by the momentum gained by studies in post-transcriptional regulation of gene expression, by the several RNA-seq-based techniques introduced in the last 2 years, and the exponential growth of datasets of this type being released. We therefore think that the effort needed to implement such proposal could be worthy and fruitful. While certainly requiring coordination between laboratories studying the topic, initiatives like OBO (Smith et al., 2007), MIAPE (Taylor et al., 2007), and BioBricks (Smolke, 2009) have shown that it is possible to implement and sustain a standardization effort aimed in our case at a better exploitation of high-throughput data. Given the pace at which these data are accumulating, we need for sure to urgently push their integrated exploitation to its fullest extent.

REFERENCES

Anders, G., Mackowiak, S. D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., et al. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.* 40, D180–D186. doi: 10.1093/nar/gkr1007

Arava, Y. (2003). Isolation of polysomal RNA for microarray analysis. *Methods Mol. Biol.* 224, 79–87. doi: 10.1385/1-59259-364-X:79

Ascano, M., Gerstberger, S., and Tuschl, T. (2013). Multi-disciplinary methods to define RNA-protein interactions and regulatory networks. *Curr. Opin. Genet. Dev.* 23, 20–28. doi: 10.1016/j.gde.2013.01.003

Avery-Kiejda, K. A., Braye, S. G., Mathe, A., Forbes, J. F., and Scott, R. J. (2014). Decreased expression of key tumour suppressor microRNAs is associated with lymph node metastases in triple negative breast cancer. *BMC Cancer* 14:51. doi: 10.1186/1471-2407-14-51

Baltz, A. G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* 46, 674–690. doi: 10.1016/j.molcel.2012.05.021

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bates, J. G., Salzman, J., May, D., Garcia, P. B., Hogan, G. J., McIntosh, M., et al. (2012). Extensive gene-specific translational reprogramming in a model of B cell differentiation and Abl-dependent transformation. *PLoS ONE* 7:e37108. doi: 10.1371/journal.pone.0037108

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

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371. doi: 10.1038/ng1201-365

Chang, H., Lim, J., Ha, M., and Kim, V. N. (2014). TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications. *Mol. Cell* 53, 1044–1052. doi: 10.1016/j.molcel.2014.02.007

Clarke, C., Henry, M., Doolan, P., Kelly, S., Aherne, S., Sanchez, N., et al. (2012). Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics* 13:656. doi: 10.1186/1471-2164-13-656

Courtes, F. C., Vardy, L., Wong, N. S., Bardor, M., Yap, M. G., and Lee, D. Y. (2013). Understanding translational control mechanisms of the mTOR pathway in CHO cells by polysome profiling. *N. Biotechnol.* doi: 10.1016/j.nbt.2013.10.003. [Epub ahead of print].

Dassi, E., Malossini, A., Re, A., Mazza, T., Tebaldi, T., Caputi, L., et al. (2012). AURA: atlas of UTR regulatory activity. *Bioinformatics* 28, 142–144. doi: 10.1093/bioinformatics/btr608

Dassi, E., Re, A., Leo, S., Tebaldi, T., Pasini, L., Peroni, D., et al. (2014). AURA 2: Empowering discovery of post-transcriptional networks. *Translation* 2:e27738. doi: 10.4161/trla.27738

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505, 696–700. doi: 10.1038/nature12756

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Dudek, K. M., Suter, L., Darras, V. M., Marczylo, E. L., and Gant, T. W. (2013). Decreased translation of Dio3 mRNA is associated with drug-induced hepatotoxicity. *Biochem. J.* 453, 71–82. doi: 10.1042/BJ20130049

Edelheit, S., Schwartz, S., Mumbach, M. R., Wurtzel, O., and Sorek, R. (2013). Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet.* 9:e1003602. doi: 10.1371/journal.pgen.1003602

Fu, S., Fan, J., Blanco, J., Gimenez-Cassina, A., Danial, N. N., Watkins, S. M., et al. (2012). Polysome profiling in liver identifies dynamic regulation of endoplasmic reticulum translatome by obesity and fasting. *PLoS Genet.* 8:e1002902. doi: 10.1371/journal.pgen.1002902

Gandin, V., Sikstrom, K., Alain, T., Morita, M., McLaughlan, S., Larsson, O., et al. (2014). Polysome fractionation and analysis of mammalian translatomes on a genome-wide scale. *J. Vis. Exp.* doi: 10.3791/51455

Genolet, R., Rahim, G., Gubler-Jaquier, P., and Curran, J. (2011). The translational response of the human mdm2 gene in HEK293T cells exposed to rapamycin: a role for the 5′-UTRs. *Nucleic Acids Res.* 39, 989–1003. doi: 10.1093/nar/gkq805

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9613–9618. doi: 10.1073/pnas.0901997106

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.*Cell* 141, 129–41. doi: 10.1016/j.cell.2010.03.009

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

Hussain, S., Sajini, A. A., Blanco, S., Dietmann, S., Lombard, P., Sugimoto, Y., et al. (2013). NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.* 4, 255–261. doi: 10.1016/j.celrep.2013.06.029

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15, 205–213. doi: 10.1038/nrg3645

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802. doi: 10.1016/j.cell.2011.10.002

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26, 2204–2207. doi: 10.1093/bioinformatics/btq351

Khoddami, V., and Cairns, B. R. (2013). Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.* 31, 458–464. doi: 10.1038/nbt.2566

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* 155, 27–38. doi: 10.1016/j.cell.2013.09.006

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., et al. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909–915. doi: 10.1038/nsmb.1838

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Li, Q., Makri, A., Lu, Y., Marchand, L., Grabs, R., Rousseau, M., et al. (2013). Genome-wide search for exonic variants affecting translational efficiency. *Nat. Commun.* 4, 2260. doi: 10.1038/ncomms3260

Liu, N., Parisien, M., Dai, Q., Zheng, G., He, C., and Pan, T. (2013). Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* 19, 1848–1856. doi: 10.1261/rna.041178.113

Mao, Y., Liu, H., Liu, Y., and Tao, S. (2014). Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 42, 4813–4822. doi: 10.1093/nar/gku159

Mazza, T., Castellana, S., Andriulli, A., Auffray, C., Vinciguerra, M., and Pazienza, V. (2013). Affinity analysis of differentially expressed genes in hepatocytes expressing HCV core genotype 1b or 3a. *Biosystems* 114, 64–68. doi: 10.1016/j.biosystems.2013.05.009

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Mutz, K. O., Heilkenbrinker, A., Lonne, M., Walter, J. G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.* 24, 22–30. doi: 10.1016/j.copbio.2012.09.004

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., et al. (2005). ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 33, D553–D555. doi: 10.1093/nar/gki056

Rayner, T. F., Rocca-Serra, P., Spellman, P. T., Causton, H. C., Farne, A., Holloway, E., et al. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* 7:489. doi: 10.1186/1471-2105-7-489

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* 505, 701–705. doi: 10.1038/nature12894

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

Schueler, M., Munschauer, M., Gregersen, L. H., Finzel, A., Loewer, A., Chen, W., et al. (2014). Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol.* 15:R15. doi: 10.1186/gb-2014-15-1-r15

Silverman, I. M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J. L., et al. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol.* 15:R3. doi: 10.1186/gb-2014-15-1-r3

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255. doi: 10.1038/nbt1346

Smolke, C. D. (2009). Building outside of the box: iGEM and the BioBricks Foundation. *Nat. Biotechnol.* 27, 1099–1102. doi: 10.1038/nbt1209-1099

Squires, J. E., Patel, H. R., Nousch, M., Sibbritt, T., Humphreys, D. T., Parker, B. J., et al. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* 40, 5023–5033. doi: 10.1093/nar/gks144

Sterne-Weiler, T., Martinez-Nunez, R. T., Howard, J. M., Cvitovik, I., Katzman, S., Tariq, M. A., et al. (2013). Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* 23, 1615–1623. doi: 10.1101/gr.148585.112

Stevens, S. G., and Brown, C. M. (2013). *In silico* estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. *PLoS ONE* 8:e57625. doi: 10.1371/journal.pone.0057625

St. Laurent, G., Tackett, M. R., Nechkin, S., Shtokalo, D., Antonets, D., Savva, Y. A., et al. (2013). Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. *Nat. Struct. Mol. Biol.* 20, 1333–1339. doi: 10.1038/nsmb.2675

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., and Bartel, D. P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* 508, 66–71 doi: 10.1038/nature13007

Talkish, J., May, G., Lin, Y., Woolford, J. L. Jr., and McManus, C. J. (2014). Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* 20, 713–720. doi: 10.1261/rna.042218.113

Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., Julian, R. K. Jr., Jones, A. R., et al. (2007). The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 25, 887–893. doi: 10.1038/nbt1329

Tebaldi, T., Re, A., Viero, G., Pegoretti, I., Passerini, A., Blanzieri, E., et al. (2012). Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC Genomics* 13:220. doi: 10.1186/1471-2164-13-220

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302, 1212–1215. doi: 10.1126/science.1090095

Vogel, C., Abreu Rde, S., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., et al. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 6, 400. doi: 10.1038/msb.2010.59

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505, 706–709. doi: 10.1038/nature12946

Wang, W. T., Zhao, Y. N., Yan, J. X., Weng, M. Y., Wang, Y., Chen, Y. Q., et al. (2014b). Differentially expressed microRNAs in the serum of cervical squamous cell carcinoma patients before and after surgery. *J. Hematol. Oncol.* 7:6. doi: 10.1186/1756-8722-7-6

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., and Zhao, J. C. (2014a). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* 16, 191–198. doi: 10.1038/ncb2902

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36, D13–D21. doi: 10.1093/nar/gkm1000

Willimott, S., Beck, D., Ahearne, M. J., Adams, V. C., and Wagner, S. D. (2013). Cap-translation inhibitor, 4EGI-1, restores sensitivity to ABT-737 apoptosis through cap-dependent and -independent mechanisms in chronic lymphocytic leukemia. *Clin. Cancer Res.* 19, 3212–3223. doi: 10.1158/1078-0432.CCR-12-2185

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C. M., Li, C. J., et al. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell* 49, 18–29. doi: 10.1016/j.molcel.2012.10.015

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

Received: 11 June 2014; accepted: 30 July 2014; published online: 14 August 2014.

Citation: Dassi E and Quattrone A (2014) Fingerprints of a message: integrating positional information on the transcriptome. Front. Cell Dev. Biol. ***2****:39. doi: 10.3389/fcell.2014.00039*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 11 February 2015
doi: 10.3389/fgene.2015.00040

Integrating multi-omic features exploiting Chromosome Conformation Capture data

Ivan Merelli[1]*, Fabio Tordini[2], Maurizio Drocco[2], Marco Aldinucci[2], Pietro Liò[3] and Luciano Milanesi[1]

[1] Bioinformatics Unit, Institute of Biomedical Technologies, Italian National Research Council, Milan, Italy
[2] Computer Science Department, University of Torino, Torino, Italy
[3] Computer Laboratory, University of Cambridge, Cambridge, UK

Edited by:
Christine Nardini, Partner Institute for Computational Biology, China

Reviewed by:
Jianlin Cheng, University of Missouri, USA
Zong Wei, Salk Institute for Biological Studies, USA

***Correspondence:**
Ivan Merelli, Bioinformatics Unit, Institute of Biomedical Technologies, Italian National Research Council, via Fratelli Cervi 93, Segrate, Milan 20090, Italy
e-mail: ivan.merelli@itb.cnr.it

The representation, integration, and interpretation of omic data is a complex task, in particular considering the huge amount of information that is daily produced in molecular biology laboratories all around the world. The reason is that sequencing data regarding expression profiles, methylation patterns, and chromatin domains is difficult to harmonize in a systems biology view, since genome browsers only allow coordinate-based representations, discarding functional clusters created by the spatial conformation of the DNA in the nucleus. In this context, recent progresses in high throughput molecular biology techniques and bioinformatics have provided insights into chromatin interactions on a larger scale and offer a formidable support for the interpretation of multi-omic data. In particular, a novel sequencing technique called Chromosome Conformation Capture allows the analysis of the chromosome organization in the cell's natural state. While performed genome wide, this technique is usually called Hi–C. Inspired by service applications such as Google Maps, we developed NuChart, an R package that integrates Hi–C data to describe the chromosomal neighborhood starting from the information about gene positions, with the possibility of mapping on the achieved graphs genomic features such as methylation patterns and histone modifications, along with expression profiles. In this paper we show the importance of the NuChart application for the integration of multi-omic data in a systems biology fashion, with particular interest in cytogenetic applications of these techniques. Moreover, we demonstrate how the integration of multi-omic data can provide useful information in understanding why genes are in certain specific positions inside the nucleus and how epigenetic patterns correlate with their expression.

Keywords: multi-omic data integration, Chromosome Conformation Capture, gene neighborhood map, chromatin spatial organization, linking gene regulatory elements

INTRODUCTION

What is the best way to integrate and represent omic data? This inquiry results critical in an era that is witnessing an explosion of the available molecular biology information. In particular, the integration and the interpretation of omic data in a systems biology view is complex, because actual representations rely on genomic coordinates, discarding at first gene spatial cooperation and renouncing to exploit the real conformation of the DNA in the nucleus. Moreover, approaches that are commonly used to annotate and analyze molecular biology experiments, such as ontology mapping and enrichment analysis, assume as prerequisite an independent sampling of features, which is clearly not satisfied while looking at long-range chromatin interactions (de Wit and de Laat, 2012), since they associate regions that are known to be functionally correlated.

Considering the number of experiments that highlight the importance of co-localization and co-expression of genes (Di Stefano et al., 2013), the possibility of mapping multi-omic features on a map capable of representing the effective disposition of genes in the nucleus can be of great utility. Moreover, the possibility of introducing network concepts to represent the behavior of genomic actors seems a suitable solution for the interpretation of this kind of data, since they allow to map a lot of information in complex, dynamical structures that organize items in an integrated way.

Recent advances in high throughput molecular biology techniques and bioinformatics have provided insights into chromatin interactions on a larger scale (Lieberman-Aiden et al., 2009). A novel technique called Chromosome Conformation Capture (3C) allows the analysis of chromosome organization in the cell's natural state (Duan et al., 2012). The combination of high-throughput sequencing with this technique, generally called Hi–C, allows the characterization of long-range chromosomal interactions genome-wide (Lin et al., 2012). Hi–C gives information about coupled DNA fragments that are cross-linked together due to spatial proximity, providing data of the chromosomal arrangement in the 3D space of the nucleus. If used in combination with chromatin immunoprecipitation, 3C can be employed for the analyses of interactions between DNA and particular proteins, in a technique called ChIA-pet (Fullwood et al., 2009; Dixon et al., 2012; Li et al., 2012; Papantonis et al., 2012).

These techniques allow the description of the nucleus organization at unprecedented resolution, offering the possibility to study the structural properties and spatial organization of chromosomes. This is of critical importance for understanding and evaluating the regulation of gene expression, DNA replication, repair, and recombination (Chepelev et al., 2012). Moreover, using the Hi–C approach, the possibility of comparing the three-dimensional organization of the DNA in physiological and pathological conditions is achievable. The capability of describing how diseases reorganize the chromatin conformation to originate novel co-localized gene clusters of co-expression would be of primary importance.

To fully exploit the potential of this technique, many issues have to be faced. First of all the huge amount of data that should be produced for describing the conformation of the DNA in the nucleus. Considering that there are more than 200 different cell types with different profiles, which also change depending on the cell's actual state, the sequencing effort required to describe the three-dimensional configuration of genes in the nucleus is huge. Moreover, the integration of epigenetic information that is strictly correlated to the DNA conformation in the cell in a mutual cross-regulation (since the expression of proteins that organize the chromatin in the nucleus is correlated to the conformation of the chromatin itself), making the data problem explosive.

In this paper we describe a initial attempt to analyze Hi–C data and related multi-omic features using a network approach to represent gene co-localization and co-regulation. In particular, we describe how the R package NuChart, with its algorithmic features that have been previously presented (Merelli et al., 2013), can be used to interpret 3C data for creating a map that represents multi-omic information. Here, we present the possibilities that can be opened by using systems biology concepts for the analysis of 3C data, in particular highlighting how this procedure has the potential to enter into clinical practice, because it provides information that can be interpreted in a cytogenetic view, with incomparable resolution and richness of details.

MATERIALS AND METHODS

Inspired by web applications such as Google Maps, we developed NuChart (Merelli et al., 2013), an R package that elaborates Hi–C information to provide a systems biology oriented, gene-centric view of the three-dimensional organization of the DNA in the nucleus (the software, the manual, and all the supporting materials are freely available at ftp://fileserver.itb.cnr.it/nuchart). NuChart can be used to describe the DNA conformation in the neighborhood of selected genes by mapping on the achieved graph genomic features that are important for controlling gene expression at epigenetic level.

Although NuChart is the first R package allowing both visualization and analysis of Hi–C data in a gene-centric fashion [other software are CytoHi–C (Shavit and Lio', 2013) and Homer (Heinz et al., 2010), which both rely on Cytoscape], a similar approach was initially presented by Wang et al. (2013), for the analysis of chromatin conformation data in experiments concerning acute lymphoblastic leukemia (ALL) and Lymphoma cells. This work pioneered the idea of analyzing the social behavior of genes by using a graph-based approach. A similar method has been exploited in NuChart, which in addition allows a statistical interpretation of both expression and epigenetic data in comparison to the topology of the graph, thus allows a deep integration of this kind of information.

For example, it is possible to map on the neighborhood graph Linking Gene Regulatory Elements [in particular, the predicted binding sites for the CTCF or Cohesin proteins (Botta et al., 2011)], isochores [that describe the variations in the GC content and are important for the genome organization (Varriale and Bernardi, 2009)], potential cryptic Recombination Signal Sequences [cRSSs, which are important for generating the antigen receptor diversity (Marculescu et al., 2006)], and other user desired genomic features (using the bed file format), such as methylation profiles and histone modifications, to infer how epigenetic features and the three-dimensional nuclear organization of DNA cooperate in controlling gene expression. This can be very useful while studying the differentiation of stem cells or for identifying chromosomal reorganizations in cancer cells.

The package is built upon the functionality of Bioconductor packages such as biomaRt, Biostrings, ArrayExpress, GEOquery, KEGGREST, limma, samr, igraph, and ergm, providing a novel method to exploit Hi–C data in a systems biology context. NuChart, used in combination with the Hicup software, processes Hi–C data in FASTQ format, performs some preliminary normalizations relying on the fragment distances from the enzymatic cut sites. The output is a detailed table concerning the chromosomal spatial neighborhood of the input genes, providing a related graph on which it is possible to map multi-omic features.

The idea behind this package is to provide a complete suite of tools for the analysis of Hi–C data using a gene-centric point of view to provide a map on which other omic data can be mapped (see **Figure 1**). Contact matrices, or better their probabilistic models, allow to create representations that only involve two chromosomes, while we are able to describe the interactions of all the chromosomes together using a graph-based approach. This representation gives more importance to the physical proximity of genes in the nucleus in comparison to coordinate-based representations. This is the same problem that impairs representations based on Circos, which are able to characterize the whole genome in one shot, but fail to describe the physical proximity of genes.

A typical analysis performed with NuChart starts with the pre-processing of the FASTQ file using Hicup, which provides as output a SAM file (see the Hicup documentation for more details). Then, data can be loaded into the R environment and normalized using a generalized linear model relying on a Poisson distribution (taking into account Hi–C fragment length, mappability and GC-content). Considering that this normalization approach is well-established (Hu et al., 2012), the algorithm returns the same results of other approaches relying on the computation of the contact matrices (Servant et al., 2012; Seitan et al., 2013; Ay et al., 2014), providing a probability score at each edge of the neighborhood graph.

This method allows to estimate, at the same way of the contact maps, the probability that different genomic regions are proximal one to the other, with the advantage of allowing an iterative

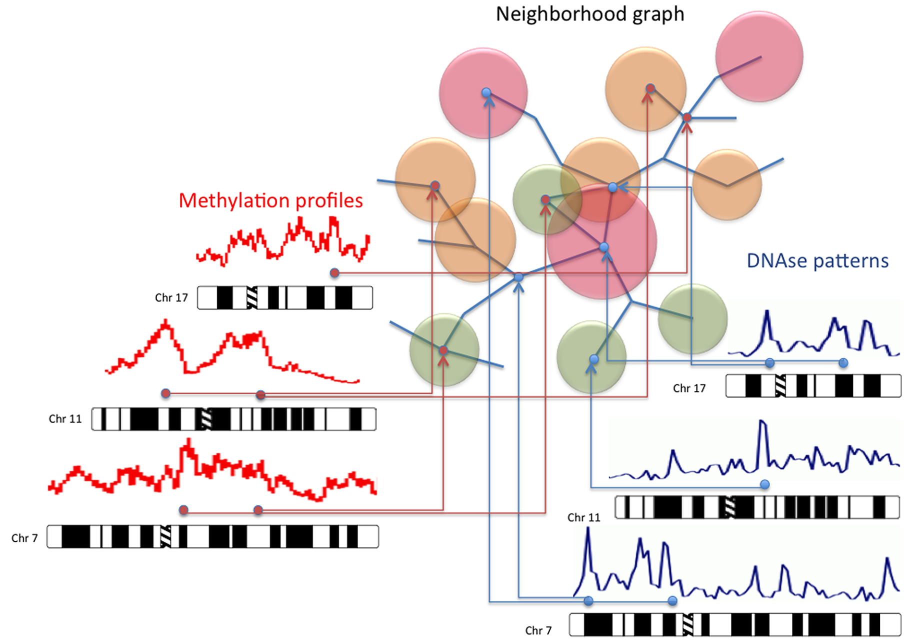


FIGURE 1 | Multi-omic integration of data using Hi–C maps. The possibility of employing a map describing the chromatin organization in the nucleus to represent multi-omic data can be very useful for the interpretation of this kind of information. This approach recall the use of Google maps to display, on a cartographic representation of a city, features such as shops and places of interest. By using the same idea, we can map on the neighborhood graph of the genes inside the nucleus omic data such as methylation patterns, immunoprecipitation information, open-chromatin conformations, and expression profiles.

analysis of the space: it is therefore possible to calculate the probability that two genes are distant a specific number of contacts. Moreover, the graph-based description of gene positions in the nucleus is extremely useful for mapping other multi-omic features, since analyzing data through this spatial description of the DNA conformation allows the identification of long-range interactions, cooperative genes and common epigenetic patterns, which are more difficult to identify relying on chromosomal coordinates.

The core procedure starts from one or more input genes from which a graph of adjacent genes is constructed. The identification of neighbor genes begins searching chromosome fragments that belong to the input genes. These fragments are then compared with other chromosome fragments located in a different genomic region, as reported by coupled reads. When a match is found and a new fragment is identified within a specific gene region, an edge between the starting gene and the novel detected one is created. A very important feature of the algorithm is the possibility to specify the number of iterations to accomplish for creating the neighborhood graph, which means to specify the maximum span that the graph can reach starting from the input genes (correlated to the diameter of the graph or, using the graph theory terminology, to the "longest shortest path").

By default this value is set to one, which means that, considering the list of genes given as input and taking into account the desired normalization, only genes that are directly in contact are mapped on the graph. If this parameter is set to two the procedure is iterated twice, meaning that all the genes identified in the neighborhood of the input genes at the first iteration of the algorithm are searched again for Hi–C interactions with other genes. And so on. This is of critical importance because it allows to overcome the limit of the contact matrix representation, which is limited by definition at representing only the interactions just one step away from the considered gene, while here we can identify paths also between distant genes.

The added value of this package is to provide the possibility of analyzing Hi–C data in a multi-omic context, by enabling the capability of mapping on the graph vertices expression data, according to a particular transcriptomic experiment, and on the edges genomic features that are known to be involved in chromosomal recombination, looping, and stability. If the user is interested in mapping on the neighborhood graph also gene expression data, there are functions for downloading microarray experiment results from ArrayExpress and GEO. Moreover, using NuChart it is possible to map on the neighborhood graph many genomic features such as data concerning cryptic RSSs, isochores, and CTCF binding sites, which are embedded in the package, but also any other omic information using the common bed file format.

NuChart also provides three functions to describe, compare, and statistically analyze neighborhood graphs once they have been created, which can be useful to highlight local and global characteristics of the fragment distribution in the context of the three-dimensional DNA topology inside the nucleus. In particular, there is the possibility to create general statistics about the graphs, which can be useful to describe physiological and pathological conditions of the cells, verifying the differences in the spatial distribution of genes. Then, neighborhood graphs can be compared by applying a conversion in adjacency matrices and then employing the Pearson correlation to check their similarity (for example to see intra and inter experiments variability).

The last set of functions available in NuChart enables the user to analyze, from a statistical point of view, the neighborhood graphs in relation to the mapped multi-omic features. In particular, these functions rely on the R package Exponential-family Random Graph Models (ERGMs) that provides an integrated set of tools for creating an estimator of the network through a stochastic modeling approach. In particular, the ERGM functions are able to extrapolate the salient characteristics of a network by implementing a maximum likelihood estimator.

Operatively, the software generates a huge number of networks, selects the ones having characteristics similar to the graph under analysis (i.e., degree distribution, connected components, topological conformation), and tries iteratively to optimize the generation parameters until all the created graphs have characteristics similar to those processed. This estimator is extremely useful, since it allows to create a probability distribution by which some peculiarities of the graph can be extrapolated, concerning both its intrinsic topology and specific attributes of the nodes (Admiraal and Handcock, 2007). In particular, the package allows to compute simple statistics about the topology of the graph, such as the significance of the vertex clustering attitude (triangle), or the significance of the network tendency to create multiple paths between two vertices (twopath). On the other hand, by choosing more complex modeling functions and exploiting the mapped multi-omic features, NuChart allows to test, for example, the probabilities that edges are a function of a specific genomic feature (nodecov) or the significance of having edges in relation to the absolute difference of a vertices' property (absdiff). The possibility of analyzing data to infer structural-activity relationships in a network is of critical importance (Reagans and McEvily, 2003).

RESULTS AND DISCUSSION

In this section we present some applications of the NuChart package. In particular, we show some interesting results relying on the possibility of creating metrics for defining how far two genes are one from the other, with possible applications to cytogenetic profiling, to the analysis of the DNA conformation in the proximity of the nucleolus, and for describing the social behavior of genes.

APPLICATION TO CYTOGENETIC

Applications of 3C techniques to cytogenetics are becoming very appealing, because the relative position of genes can be identified using high-throughput experiments. An example can be found in the work of Naumova et al. (2013), which concerns the analysis of the mitotic chromosome organization, while other studies showed how it is possible to identify translocations in Hi–C data (Rusk, 2014). Here we show how Hi–C can be used for diseased versus normal cells comparisons, with particular interest in leukemias, since it reproduces results achieved by Fluorescence *in situ* hybridization (FISH) experiments.

Although Hi–C is intended to estimate the contact frequencies between different genomic regions, there is a clear correlation with chromosomal translocations, since recombinations are largely influenced by the distance between fragments in which DNA breaks, necessary for translocations, occur. There are already many evidences in this sense (Meaburn et al., 2007; Engreitz et al., 2012; Shugay et al., 2012; Zhang et al., 2012; Kenter et al., 2013), which demonstrate how the physical distance plays a leading role for recombinations, in particular when the frequency of DNA breaks are physiological (while in cellular models where a high number of translocation are artificially induced the frequency becomes the dominant factor). Considering the association between contact frequencies and translocations, we think that a graph-based approach may be useful for data analysis from a recombination point of view. NuChart is capable of providing an immediate representation of genomic segments that are more likely to translocate with a specific gene, taking into account that the recombination probability is proportional to the weight of the connecting edges, according to the employed normalization.

The first example we present concerns the Philadelphia translocation, which is a specific chromosomal abnormality associated with chronic myelogenous leukemia (CML). The presence of this translocation is a highly sensitive test for CML, since 95% of people with CML have this abnormality, although occasionally it may occur in acute myelogenous leukemia (AML). The result of this translocation is that a fusion gene created from the juxtaposition of the ABL1 gene on chromosome 9 (region q34) to part of the BCR ("breakpoint cluster region") gene on chromosome 22 (region q11). This is a reciprocal translocation, creating an elongated chromosome 9 (called der 9), and a truncated chromosome 22 (called the Philadelphia chromosome).

Using NuChart we compared the distance of some couples of genes that are known to create translocation in CML/AML. In particular, our analysis relies on data from the experiments of Lieberman-Aiden et al. (2009), which consist in four lines of karyotypically normal human lymphoblastoid cell line (GM06990) sequenced with Illumina Genome Analyzer, compared with two lines of K562 cells, an erythroleukemia cell line with an aberrant karyotype. Starting from well-established data related to the cytogenetic experiments (Dewald, 2002), we tried to understand if the Hi–C technology, in combination with NuChart, can successfully be applied in this context, by verifying if translocations normally identified by using FISH can also be studied using 3C data. Therefore, we identified five couples of genes that are know to be involved in translocations and we compared their Hi–C interactions in physiological and diseased cells.

The very interesting result is that ABL1 and BCR, considered a normalization equivalent to the one achieved with HicNorm, are likely to be distant 1 or 2 contacts ($p < 0.05$) in sequencing runs concerning GM06990 with HindIII as digestion enzyme (SRA:SRR027956, SRA:SRR027957, SRA:SRR027958, SRA:SRR027959), while they are directly in contact ($p < 0.05$) in sequencing runs related to K562 with digestion enzyme HindIII (SRA:SRR027962 and SRA:SRR027963). Therefore, there is a perfect agreement between the positive and the negative presence of Hi–C interactions and FISH data (see **Figure 2**). At the same way, AML1 and ETO are in close proximity ($p < 0.05$) in leukemia cells

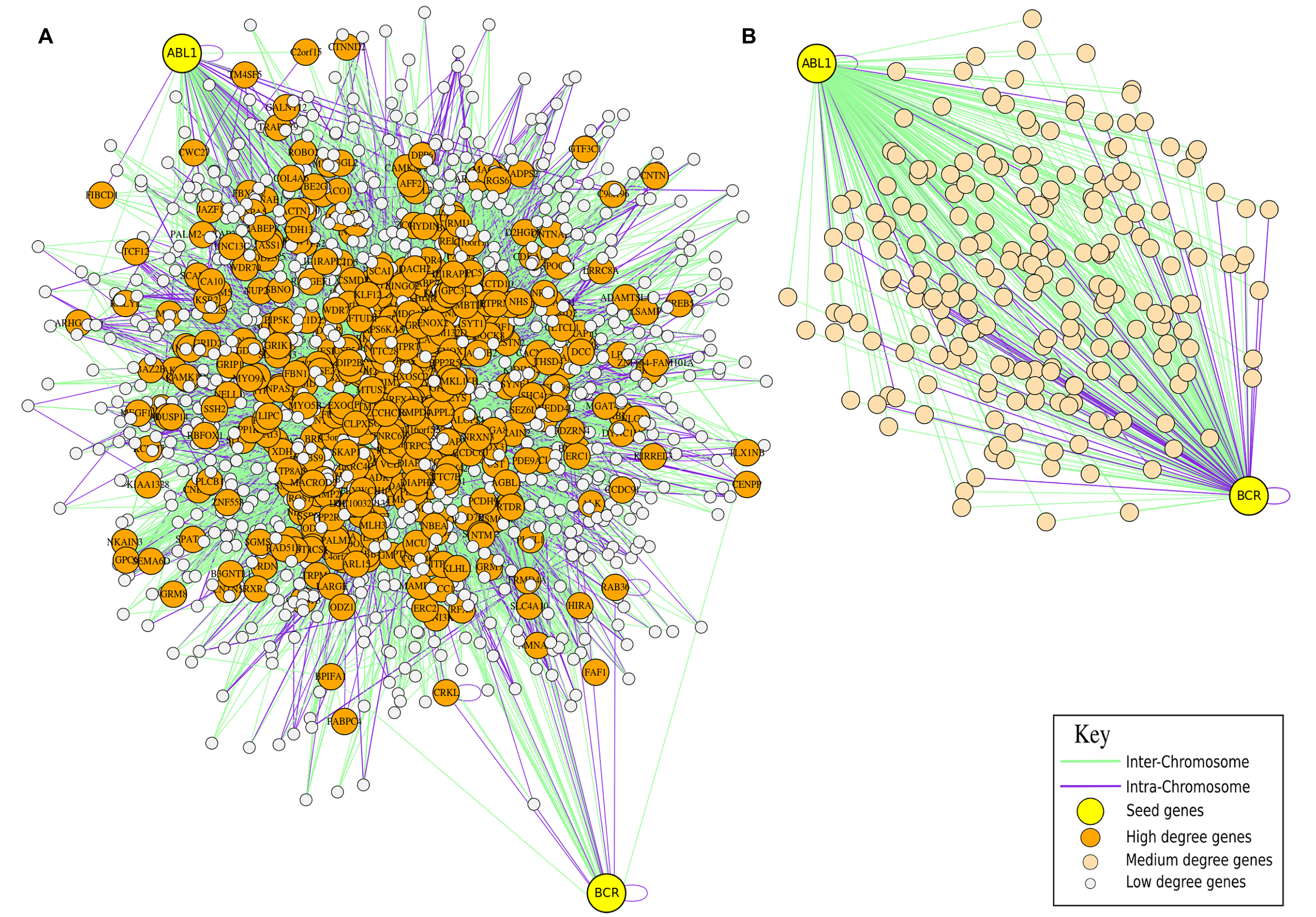


FIGURE 2 | Hi–C for Cytogenetic application. Comparison of FISH data and Hi–C information for the cytogenetic analysis of translocation in leukemia. In panel **(A)** the neighborhood graph of the genes ABL1 and BCR generated from the GM06990 cell line (Lieberman-Aiden experiment SRA:SRR027956) is presented. The graph shows that in normal cells the genes are not directly in contact since other genes are placed between them. In panel **(B)** the neighborhood graph of the genes ABL1 and BCR from the K562 cell line (Lieberman-Aiden experiment SRA:SRR027962) is presented. The graph shows that in cancer cells the genes are closer one to the other, which is a preliminary evidence of a possible translocation.

(SRA:SRR027962 and SRA:SRR027963), while they are likely to be far 2 or 3 contacts ($p < 0.05$) in normal cells (SRA:SRR027956, SRA:SRR027957, SRA:SRR027958, SRA:SRR027959). Considering the translocation CBFβ-MYH11, these genes are distant 2 or 3 contacts ($p < 0.05$) in GM06990 (SRA:SRR027956, SRA:SRR027957, SRA:SRR027958, SRA:SRR027959), while they are proximal with high probability ($p < 0.05$) in K562 (SRA:SRR027962, but not in SRA:SRR027963). We had no significant results for NUP214-DEK and PML-RARα translocations, which, however, are more rare in this kind of disease.

A second example of Hi–C cytogenetic application concerns the experiments of Wang et al. (2013) about B-cell ALL. Also in this disease there are well-characterized translocations, the most important of which is the TEL-AML1 fusion gene (Stams et al., 2005) that is present in about 25% of patients. This translocation of chromosome 12 (region q34) and chromosome 21 (region q22) results in the expression of chimeric transcription factors, which block both differentiation and apoptosis by interfering with the function of their wild-type counterparts.

As before, we employed NuChart to characterize the distance between some couples of genes in the cells' physiological and pathological state. In detail, we used the results of the 4 karyotypically normal human lymphoblastoid cell line (GM06990) from the experiments of Lieberman-Aiden as control data (as in the Wang's paper), while pathological profiles are directly taken from the experiments performed by Wang et al. (2013) (private communication). This dataset consists of 2 highly overlapping Hi–C experiments, the first concerning a case of primary human B-Cell ALL (B-ALL) and the second regarding the MHH-CALL-4 B-Cell ALL cell line (CALL4). Also in this case, starting from some well-established translocations, we tested the capability of the Hi–C technique, in combination with NuChart, to capture some genomic rearrangements usually identified using FISH.

The first result is that TEL and AML1, considered a normalization equivalent to the one achieved with HicNorm, are always distant 2 contacts ($p < 0.05$) in sequencing runs concerning GM06990 with HindIII as digestion enzyme (SRA:SRR027956, SRA:SRR027957, SRA:SRR027958, SRA:SRR027959), while they are directly in contact ($p < 0.05$) in sequencing runs related to B-ALL and CALL4. Other tests were performed on the E2A-PBX translocation: these genes are in close proximity ($p < 0.05$) in cancer cells (B-ALL and CALL4), while they are likely to be far 2 or 3 contacts ($p < 0.05$) in three out four control cell lines (SRA:SRR027956, SRA:SRR027958, SRA:SRR027959). Following the results discussed in the work of Taylor et al. (2013) we also tested the proximity of genes IGH and miR125b1 (related to a microRNA), which are distant 2 or 3 contacts ($p < 0.05$) in GM06990 (SRA:SRR027956, SRA:SRR027958, SRA:SRR027959), while they are proximal with high probability ($p < 0.05$) in leukemias cells (CALL4, but not in B-ALL, which presents a lower reads density). Considering the translocation BCR–ABL1, genes are distant 2 or 3 contacts ($p < 0.05$) in GM06990 (SRA:SRR027956, SRA:SRR027957, SRA:SRR027958, SRA:SRR027959), while they are proximal with high probability ($p < 0.05$) in leukemias cells (CALL4, but not in B-ALL, which presents a lower reads density). We had no results for the MLL and AF4 translocation.

These results are of significant importance, because with the decreasing of sequencing costs the Hi–C technique can be an effective diagnostic option for cytogenetic analysis, with the possibility of improving the knowledge regarding the correlation between the genome architecture and translocations. For example, Hi–C can be used to infer non trivial risk markers related to aberrant chromosomal conformation, like the Msc5a loci for breast cancer, which is known to play a critical role in the re-organization of a portion of chromosome 9 by CTCF proteins.

RNA POLYMERASES

In the following example, we discuss an interesting analysis regarding the internal organization of the DNA in the nucleus, working on the data produced in the Dixon et al. (2012) experiments. The intention is to show the different chromosomal organizations that occur in the nucleolus, while gene expression is heavily characterizing the differentiation of stem cells, since this part of the nucleus is involved in the transcription of ribosomal RNA (rRNA) subunits and in their combination with proteins to form complete ribosomes. Therefore, at the border of the nucleolus are exposed transcriptional units ready to express genes, and it would be very useful to understand the organization of these structures in relation to genomic regions that are going to be transcribed.

For this reason, we performed an Hi–C analysis of some specific subunits of the RNA Polymerase I (that only transcribes rRNA), RNA Polymerase II (directly involved in microRNA and gene expression), and RNA Polymerase III (mainly required to express tRNA) to shed light in their different configurations in different cell types. While most of the subunits are shared, some are peculiar of a particular RNA Polymerase and we choose to use these subunits to verify if there is correlation between their position in the nucleus and their activities. Respectively, the neighborhood graphs have been produced according to two different sequencing runs performed on human embryonic stem cells (SRA:SRR400260 and SRA:SRR400261), and from human lung embryonic fibroblast (SRA:SRR400266 and SRA:SRR400267) of Dixon et al. (2012) experiments.

In **Figure 3** a detailed representation of the different RNA Polymerase II neighborhood graphs is shown. In particular, these graphs show the neighborhood of the POLR2A gene that encodes for RPB1 (Strachan and Read, 1999), the largest subunit of the RNA polymerase II, which catalyzes the transcription of DNA to synthesize precursors of mRNA, most snRNA and microRNA, in the different cell lines. The representation shows that there are a wide range of genes involved in cell differentiation, with an enrichment of genes related to the cell cycle process (such as CDC45 and CCNE1, CCNB1) and many transcription factors (such as EBF1, TFEC, TFAP2A, TFB1M). Concerning POLR1A, that encodes for the A190 protein of the RNA Polymerase I, in the different experiments, as expected, it has in its neighborhood genes that are correlated to the rRNA subunits, such as RPL31, MPRS5, MRPS9, MRPS24, MRPS27, and MRPL35. Regarding POLR3B, which encodes for the subunit C128 of the RNA Polymerase III, we found in its neighborhood a couple of genes related to tRNA, in particular TRNAD1 (transfer RNA aspartic acid 1 – anticodon GUC) and TRNAS26 (transfer RNA serine 26 – anticodon AGA).

Considering the variability in the neighborhood of these genes, computed as correlation between lists of adjacent genes, there is a wide changeability looking at the RNA Polymerase II, while the differences considering RNA Polymerase I and III are considerably smaller. In particular, the similarity between two different runs of sequencing performed on the same cell type is relatively high for DNA Polymerase II (respectively, 60 and 67%), while there are very important differences between the two cell lines (correlation below 30%), which witnesses the importance (and the variability) that chromosomal re-organizations have at the nucleus/nucleolus level for co-expression. Considering DNA Polymerase I and III, there is a high reproducibility for runs performed on the same samples (respectively, 85 and 87% for POLR1A and 80 and 83% for POLR3B) and a relative increase in the analyses performed in different cell lines (correlation around the 40%). This kind of analysis is very important for understanding, in a particular moment, what the cells are going to express by reorganizing their chromosomal structure in the three-dimensional space of the nucleus.

NETWORK MODELING

The power of NuChart relies on the capability of capturing and describing the co-localization and co-activation of single entities in a gene network, exploiting a systems biology approach. Moreover, the interaction of the actor genes with the environment is of critical importance for understanding the entire system. This can be performed using the modeling functions of the package, which allow to statistically characterize the distribution of the edges in relation to the characteristics of the nodes that are the mapped multi-omic features. In order to show the possibilities of NuChart in terms of statistical inference on the graph, we performed the analysis of the clusters of genes Kruppel-associated box (KRAB; **Figure 4**) and human leukocyte antigen (HLA; **Figure 5**) in the context of four Dixon et al. (2012) experiments to verify the correlation of the edge distribution in relation to some genomic features

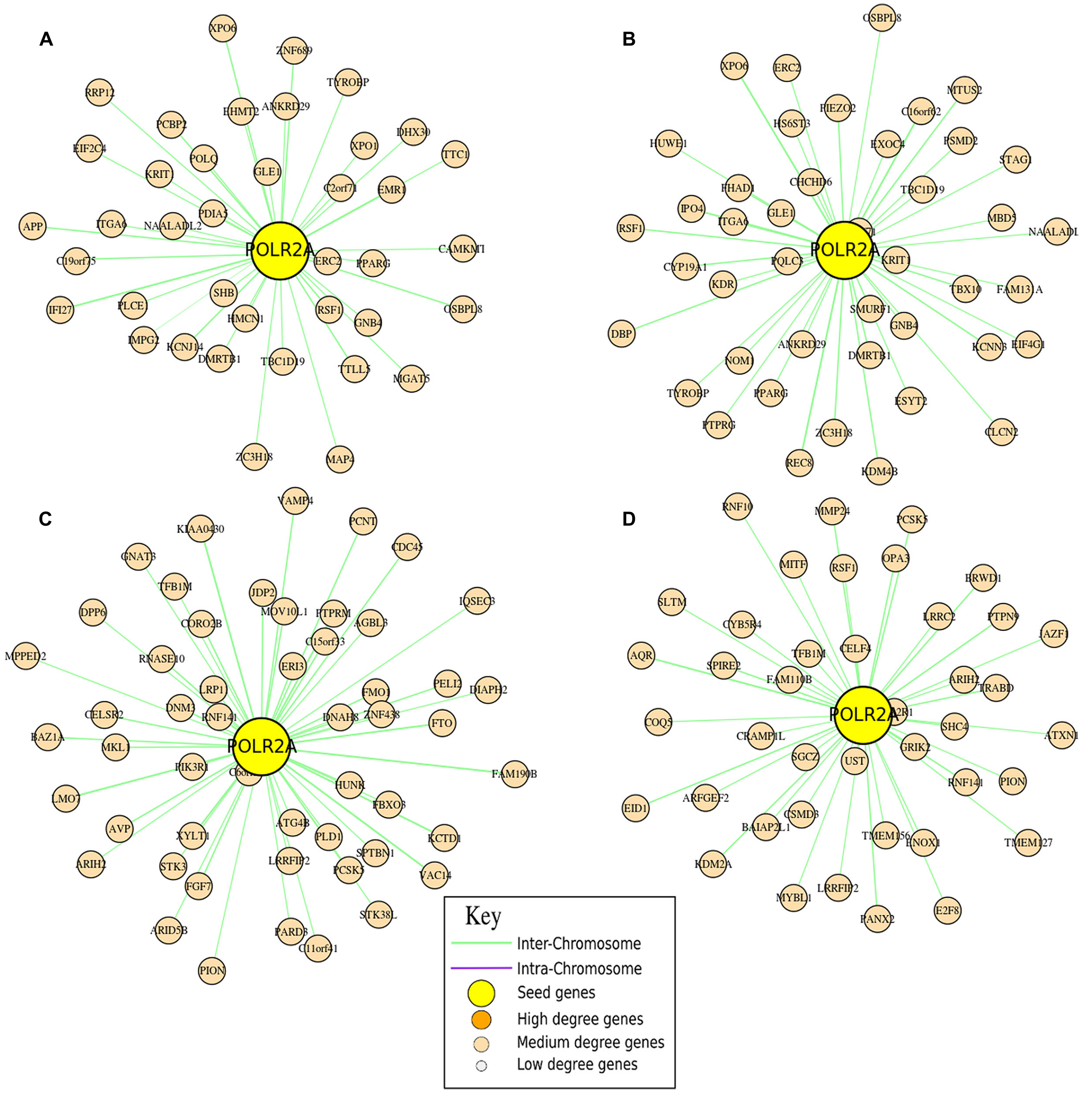


FIGURE 3 | Analysis of the nucleolus. Neighborhood graphs of the gene POLR2A in four different runs from the Hi–C experiments of Dixon et al. (2012) to show inter and intra run modifications of the chromatin conformation. Panels **(A,B)** concern sequencing runs from hESC (SRA:SRR400260, SRA:SRR400261), while data in panels **(C,D)** are from IMR90 (SRA: SRR400266, SRA:SRR400267). In these graphs, seed genes are the genes given as input to the algorithm, while output genes are differentially represented according to their importance (in terms of degree).

(hypersensitive sites, CTCF binding sites, isochores, RSSs), whose data are embedded in the NuChart package.

The first analyzed locus is located in cytoband chr19.q13.12 and concerns the clusters of Kruppel-type zinc finger genes, related to the KRAB, that are distinctive for their tandem organization (Huntley et al., 2006). Zinc finger proteins are a family of transcription factors that regulate the gene expression, and most of these proteins are members of the KZNF family. There are seven human-specific novel KZNFs and 10 KZNFs that have undergone pseudo-gene transformation specifically in the human lineage. 30 additional KZNFs have experienced human-specific sequence changes that are presumed to be of functional significance. Members of the KZNF family are often in regions of segmental duplications, and multiple KZNFs have undergone human-specific duplications and inversions.

The second analyzed gene cluster concerns the HLA system, which is the name of the locus containing the genes that encode for major histocompatibility complex (MHC) in humans. The proteins encoded by these genes are also known as antigens, as a result of their historic discovery as factors in organ transplants. The HLA belongs to a super-locus that contains a large number of genes related to the immune system function in

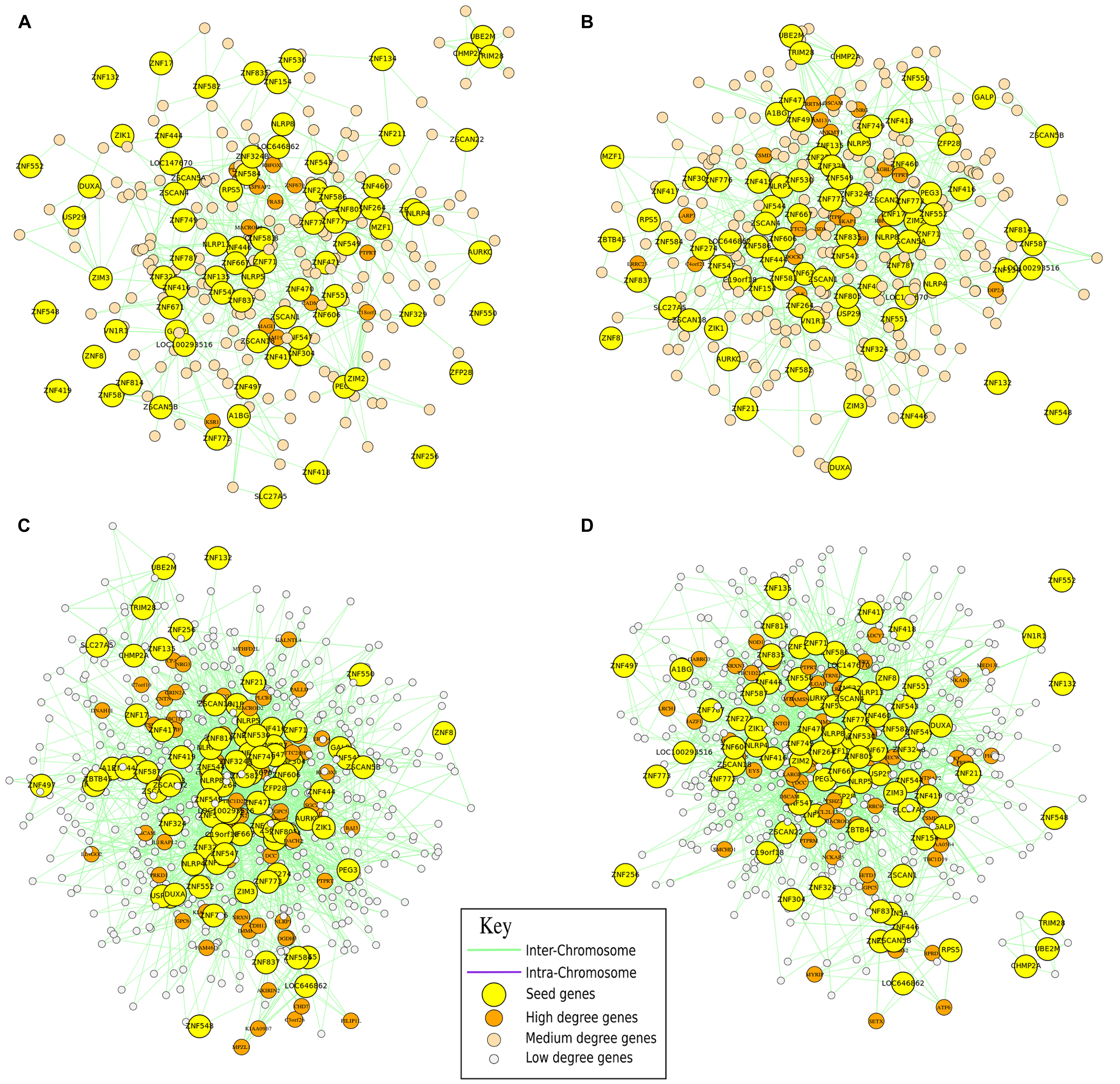


FIGURE 4 | Gene clustering analysis. Neighborhood graphs of the Kruppel-type zinc finger cluster of genes (cytoband chr19.q13.12), related to the Kruppel-associated box (KRAB), in four different runs from the Hi–C experiments of Dixon et al. (2012) to show inter and intra run modifications of the chromatin conformation. Panels **(A,B)** concern sequencing runs from hESC (SRA:SRR400260, SRA:SRR400261), while data in panels **(C,D)** are from IMR90 (SRA: SRR400266, SRA:SRR400267). In these graphs, seed genes are the genes given as input to the algorithm, while output genes are differentially represented according to their importance (in terms of degree).

humans. In particular, this group of genes resides on cytoband chr6.p31.21 and encodes for cell-surface antigen-presenting proteins, which have many different functions. Primarily, the HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria.

These statistical results are quite intriguing to analyze (**Table 1**). From one side, the correlation between the presence of CTCF binding sites and edges was predictable since Linking Gene Regulatory Elements are demanded to keep different regions of the genome close to each other, but is very interesting to quantify this association. On the other hand, regions with isochores seem less involved in long-range interactions, which can be quite surprising considering that these portions of the genome are considered gene-rich. The correlation between cryptic RSS sites and edges is more pronounced in the HLA cluster in comparison to the KRAB

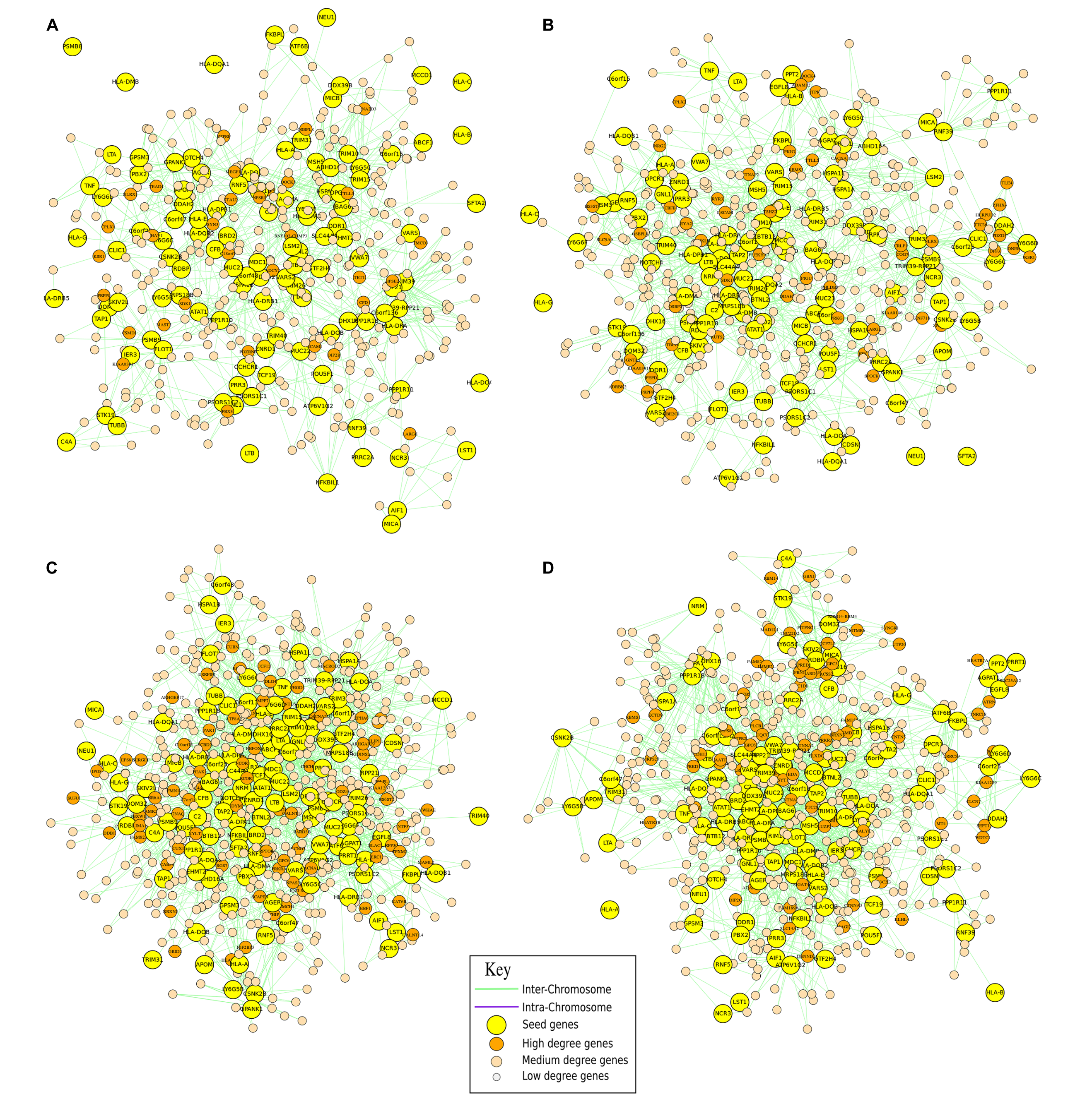


FIGURE 5 | Gene clustering analysis. Neighborhood graphs of the human leukocyte antigen (HLA) cluster of genes (cytoband chr6.p31.21) in four different runs from the Hi–C experiments of Dixon et al. (2012) to show inter and intra run modifications of the chromatin conformation. Panels **(A,B)** concern sequencing runs from hESC (SRA:SRR400260, SRA:SRR400261), while data in **(C,D)** are from IMR90 (SRA: SRR400266, SRA:SRR400267). In these graphs, seed genes are the genes given as input to the algorithm, while output genes are differentially represented according to their importance (in terms of degree).

cluster, probably due to a more consistent presence of this kind of sequences in genes related to the immune system. Finally, the correlation between hypersensitive sites (super sensitivity to cleavage by DNase) and edges, although positive, is poor, probably because the accessibility of these regions are impaired by a large number of long-range interactions.

CONCLUSION

The integration and visualization of omic data is a critical issue and they really represent challenges for scientists that work on Big Data paradigms in the 21st century. Tools to integrate a cascade of multi-omic data with the information about the structure of the nucleus require a cartographic approach such as Google

Table 1 | Analyses of CTCF binding sites, isochores, cryptic RSSs, and hypersensitive sites (super sensitivity to cleavage by DNase) impact on the edge distribution of the KRAB cluster of genes and of the HLA cluster of genes.

	KRAB		HLA	
	Estimate	**SE**	**Estimate**	**SE**
SRA:SRR400260				
Edges + nodecov("dnase")	0.2867	0.08451	0.1751	0.07961
Edges + nodecov("ctcf")	0.6531	0.01157	0.5845	0.01253
Edges + nodecov("rss")	0.5804	0.06176	0.6304	0.08196
Edges + nodecov("iso")	−1.0470	0.09269	−0.9406	0.09156
SRA:SRR400261				
Edges + nodecov("dnase")	0.2042	0.06782	0.1706	0.08022
Edges + nodecov("ctcf")	0.6629	0.04158	0.6287	0.03225
Edges + nodecov("rss")	0.5378	0.03566	0.6419	0.03776
Edges + nodecov("iso")	−1.0151	0.09566	−0.9335	0.08969
SRA:SRR400266				
Edges + nodecov("dnase")	0.3042	0.05962	0.1818	0.07822
Edges + nodecov("ctcf")	0.6738	0.03744	0.5678	0.02113
Edges + nodecov("rss")	0.5569	0.02996	0.6617	0.03776
Edges + nodecov("iso")	−1.1000	0.09655	−0.8305	0.08969
SRA:SRR400267				
Edges + nodecov("dnase")	0.3272	0.07932	0.1901	0.05925
Edges + nodecov("ctcf")	0.6645	0.04158	0.4677	0.02005
Edges + nodecov("rss")	0.5378	0.02755	0.6520	0.03883
Edges + nodecov("iso")	−0.9501	0.09076	−0.8707	0.09050

SE, Standard Error.

It's very interesting to highlight the high similarities between the four sequencing runs. In particular, data demonstrates that CTCF binding sites and cryptic RSSs have a positive influence on the presence of edges. At the same way DNase hypersensitive sites are positively correlated with edges although with less impact, while isochores are negatively correlated with the edge distribution.

maps, because genome browsers only work at the coordinate level, discarding long-range interactions and associations.

Changing the point of view into a more systems biology fashion, we think that the information about the chromatin organization may also be the key to interpret this multi-omic cascade of data, since they are capable of providing genetic maps to make clearer the collective behavior of genes. The cooperation among genes can probably be better interpreted using tools that are typical of the social network era and the possibility to use tools like NuChart supports this concept. In particular, the possibility of having suitable descriptions of how genes are localized in the nucleus, enriched by genomic features that can characterize the way they are capable of interacting, and combined with statistical analysis and semantic tools may result extremely useful in the years to come.

The interpretation of epigenetic features, genomic patterns, DNA binding sites, co-expression patterns could take an incredible advantage from the availability of distance matrices between genes, which can provide a measure of their correlation. Vice versa, due to the close connection between the three-dimensional organization of the DNA in the nucleus and the multi-omic features that regulate the cellular machinery, distance information can provide new hints about clusters of genes that cooperate under the control of the same transcription factors for specific biological processes.

ACKNOWLEDGMENTS

We thank John Hatton of Institute for Biomedical Technologies (CNR-ITB) for proofreading the manuscript. This work has been supported by the Italian Ministry of Education and Research (MIUR) through the Flagship (PB05) InterOmics, HIRMA (RBAP11YS7K), and the European MIMOMICS projects.

REFERENCES

Admiraal, R., and Handcock, M. S. (2007). Networksis: a package to simulate bipartite graphs with fixed marginals through sequential importance sampling. *J. Stat. Softw.* 24, 1–21.

Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* 24, 999–1011. doi: 10.1101/gr.160374.113

Botta, M., Haider, S., Leung, I. X., Liò, P., and Mozziconacci, J. (2011). Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.* 6, 426. doi: 10.1038/msb.2010.79

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res* 22, 490–503. doi: 10.1038/cr.2012.15

Dewald, G. W. (2002). Cytogenetic and FISH studies in myelodysplasia, acute myeloid leukemia, chronic lymphocytic leukemia and lymphoma. *Int. J. Hematol.* 76(Suppl. 2), 65–74. doi: 10.1007/BF03165090

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26, 11–24. doi: 10.1101/gad.179804.111

Di Stefano, M., Rosa, A., Belcastro, V., di Bernardo, D., and Micheletti, C. (2013). Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19. *PLoS Comput. Biol.* 9:e1003019. doi: 10.1371/journal.pcbi.1003019

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380. doi: 10.1038/nature11082

Duan, Z., Andronescu, M., Schultz, K., Lee, C., Shendure, J., Fields, S., et al. (2012). A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods* 58, 277–288. doi: 10.1016/j.ymeth.2012.06.018

Engreitz, J. M., Agarwala, V., and Mirny, L. A. (2012). Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS ONE* 7:e44196. doi: 10.1371/journal.pone.0044196

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 462, 58–64. doi: 10.1038/nature08497

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 28, 3131–3033. doi: 10.1093/bioinformatics/bts570

Huntley, S., Baggott, D. M., Hamilton, A. T., Tran-Gyamfi, M., Yang, S., Kim, J., et al. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.* 16, 669–677. doi: 10.1101/gr.4842106

Kenter, A. L., Wuerffel, R., Kumar, S., and Grigera, F. (2013). Genomic architecture may influence recurrent chromosomal translocation frequency in the Igh locus. *Front. Immunol.* 4:500. doi: 10.3389/fimmu.2013.00500

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., and Zheng, M. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98. doi: 10.1016/j.cell.2011.12.014

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369

Lin, Y. C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., et al. (2012). Global changes in the nuclear positioning of genes and intra- and inter-domain genomic interactions that orchestrate B cell fate. *Nat. Immunol.* 13, 1196–1204. doi: 10.1038/ni.2432

Marculescu, R., Vanura, K., Montpellier, B., Roulland, S., Le, T., Navarro, J. M., et al. (2006). Recombinase, chromosomal translocations and lymphoid neoplasia: targeting mistakes and repair failures. *DNA Repair.* 5, 1246–1258. doi: 10.1016/j.dnarep.2006.05.015

Meaburn, K. J., Misteli, T., and Soutoglou, E. (2007). Spatial genome organization in the formation of chromosomal translocations. *Semin. Cancer Biol.* 17, 80–90. doi: 10.1016/j.semcancer.2006.10.008

Merelli, I., Liò, P., and Milanesi, L. (2013). NuChart: chromosomal spatial neighbourhood and multi-omics annotation. *PLoS ONE* 8:e75146. doi: 10.1371/journal.pone.0075146

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., et al. (2013). Organization of the mitotic chromosome. *Science* 342, 948–953. doi: 10.1126/science.1236083

Papantonis, A., Kohro, T., Baboo, S., Larkin, J. D., Deng, B., Short, P., et al. (2012). TNFα signals through specialized factories where responsive coding and miRNA genes are transcribed. *EMBO J.* 31, 4404–4414. doi: 10.1038/emboj.2012.288

Reagans, R., and McEvily, B. (2003). Network structure and knowledge transfer: the effects of cohesion and range. *Adm. Sci. Q.* 48, 240–267. doi: 10.2307/3556658

Rusk, N. (2014). Genomics: genomes in 3D improve one-dimensional assemblies. *Nat. Methods* 11, 5. doi: 10.1038/nmeth.2795

Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P. P., Lajoie, B. R., Ing-Simmons, E., et al. (2013). Cohesin based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.* 23, 2066–2077. doi: 10.1101/gr.161620.113

Servant, N., Lajoie, B. R., Nora, E. P., Giorgetti, L., Chen, C. J., Heard, E., et al. (2012). HiTC: exploration of highthroughput 'C' experiments. *Bioinformatics* 28, 2843–2844. doi: 10.1093/bioinformatics/bts521

Shavit, Y., and Lio', P. (2013). CytoHiC: a cytoscape plugin for visual comparison of Hi-C networks. *Bioinformatics* 29, 1206–1207. doi: 10.1093/bioinformatics/btt120

Shugay, M., de Mendíbil, I. O., Vizmanos, J. L., and Novo, F. J. (2012). Genomic hallmarks of genes involved in chromosomal translocations in hematological cancer. *PLoS Comput. Biol.* 8:e1002797. doi: 10.1371/journal.pcbi.1002797

Stams, W. A., den Boer, M. L., Beverloo, H. B., Meijerink, J. P., van Wering, E. R., Janka-Schaub, G. E., et al. (2005). Expression levels of TEL, AML1, and the fusion products TEL-AML1 and AML1-TEL versus drug sensitivity and clinical outcome in t(12;21)-positive pediatric acute lymphoblastic leukemia. *Clin. Cancer Res.* 11, 2974–2980. doi: 10.1158/1078-0432.CCR-04-1829

Strachan, T., and Read, A. P. (1999). *Human Molecular Genetics*, Section 7.2. New York: Garland Science

Taylor, K. H., Briley, A., Wang, Z., Cheng, J., Shi, H., and Caldwell, C. W. (2013). Aberrant epigenetic gene regulation in lymphoid malignancies. *Semin Hematol.* 50, 38–47. doi: 10.1053/j.seminhematol.2013.01.003

Varriale, A., and Bernardi, G. (2009). Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* 95, 25–28. doi: 10.1016/j.ygeno.2009.09.006

Wang, Z., Cao, R., Taylor, K., Briley, A., Caldwell, C., Cheng, J., et al. (2013). The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PLoS ONE* 8:e58793. doi: 10.1371/journal.pone.0058793

Zhang, Y., McCord, R. P., Ho, Y. J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C., et al. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* 148, 908–921. doi: 10.1016/j.cell.2012.02.002

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

Received: 30 September 2014; accepted: 27 January 2015; published online: 11 February 2015.

Citation: Merelli I, Tordini F, Drocco M, Aldinucci M, Liò P and Milanesi L (2015) Integrating multi-omic features exploiting Chromosome Conformation Capture data. Front. Genet. ***6****:40. doi: 10.3389/fgene.2015.00040*

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

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

REVIEW ARTICLE
published: 29 July 2014
doi: 10.3389/fcell.2014.00031

Computational modeling of heterogeneity and function of CD4+ T cells

Adria Carbo[1,2], Raquel Hontecillas[1,2], Tricity Andrew[1,2], Kristin Eden[1,2], Yongguo Mei[1,2], Stefan Hoops[2] and Josep Bassaganya-Riera[1,2,3]*

[1] Nutritional Immunology and Molecular Medicine Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA
[2] Center for Modeling Immunity to Enteric Pathogens, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA
[3] Department of Biomedical Sciences and Pathobiology, Virginia-Maryland Regional College of Veterinary Medicine, Virginia Tech, Blacksburg, VA, USA

Edited by:
Paolo Tieri, Consiglio Nazionale delle Ricerche, Italy

Reviewed by:
Dougu Nam, UNIST, South Korea
William OConnor, Albany Medical College, USA

***Correspondence:**
Josep Bassaganya-Riera, Nutritional Immunology and Molecular Medicine Laboratory, Virginia Bioinformatics Institute, Virginia Tech University, 1015 Life Science Circle, Blacksburg, VA 24061, USA
e-mail: jbassaga@vt.edu

The immune system is composed of many different cell types and hundreds of intersecting molecular pathways and signals. This large biological complexity requires coordination between distinct pro-inflammatory and regulatory cell subsets to respond to infection while maintaining tissue homeostasis. CD4+ T cells play a central role in orchestrating immune responses and in maintaining a balance between pro- and anti- inflammatory responses. This tight balance between regulatory and effector reactions depends on the ability of CD4+ T cells to modulate distinct pathways within large molecular networks, since dysregulated CD4+ T cell responses may result in chronic inflammatory and autoimmune diseases. The CD4+ T cell differentiation process comprises an intricate interplay between cytokines, their receptors, adaptor molecules, signaling cascades and transcription factors that help delineate cell fate and function. Computational modeling can help to describe, simulate, analyze, and predict some of the behaviors in this complicated differentiation network. This review provides a comprehensive overview of existing computational immunology methods as well as novel strategies used to model immune responses with a particular focus on CD4+ T cell differentiation.

Keywords: CD4+ T cell differentiation, computational modeling, immunoinformatics, computational immunology, CD4+ T cell dogma

INTRODUCTION

The human immune system consists of two main behavioral and functional waves: first, the innate immune response provides a first barrier against foreign elements and second, the adaptive immune system builds an effective and specific immune response to combat such elements. The principal function of the adaptive responses is not only the specific recognition to foreign antigens, but also the formation of immunologic memory, and the development of tolerance to self-antigens (Luckheeram et al., 2012). Originated in the bone marrow and matured in the thymus, CD4+ T cells are part of the specific adaptive immunity compartment. T cell selection in the thymus allows creating an array of T cell repertoire for antigen recognition, as well as allowing the selection process through MHC-II and the expression of surface markers, such as CD4 or CD8 (Klein et al., 2009). Mature CD4+ T cells then translocate into the secondary lymphoid organs, such as the lymph nodes or the spleen, where they are involved in immune surveillance through interaction with MHC-II molecules expressed on the surface of antigen-presenting cells (Drayton et al., 2006). In this inductive site, naïve CD4+ T cells sample the tissue environment and depending on the cytokine milieu, they differentiate into functionally distinct regulatory or effector subsets.

The central dogma of CD4+ T cell differentiation has evolved over the past decades as new studies have unveiled differentiation pathways and novel mechanisms shaping the CD4+ T cell compartment. The Th1 vs. Th2 conceptual framework that Mossman and Coffman provided (Mosmann and Coffman, 1989) was largely expanded when novel discoveries on RORγt and IL-17A producing T cells defined the Th17 phenotype (Ivanov et al., 2006) and with the identification of FOXP3 raised as a key transcription factor in charge of driving the regulatory response in CD4+ T cells (Fontenot et al., 2003; Hori et al., 2003). Recent in-depth characterization of CD4+ T cell lineages has resulted in the discovery of new phenotypes, positioning the CD4+ T cell population as one of the most heterogeneous immune cell subsets. Furthermore, the latest discoveries are pushing the understanding of CD4+ T cell differentiation from a 4-player game to a multi-pronged interplay of complex networks and common transcription factors and cytokines with highly plastic functionalities. As an example, the production of IL-9 by the transcription factor PU.1 leads to the establishment of the Th9 phenotype (Ma et al., 2010). Furthermore, other phenotypes, such as Th17, are now under scrutiny since IL-17 and IL-22 are co-expressed in an IL-23 dependent manner (Trifari and Spits, 2010; Sonnenberg et al., 2011). New studies are pointing out to the aryl hydrocarbon receptor (AhR) as the master transcription factor responsible for IL-22 secretion (Ramirez et al., 2010), leading to the designation of a new CD4+ T cell phenotype, Th22, which has been also identified in humans (Eyerich et al., 2009; Fujita et al.,

2009) Moreover, FOXP3-independent IL-10 upregulation has been implicated in the activation of the regulatory axis under the regulatory type 1 (Tr1) CD4+ T cells (Pot et al., 2011). Lastly, follicular T helper cells (Tfh) have become an object of intense study since they have been described as a very plastic subset that could swift the CD4+ T cell balance. Tfh cells can leave the T cell areas and localize in the B cell follicle, a migration that is facilitated by their concurrent expression of the B cell zone homing chemokine receptor CXCR5 and downregulation of the T cell zone homing chemokine receptor CCR7 (Ansel et al., 1999; Hardtke et al., 2005). Thus, this close proximity to B cells allows Tfh cells to support their activation, expansion and differentiation. To help promote this crosstalk with B cells, Tfh cells produce IL-21 via activation of the transcription factor BCL-6, thereby promoting a Th1/Th17 profile. Also, IL-2 is emerging as a trigger for Th1 differentiated cells to adopt a Tfh-like phenotype by downregulating BLIMP1 and interacting with STAT proteins (Breitfeld et al., 2000). Since the BCL-6 pathway is linked to STAT factors induced by IL-6 that in turn promotes IL-21 and TNFα production, the study of the role of Tfh is important in the context of infectious, immune-mediated, or chronic inflammatory diseases.

Computational modeling has become an indispensable tool to synthesize, organize, and integrate diverse data types and theoretical frameworks to help generate new knowledge and guide *in vivo* experimentation. This review highlights how computational modeling has helped advancing the understanding of signaling events controlling CD4+ T heterogeneity and it also discusses new opportunities in the context of modeling strategies and tools.

MATHEMATICAL MODELING AND CD4+ T CELL DIFFERENTIATION

Initial attempts to apply computational modeling approaches to study CD4+ T cell differentiation only focused on the Th1 and Th2 phenotypes. Indeed the well-established dichotomy between these two phenotypes is supported by extensive information on how T-bet (Th1) and GATA3 (Th2) interact. One of the first published studies extrapolated the Th1/Th2 experimental facts into systemic behavior during an immune response, indicating that suppression and domination of one phenotype over the other could dictate the final differentiation outcome (Fishman and Perelson, 1999). In this study, the model encompassed not only Th1 and Th2, but also the effect of antigen presentation via APCs. This mathematical model illustrated how the final differentiation of Th1 or Th2 depends in both the competition for antigenic stimulation and the cytokine-mediated cross suppression between phenotypes. Subsequent studies applied mathematical modeling to study the Th1 and Th2 phenotypes in the presence of other cytokines such as IL-10 or TGFβ (Yates et al., 2000), antigen availability and instructional intracellular feedbacks (Bergmann et al., 2001, 2002), upregulation of the master transcription factors T-bet and GATA3 (Mariani et al., 2004; Yates et al., 2004) or in the context of cancer and rejection of melanomas (Eftimie et al., 2010). These modeling efforts highlighted the differences between instructive and feedback mechanisms as well as activated pathways in both phenotypes. Other studies solely focused on a single phenotype, such as the work published by Schulz et al. (2009) where the computational model revealed that Th1 differentiation is a two-step process in which the early Th1 cell-polarizing phase is followed by a later phase showing expression of T-bet. Hofer et al. (2002) published a mathematical model showing that GATA-3 transcriptional activation creates a threshold for autoactivation, resulting in two GATA-3 expression states: one for basal expression and one of high expression sustained by its autoactivation.

As new data became available, the increasing complexity of the CD4+ T cell paradigm became evident and new computational approaches were developed to ascertain the regulatory mechanisms controlling differentiation, plasticity, and heterogeneity. van den Ham and de Boer (2008) developed an ODE-based model that describes important regulators and allows for stable switches between several different phenotypes. Other studies focused on the interaction of Th17 and iTreg since Bettelli et al. (2006) described the functional antagonism of Th17 and iTreg. For instance, Hong et al. (2011) constructed a mathematical model of Th17/Treg differentiation that exhibited functionally distinct states, including a RORγt+ FOXP3+. While reductionist approaches have improved our ability to understand small components of the system, studying CD4+ T cell heterogeneity often requires implementing systems approaches and computational methods that can help deciphering complexity. Computational models of CD4+ T cell differentiation and heterogeneity are needed to accurately represent how CD4+ T cells are differentiated and accurately predict sensitivities to determine which pathways and molecules can be most critical to switch from one phenotype to another. A major challenge in systems-level models is the calibration process. Estimation of parameters of large-scale CD4+ T cell differentiation models have proven successful (Carbo et al., 2013b) by following a "divide-and-conquer approach." This approach is highly useful when parameterizing large models with more than one parameter estimation. First the parameter calibration is divided into smaller parameter estimations: one estimation per phenotype represented in the model. If necessary, other parameter estimations involving specific interactions, such as the Th1/Th2 or the Th17/Treg crosstalk, can be performed. Once parameters are located in a more targeted parameter space, a global parameter estimation is run with all the parameters in the model, allowing us to identify a good global parameter set. These approaches can be easily performed using modeling software such as COPASI (Hoops et al., 2006).

The CD4+ T cell differentiation model described in Carbo et al. (2013b) allows the user to have a global understanding with four CD4+ T cell phenotypes represented. The most recent systems biology markup language (SBML)-compliant network (Hucka et al., 2003) provides a structured understanding on different pathways involved in CD4+ T cell differentiation (**Figure 1**). SBML-based models are indeed highly portable between different simulation platforms. Of note, SBML-based topologies allow standardization in the modeling community and promote cross-transfer of several computational models in an efficient manner. The SBML standards are an essential step toward integrating an ensemble of distributed immunological models (within cells, between cells, at the cell population level, tissue-level, whole organism and human populations).

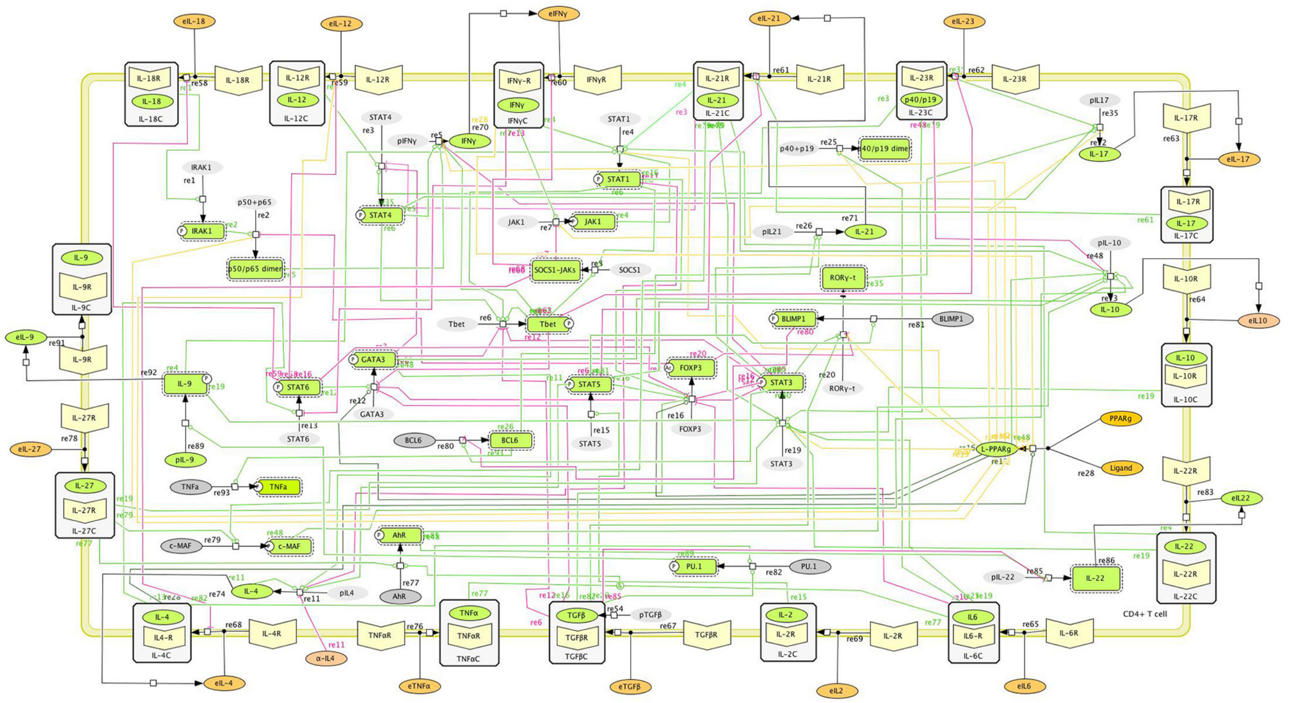


FIGURE 1 | Main intracellular differentiation pathways of a single CD4+ T cell. Systems Biology Markup Language (SBML)-compliant network model of CD4+ T cell differentiation, including cytokines, receptors, and intracellular signaling pathways controlling CD4+ T cell fate and function.

Another example of CD4+ T cell modeling would be the model by Mendoza and Pardo (2010). In this model, a continuous dynamical system, in the form of a set of coupled ordinary differential equations, was used. Such strategy was then applied to a regulatory network of 36 nodes, representing four CD4+ T cell phenotypes (Th1, Th2, Th17, and Treg). Although this model creates a framework for four phenotypes, the calibration of this larger network, however, was not conducted with experimental data but with default parameters that enabled the differentiation of the four phenotypes, not taking in consideration if reactions occur in a rapid or slow fashion. In addition, the model was not SBML compliant.

Others have explored the contribution of different CD4+ T cell phenotypes to the modulation of immune responses toward *Helicobacter pylori* infection (Carbo et al., 2013a). This study aimed to provide new mechanistic insights on the dynamics of mucosal Th1, Th17, and Treg cells by using both an ODE- and agent-based (ABM) cellular model of the mucosal immune responses during *H. pylori* infection. Alternatively, the logical model strategy has also been used to explore CD4+ T cell differentiation (Mei et al., 2013b; Mendoza, 2013). Mendoza et al. applied either continuous or discrete dynamical systems, regulatory networks of Th1/Th2 or of a combination of different transcription factors adding Th17 and iTreg to represent different states. Even though network modeling has shown to be appropriate, as the production of high-dimensional experimental data is increasingly becoming available, other methods, such as ODE- or agent-based modeling, could help understanding the mechanisms of CD4+ T cell differentiation at the systems level (Hoops et al., 2006; Mei et al., 2012; Wendelsdorf et al., 2012).

DIVING INTO CD4+ T CELL LINEAGES: PHENOTYPE OR FUNCTION?

CD4+ T cells form a complex and highly specialized network, representing a major population implicated in mediating host protective and homeostatic responses. However, their excessive or uncontrolled accumulation can also represent a feature in different diseases such as Inflammatory Bowel Disease (IBD) (Abraham and Cho, 2009), Alzheimer's disease (Monsonego et al., 2013), multiple sclerosis (Chitnis, 2007), or allergic disease (Islam and Luster, 2012), among many others. Therefore, their function is closely guided by external signals that are captured from the environment. Also, CD4+ T cells orchestrate immune responses by modulating the function of other cell subsets, such as dendritic cells or macrophages, through secretion of an array of soluble factors, cytokines, and chemokines into the environment. The cytokine profile secreted by each CD4+ T cell will directly depend on which intracellular molecular pathways have been activated, which cytokines are released and how the priming of the single CD4+ T cell has occurred. As an example, IL-6 and TGFβ will activate the Th17 transcriptional machinery, mainly composed by RORγt, RORα, and the phosphorylated form of STAT3. These molecules will activate the transcription of IL-21 and IL-17 and will direct the cell into a Th17 phenotype. However, when a CD4+ T cell is located in an environment rich in TGFβ, lacking IL-6 or other pro-inflammatory cytokines, TGFβ will promote FOXP3 and the phosphorylated STAT5, resulting in the secretion of IL-10 and TGFβ that will activate the regulatory axis. This differentiation dichotomy also depends in part on the T-cell receptor (TCR) engagement and a co-stimulatory signal, frequently involving the CD28 receptor: two basic signals required for a full CD4+ differentiation process. Indeed, Miskov-Kizanov et al. showed how the duration of T cell stimulation through the TCR receptor is a critical determinant of cell date and plasticity by constructing a logic circuit model of TCR signaling pathways in CD4+ T cells (Miskov-Zivanov et al., 2013).

CD4+ T cells have a strong predisposition to certain programming and developmental programs enabled by the cytokine environment. However, in the context of disease, where plasticity between phenotypes appears to be the norm, rather than the exception, double positives, such as IFNγ/IL17A often appear in pathological states such as in the context of murine colitis, where the accumulation of IL-17A+ IFNγ+ seems to occur in an IL-23 dependent manner (Ahern et al., 2010). Indeed, IL-23 has been shown to drive CD4+ T helper cell populations into a pathogenic state capable to drive autoimmune population by using passive transfer studies (Langrish et al., 2005), pinpointing IL-23 as a critical player in CD4+ T cell pathogenicity. Moreover, several studies showed that IL-17A could potently induce type 2 diabetes (Arababadi et al., 2010; Jagannathan-Bogdan et al., 2011; Zeng et al., 2012) potentially by modulating the pathogenesis of insulin resistance induced by angiotensin II type 1 receptor (Ohshima et al., 2012) hence increasing the production of renal nitric oxide (Imanishi et al., 2013). Th17 also showed a pleiotropic functionality, since intestine IL-17A+ IL-10+ T cells were found in the small intestine following treatment with anti-CD3 antibody, known to induce an immunosuppressive environment (Esplugues et al., 2011). Furthermore, intestinal epithelial lesions were accentuated in IL-17A null mice (Yang et al., 2008). These implications support a theory, whereby CD4+ T cells are not defined by its inflammatory status but by the functions they accomplish after being exposed to the cytokine milieu. The CD4+ T cell compartment has been demonstrated to be governed, not only by phenotype, but also by function, therefore forcing the distinction between a stable T cell lineage and a T cell differentiation state. Indeed, the ability of a CD4+ T cell to choose a predetermined differentiation program has been shown to be more complex than expected. This determination seems to now bow down to a more functional approach, where CD4+ T cells are not determined by phenotype, but by function, as needed. The functionality of CD4+ T cells as a means of classifying and determining their operational status has already been discussed in O'Connor et al. (2010) and Basu et al. (2013). The traditional view on the CD4+ T cell dogma has now changed into a more comprehensive vision, where the innate immune compartment influences differentiation on CD4+ T cells and not only 2 or 4, but 8 known phenotypes are represented and new phenotypes or states are likely to emerge (**Figure 2**).

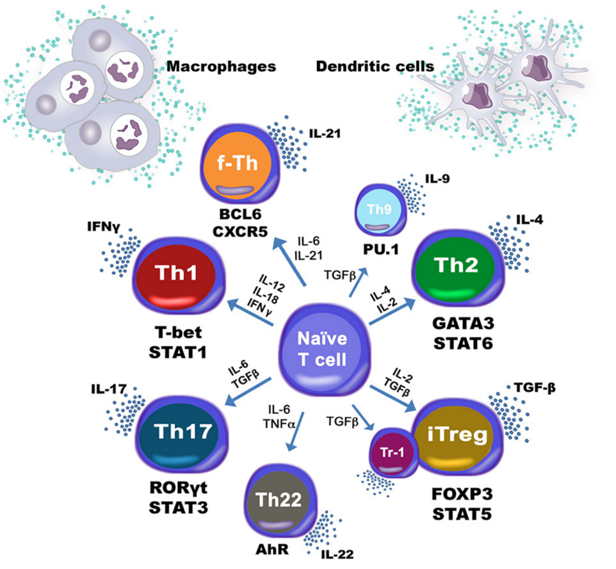


FIGURE 2 | Heterogeneity of CD4+ T cell subsets. T helper type 1 (Th1), type 2 (Th2), type 17 (Th17), type 9 (Th9) and type 22 (Th22), Follicular T helper cells (Tfh), and induced regulatory T cells (iTreg) as well as type 1 regulatory T cells (Tr1) are induced based on multiple cytokines being produced by dendritic cells and macrophages among other immune subsets.

DECIPHERING CD4+ T CELL PLASTICITY BY USING COMPUTATIONAL MODELING APPROACHES

Transcription factors, TCR, chemokines, surface receptors, and cytokines determine how CD4+ T cells become activated, maintained and how they can mature into distinguishable featured profiles. However, an increasing understanding on how the mechanisms of differentiation work is revealing increased flexibility and plasticity between different CD4+ T cell phenotypes that allow functional heterogeneity. As discussed above, the functional plasticity between Th1 and Th17 cells resulting in IFNγ+ IL-17A+ CD4+ T cells (Lee et al., 2009; Kurschus et al., 2010) has already been investigated. Indeed, Th17 has been shown to be a very unstable phenotype (Mathur et al., 2006). Functionally, Th17 cells during mucosal inflammation seem significantly different than those Th17 cells involved in regulating homeostasis at the steady state. Whereas IL-17A single positive Th17 cells produce IL-22, which may provide a mechanisms through which Tregs cells reinforce the epithelial barrier (Lin et al., 2014), this same Th17 population can accumulate and produce additional mediators such as IFNγ or GM-CSF during gut inflammatory disorders (Ahern et al., 2010; Codarri et al., 2011; El-Behi et al., 2011). CD4+ T cell plasticity is not only initiated by a change within the intracellular compartment, but also by a change in the extracellular environment. Th1 cells have been demonstrated to acquire plasticity toward a follicular T helper (Tfh)-like phenotype when they encounter a cytokine milieu that is not rich in IL-2 (Liao et al., 2011; Oestreich et al., 2012). Other studies also suggest that early Th1 differentiation is marked by a Tfh cell-like transition highlighting the role of Tbet and STAT4 in mediating these transitions (Nakayamada et al., 2011). The regulatory phenotype iTreg has also been reported to adopt plasticity mechanisms. Several studies have identified, for example, a double RORγt+ FOXP3+ (Lochner et al., 2008; Zhou et al., 2008) that can further differentiate into a pathogenic IL-17-expressing CD4+ T cell (Osorio et al., 2008). These examples illustrate the need for improving our mechanistic understanding at the systems level, where plasticity in the *in vivo* setting needs to be at focus.

Computational methods have also been applied to study the plasticity of CD4+ T cells. Magombedze et al. considered a population plasticity mechanism between Th1 and Th2 during *Mycobacterium avium* infection by using a reduced ODE-based model where the phenotype change of MAP-specific T cells occurred due to differences in the rates of differentiation, proliferation, and death at the site of infection (Magombedze et al., 2014). However, the cellular plasticity involving several intracellular pathways was not represented. In contrast, Pedicini et al. used computational models to analyze the cellular plasticity between Th1 and Th2 cells, extending the regular Tbet/GATA3 plasticity predictions to a broader panel of molecules, involving IRF4, STAT1 and STAT6, MAF, NFAT, and SOCS1 (Pedicini et al., 2010). More comprehensive approaches have also been explored by using extended logical formalisms with Boolean variables to assess the effect of different cytokines in making a CD4+ T cell evolve toward a specific state (Naldi et al., 2010). As a general rule, validation studies are performed to endorse and corroborate the usefulness of computational models. Whereas computational models may be used for *in silico* experimentation, *in vivo* and *in vitro* validation needs to be performed in order to ensure its predictability and prove that the plasticity described *in silico* can be translated into an *in vivo* setting in those cases. To address plasticity *in vivo*, the modeling cycle needs to be completed; first, the model needs to be created based on either available data and/or theory-driven knowledge. Afterwards, calibration procedures need to ensure that a good parameter value set has been found and quality control needs to be run to check that the computational model fully represents our experimental data. Third, *in silico* experimentation, using loss-of-function, overexpression or sensitivity analysis strategies need to be performed. Finally, *in vivo* or *in vitro* validation studies will authenticate the computational model and serve as future calibration data for model refinement. These new approaches are helping immunologists to target novel experiments that will shed some light to the subjective issue of CD4+ T cell plasticity.

The computational CD4+ T cell differentiation landscape has generated several validated studies. We validated experimentally that activation of the transcription factor peroxisome proliferator activated receptor gamma (PPARγ) favored the plasticity of Th17 cells toward iTreg cells, a key prediction of our CD4+ T cell model (Carbo et al., 2013b). This model consisted of 60 differential equations, representing 52 reactions and 93 species, computing the differentiation of a CD4+ T cell into Th1, Th2, Th17, and Treg. The model included cytokines, nuclear receptors and transcription factors that defined fate and function of CD4+ T cells. The first set of computationally derived hypotheses were centered around PPARγ and its modulatory role between Th17 and iTreg. Time course simulations illustrated how PPARγ can trigger plasticity in IL-17A+ producing Th17 cells, causing the system to become a iTreg CD4+ T cell. To validate this prediction, *in vitro* and *in vivo* experiments in the context of an IBD onset were designed with PPARγ null CD4+ T cells as well as with a treatment with pioglitazone, a PPARγ activator. The study presented in Miskov-Zivanov et al. (2013) also validated the interaction of FOXP3 and mTOR following TCR activation by purifying and activating DCs and CD4+ T cells and assessing the expression of different intracellular markers using cell staining and flow cytometry. Another example is the validation of the time-dependent, dual T-bet wave during Th1 differentiation validated using gene expression analysis in CD4+ T cells isolated from wild-type and IFNγ null mice (Schulz et al., 2009).

COMPLEMENTARITY OF THEORETICAL AND DATA-DRIVEN MODELS

In computational immunology, often times, the available knowledge about a given set of biological events is used to construct a specific mathematical model. This theoretical approach is therefore directly correlated to the amount of information that is publicly available and the model created upon these pieces of data will only represent the processes delineated within. On the other hand, models can be constructed based solely on analyzing data itself. The increasing availability of high-dimensional data to quantify signaling and cellular responses, together with the novel sequencing technology advancements, is opening a new avenue to use these data-rich datasets to build computational models and help understanding CD4+ T cell differentiation responses. This systems-biology approach, however, can be a double-edged sword: generating high-throughput datasets is part of a big-data strategy, and sometimes, without the appropriate tools, can bring more confusion than understanding to the problem (Bray, 2001). On the other side, this increased availability of data, if used correctly, can streamline the modeling approach, offering a tremendous amount of data for calibration purposes that could allow modelers to build fully calibrated, predictive and extremely comprehensive models that could help generate important hypotheses. These two opposed modeling views can actually be used as a complementary strategy. Theoretical models lack data either for network architecture construction or for model calibration. Data-driven modeling, however, is sometimes confusing, and lack general rules to guide the user and make sense of such big pieces of data. Combining the organization-based approach from theory-driven models with the amount of data and novelty from the data-driven model, highly predictive, hybrid models can be ultimately constructed. In fact, substantial evidence has been shown to understand that the just and only use of data-driven models can represent a trap. The so called "Big Data Hubris" (the often implicit assumption that big data are a substitute, rather than a supplement to, traditional data collection and analysis) already triggered an overestimation of Google's assessment on flu prevalence in 2013 (Lazer et al., 2014). This is a clear example on how data-based and data-driven results were wrongly generated due to the lack of theory underlying unstructured data integration.

The long-standing traditional theory-driven approach has been proven to provide helpful insights on how CD4+ T cells function, where modeling strategies are based on prior biological understanding of the molecular mechanisms involved (Fishman and Perelson, 1999; Hofer et al., 2002; Mendoza, 2006; Klinke, 2007; Hong et al., 2011). However, often times theory-driven modeling is intimately linked to reductionist approaches, since the availability of calibration data can become an issue if building comprehensive networks. Data-driven modeling emerges as a new and complementary approach for multivariate analysis and systems-level analyses. Often times, predictability in computational systems is linked to either the lack of data to construct the computational model or the limitations on the model topology. The combination of data-driven approaches and theoretical strategies may solve these problems therefore promoting the creation of truly predictive models. An example on how to use high-throughput data to construct a CD4+ T cell comprehensive network is the study published by Yosef et al., where they used transcriptional profiling with microarrays at high temporal resolution to build a Th17 induction system (Yosef et al., 2013). In this study, 1291 genes were differentially identified and clustered into 20 groups, depending on their temporal profiles. Another advantage highlighted in this study is the use of modules to explain the processes controlling Th17 differentiation. Four regulatory modules were identified: the positive module that increased IL-17 levels, the negative module that downregulated IL-17, the signature of Th17 genes and signature of other CD4+ T cell subtypes. This work supported the finding of 3 novel key regulators of Th17 function: *Mina*, *Fas*, and *Pou2af1*. Another study where data-driven approaches were taken was the work performed by Ciofani et al., where they combined genome-wide transcription factor occupancy, expression profiling of transcription factor mutants, and transcriptional regulatory network (Ciofani et al., 2012). Integration of several datasets allowed the inference of a Th17 network that highlighted some key regulators to Th17 plasticity, such as *Fosl2*. These two approaches have unveiled novel nodes by using a data-driven approach. However, both networks, which represent static pictures, lack dynamics running on the background. By adding dynamics to the system, a whole new dimension can be added. These data-rich models could be used to determine how the system evolves when a node is knocked-out, or how sensitive are reactions and fluxes to change by a special drug or modulator in a more mechanistic manner. A counterfactual example related to the CD4+ T cell differentiation process is the role of IL-17A in chronic inflammation during IBD. Although it has been reported increased expression of IL-17A during IBD (Fujino et al., 2003) and both IL-17R-deficient mice in TNBS-induced colitis model (Zhang et al., 2006) as well as IL-17A-deficient mice in a DSS-induced colitis model (Ito et al., 2008) were reported to worsen the clinical disease symptoms, some other opposing studies highlighted the protective role of IL-17A production *in vivo* (Ogawa et al., 2004; O'Connor et al., 2009). Very interestingly, a human anti-IL-17A monoclonal antibody to treat Crohn's disease showed that blockade of IL-17A in humans was ineffective and higher rates of adverse events were noted compared with the placebo group (Hueber et al., 2012). In this case, where it is clear there are missing pieces in this puzzle, a combined strategy with both theory-driven and data-driven modeling could shed some light by looking at other players in these intricate and complex interactions.

Data-driven modeling nicely complements and synergizes with theory-driven due to the availability of data for calibration purposes, the potential of discovering novel regulators in the network that have never been described before, and the capability to comprehensively and mechanistically understand complex systems. At the same time, hypotheses extracted from modeling need to be validated to become accepted theories by the community. The combination of theory driven models with data-driven approaches is becoming a strong, useful tool to ensure that the basic knowledge is represented, but at the same time, that novelty and higher predictability is reached. The combination of these two different strategies and multiscalability is now increasing the predictability of very comprehensive models.

DETERMINISTIC vs. STOCHASTIC APPROACHES

In complex regulatory schemas, such as the CD4+ T cell differentiation network, gene expression is controlled by transcriptional signals that determine how rapid and how often a specific gene is transcribed. This transcription process, however, depends on other signals and molecules, such as transcription factors and promoter signals that will trigger cell-to-cell variability. Often times, gene transcription is a result of a combination of other signaling cascades, therefore adding not only complexity and variability due to the differential activation of upstream molecules, but also a time delay while the signal molecule concentration either accumulates or decays.

In CD4+ T cell differentiation, variability is a key component of the process. In fact, not all the cells expressing RORγt exhibit IL-17A production even in the presence of the correct inductors TGFβ and IL-6 (Zhou et al., 2008). Furthermore, Guo et al. showed how IL-4 secreting and non-secreting cells from Th2 cultures have a similar probability of producing IL-4 upon subsequent stimulation, implying that there is stochastic element in IL-4 production by stimulated Th2 cells (Guo et al., 2004). Even after assuming that most genes are expressed from both alleles when the transcription machinery is in place, some studies point out that some cytokine genes in T cells are often expressed in a monoallelic manner (Riviere et al., 1998). Alternatively, the transcription rates also vary if agonistic transcription factors are bound (Chen et al., 2011). Given these set of premises, stochastic approaches that add this type of variability within the CD4+ T cell subset can be used to help explain biological variation. In this case, this variability offers a unique way to control regulation, by inducing stimuli but controlling the fraction of cells expressing a specific cytokine.

Deterministic models of CD4+ T cell differentiation are more prevalent than stochastic-based models. Of note, deterministic approaches have unveiled a large amount of findings that relate to single cell behavior. A fraction of these models have focused on the analysis of one phenotype only (Schulz et al., 2009; Gross et al., 2011), and other models have focused on more than one phenotype and the interactions between the resulting states (van den Ham and de Boer, 2008; Gross et al., 2011; Hong et al., 2011; Carbo et al., 2013b). Mariani et al., in contrast, used a stochastic approach to show how an IL-4 stochastic mechanism acting at the chromatin level can be integrated with transcriptional regulation to quantitatively control cell-to-cell variability (Mariani et al., 2010). Furthermore, Santoni et al. used an agent-based model to assess Th1 vs. Th2 fates in the context of hypersensitivity reactions (Santoni et al., 2008). Recently, Mei et al. assessed the role of the IL-6 receptor in controlling the balance between Th17 and iTreg using a novel, web-based stochastic modeling tool (Mei et al., 2013a). Other approaches have used the mathematical formulation of a cell population master equation (CPME) that describes population dynamics and takes into account the major sources of heterogeneity, namely stochasticity in reaction, DNA-duplication, and division, using the Montecarlo algorithm (Stamatakis and Zygourakis, 2010). Manninen et al. (2006) developed several approaches to incorporate stochasticity into deterministic differential equation models, obtaining so-called Itô stochastic differential equations, and applied them to neuronal protein kinase C signal transduction pathway modeling. Even though traditional molecular biology research has tended to composite single cell deterministic models, diversification of T cell fate during CD4+ T cell differentiation implies that the fate of any individual cell may also be acquired stochastically. Therefore, stochastic simulations within the CD4+ T cell differentiation process could help to understand the tight regulation between phenotypes as well as help identify key nodes that, when acting at higher variability, can skew the output of differentiation into a specific differentiation program.

APPLICATION OF MULTISCALE MODELING TO STUDY CD4+ T CELL DIFFERENTIATION

CD4+ T cell differentiation is a process where a change in the intracellular compartment can tremendously impact the outcome of tissue pathology and clinical disease. Distinct intracellular processes dictate the secretion of chemokines, cytokines, and other soluble factors. These components can, at the same time, modulate other CD4+ T cell nearby by binding to specific receptors. This population effect can modulate other downstream immune subsets that can ultimately affect the formation of lesions at the tissue level. Thus, CD4+ T cell differentiation is not only an intracellular process: population and cellular organization are another major mechanism that may contribute to the change in the dominant phenotype of effector CD4+ T cells during chronic pathologies (Magombedze et al., 2013). Indeed, the mucosal immune system includes hierarchical interactions between cells leading to emerging behaviors with dimensions ranging from nanometers to meters and time scales from nanoseconds to years. The spatiotemporal scales where CD4+ T cells participate can actually range from micro-seconds to months or years and to nanometers to centimeters or meters (**Figure 3A**). Complex and dynamic information processing networks transfer information across scales in immunity encoding host responses and repair measures. The architecture of such multiscale network also needs to be completely embedded in a comprehensive, integrated system. Because of this flexibility in parameter calibration and sensitivity analyses, Ordinary or Stochastic Differential Equations (ODE or SDE) are ideal candidates to encapsulate and simulate intracellular events. In addition, neural networks have also been used to classify and simulate immune cell subsets (Mei et al., 2013b). In the multiscale setting, these ODE- or SDE-based models would reproduce intracellular CD4+ T cell activation with a release of cytokines and chemokines as a result of the process of differentiation. Partial Differential Equation (PDE) modeling would be a great way to simulate the diffusion reactions of such cytokines in the environment. Ultimately, an agent-based model, adding randomness to the biological system, which helps to better represent responses at the cellular level, would encompass and organize the ODE/SDE models with the PDE simulations by simulating CD4+ T cells as objects that can change its state depending on the cytokine milieu. As a result of these premises, multiscale models are positioned as a comprehensive tool to understand not only the intracellular events happening within the CD4+ T cell compartment at a single cell level, but also understanding the interactions and sensitivities, at the cellular, population and tissue levels, that contribute to disease chronicity, tolerance, or resolution (**Figure 3B**).

All together, ODE models can calculate the intracellular concentration of different species over time, PDE models could analyze the gradient concentration of cytokines and chemokines secreted by the ODE model, ABM-based models could modulate the cell-cell interactions and spatial compartments could represent the tissue-level scale, including lesion formation. Current experimental techniques are limited in allowing immunologists to quantitatively manipulate immune responses to pathogens in

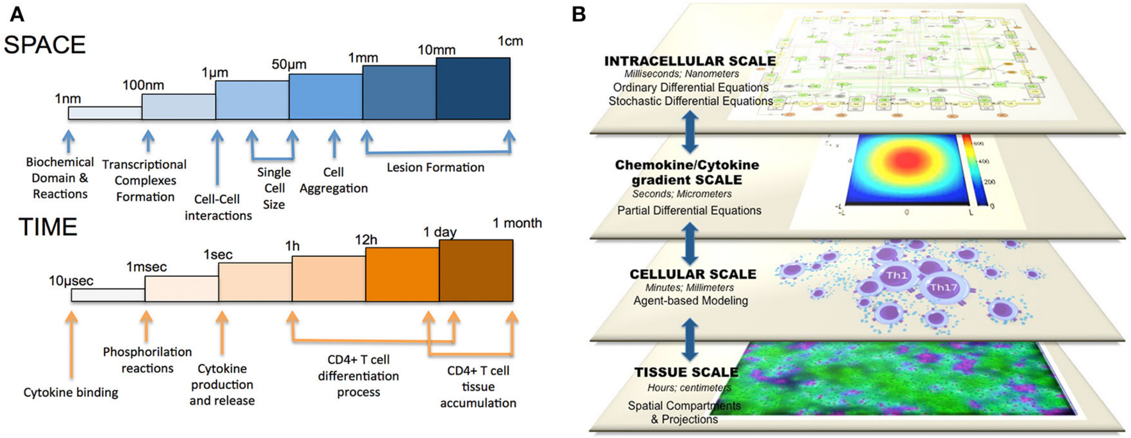

A
SPACE
1nm
100nm
1µm
50µm
1mm
10mm
1cm
Biochemical Domain & Reactions
Transcriptional Complexes Formation
Cell-Cell interactions
Single Cell Size
Cell Aggregation
Lesion Formation
TIME
10µsec
1msec
1sec
1h
12h
1 day
1 month
Cytokine binding
Phosphorilation reactions
Cytokine production and release
CD4+ T cell differentiation process
CD4+ T cell tissue accumulation
B
INTRACELLULAR SCALE
Milliseconds; Nanometers
Ordinary Differential Equations
Stochastic Differential Equations
Chemokine/Cytokine gradient SCALE
Seconds; Micrometers
Partial Differential Equations
CELLULAR SCALE
Minutes; Millimeters
Agent-based Modeling
Th1
Th17
TISSUE SCALE
Hours; centimeters
Spatial Compartments & Projections

FIGURE 3 | Multiscale modeling of CD4+ T cell differentiation. The CD4+ T cell differentiation process comprises **(A)** different spatiotemporal parameters (milliseconds to hours and nanometers to centimeters) as well as **(B)** different scales (intracellular, difussion gradient, cellular, and tissue-level scale).

a controlled manner in animal models and to trace events at the tissue level confidently back to specific cellular level interactions and molecular or signaling mechanisms. In a multiscale model, one can test whether mechanisms seen in the experimental context *in vivo* or *in vitro* are plausible explanations for phenomena observed at the clinical level. There have been several previous studies on multiscale modeling in the context of immunity: Sloot and Hoekstra (2010) proposed a multi-scale modeling methodology in computational biomedicine and presented two cases studies. Krinner et al. (2013) coupled an agent-based model of hematopoietic stem cells with an ODE model of granulopoiesis. Also, Klinke (2007) published a multiscale model of dendritic cell education and trafficking in the lung. Some very recent multiscale approaches to study the CD4+ T cell population have been performed in the context of HIV infection (Yeghiazarian et al., 2013) and also in the context of CD4+ T cell migration, signaling, and interaction with the APC compartment (Huang, 2010). Furthermore, Dwivedi et al. recently developed a multiscale systems model of IL-6–mediated immune regulation in Crohn's disease, by integrating intracellular signaling with organ-level dynamics of pharmacological markers underlying the disease (Dwivedi et al., 2014). Santoni et al. (2008) also combined an agent-based model of type I hypersensitivity reactions showing hallmarks of the response to a generic allergen with a gene regulatory network for the switch of Th1/Th2 phenoptypes. Despite all these strategies and studies, there is no comprehensive multiscale model that computes more than two phenotypes of CD4+ T cell differentiation based on the availability of certain factors in the environment and considers more than one scale in the simulation.

Multiscale modeling may also help integrate immune processes and metabolic pathways to build systems-level immunometabolic frameworks. Indeed, T cell metabolism is highly dynamic and has a tremendous impact on the ability of T cells to grow, activate and differentiate (Gerriets and Rathmell, 2012). Glucose metabolism is one of the pathways that has been targeted to explore immunometabolism. One example is the study from Maciver et al. where they found that activation of T cells causes a large increase in glucose transporter 1 (Glut1) expression and surface localization (Maciver et al., 2008). Furthermore, CD28 appeared to promote Akt-independent up-regulation of Glut1 and Akt-dependent Glut1 cell surface trafficking (Jacobs et al., 2008). Multiscale modeling analyses could also help to differentiate which are the metabolic needs to promote specific developmental programs. In fact, effector and regulatory phenotypes have distinct glycolytic and lipid oxidative metabolic programs (Michalek et al., 2011). Pearce et al. reviewed (Pearce, 2010) how activated T cells have an anabolic metabolism, whereas non-proliferating T cells had an opposed catabolic metabolism. Furthermore, autophagy has been found to be essential for T cell survival and proliferation (Pua et al., 2007). Later the same group described how the same process of autophagy may have a physiologically significant role in the clearance of mitochondria in T cells as part of normal T cell homeostasis (Pua et al., 2009), creating a clear link between immunometabolism and T cell function. By using a multiscale strategy, these metabolic programs could be integrated in differentiation simulations and more importantly, the processes could be manipulated to control anti- and pro-inflammatory development in the context of inflammatory diseases. Thus, modeling can be used to quantitatively study dynamic processes located at the interface of immunity and metabolism.

Of note, understanding the mechanisms of CD4+ T cell differentiation and plasticity across scales can lead to the identification of novel therapeutic targets for skewing effector cells into regulatory phenotypes that suppress inflammation. Therefore, multiscale modeling can, indeed, increase predictability and systems-wide mechanistic understanding as to how CD4+ T cells are activated, maintained, and transformed.

CONCLUSION

T cell immune responses are extremely heterogeneous and complex. This variability is not fully understood and there are still several questions in regards to CD4+ T cell plasticity and function. Indeed, the issue of what criteria to use to characterize distinct T cell subsets is becoming increasingly complicated. Moreover, the idea that CD4+ T cells are governed by function and not by phenotype is clearly emerging as more double positive and plastic behaviors are being unveiled. The possibility that every helper T cell process is a unique combination of molecules, however, cannot be discarded. This review highlighted how CD4+ T cells have a strong predisposition to certain developmental programs, but it also showed how, at certain times with certain environmental signals, this predisposition is skewed toward another program. Computationally, the plural CD4+ T cell scenario is still a field of interest and active investigation. As new advancements in the understanding of immune responses continue to unfold, computational modeling approaches are likely to be required to comprehensively and systematically investigate mechanisms across spatiotemporal scales and to help integrate diverse data types.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the National Institutes of Health (5R01AT004308) to Josep Bassaganya-Riera, NIAID Contract No. HHSN272201000056C to Josep Bassaganya-Riera, and funds from the Nutritional Immunology and Molecular Medicine Laboratory (URL: www.nimml.org).

REFERENCES

Abraham, C., and Cho, J. H. (2009). Inflammatory bowel disease. *N. Engl. J. Med.* 361, 2066–2078. doi: 10.1056/NEJMra0804647

Ahern, P. P., Schiering, C., Buonocore, S., McGeachy, M. J., Cua, D. J., Maloy, K. J., et al. (2010). Interleukin-23 drives intestinal inflammation through direct activity on T cells. *Immunity* 33, 279–288. doi: 10.1016/j.immuni.2010.08.010

Ansel, K. M., McHeyzer-Williams, L. J., Ngo, V. N., McHeyzer-Williams, M. G., and Cyster, J. G. (1999). In vivo-activated CD4 T cells upregulate CXC chemokine receptor 5 and reprogram their response to lymphoid chemokines. *J. Exp. Med.* 190, 1123–1134. doi: 10.1084/jem.190.8.1123

Arababadi, M. K., Nosratabadi, R., Hassanshahi, G., Yaghini, N., Pooladvand, V., Shamsizadeh, A., et al. (2010). Nephropathic complication of type-2 diabetes is following pattern of autoimmune diseases? *Diabetes Res. Clin. Pract.* 87, 33–37. doi: 10.1016/j.diabres.2009.09.027

Basu, R., Hatton, R. D., and Weaver, C. T. (2013). The Th17 family: flexibility follows function. *Immunol. Rev.* 252, 89–103. doi: 10.1111/imr.12035

Bergmann, C., Van Hemmen, J. L., and Segel, L. A. (2001). Th1 or Th2: how an appropriate T helper response can be made. *Bull. Math. Biol.* 63, 405–430. doi: 10.1006/bulm.2000.0215

Bergmann, C., van Hemmen, J. L., and Segel, L. A. (2002). How instruction and feedback can select the appropriate T helper response. *Bull. Math. Biol.* 64, 425–446. doi: 10.1006/bulm.2001.0258

Bettelli, E., Carrier, Y., Gao, W., Korn, T., Strom, T. B., Oukka, M., et al. (2006). Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells. *Nature* 441, 235–238. doi: 10.1038/nature04753

Bray, D. (2001). Reasoning for results. *Nature* 412, 863. doi: 10.1038/35091132

Breitfeld, D., Ohl, L., Kremmer, E., Ellwart, J., Sallusto, F., Lipp, M., et al. (2000). Follicular B helper T cells express CXC chemokine receptor 5, localize to B cell follicles, and support immunoglobulin production. *J. Exp. Med.* 192, 1545–1552. doi: 10.1084/jem.192.11.1545

Carbo, A., Bassaganya-Riera, J., Pedragosa, M., Viladomiu, M., Marathe, M., Eubank, S., et al. (2013a). Predictive computational modeling of the mucosal immune responses during Helicobacter pylori infection. *PLoS ONE* 8:e73365. doi: 10.1371/journal.pone.0073365

Carbo, A., Hontecillas, R., Kronsteiner, B., Viladomiu, M., Pedragosa, M., Lu, P., et al. (2013b). Systems modeling of molecular mechanisms controlling cytokine-driven CD4+ T cell differentiation and phenotype plasticity. *PLoS Comput. Biol.* 9:e1003027. doi: 10.1371/journal.pcbi.1003027

Chen, Z., Lin, F., Gao, Y., Li, Z., Zhang, J., Xing, Y., et al. (2011). FOXP3 and RORgammat: transcriptional regulation of Treg and Th17. *Int. Immunopharmacol.* 11, 536–542. doi: 10.1016/j.intimp.2010.11.008

Chitnis, T. (2007). The role of CD4 T cells in the pathogenesis of multiple sclerosis. *Int. Rev. Neurobiol.* 79, 43–72. doi: 10.1016/S0074-7742(07)79003-7

Ciofani, M., Madar, A., Galan, C., Sellars, M., Mace, K., Pauli, F., et al. (2012). A validated regulatory network for Th17 cell specification. *Cell* 151, 289–303. doi: 10.1016/j.cell.2012.09.016

Codarri, L., Gyulveszi, G., Tosevski, V., Hesske, L., Fontana, A., Magnenat, L., et al. (2011). RORgammat drives production of the cytokine GM-CSF in helper T cells, which is essential for the effector phase of autoimmune neuroinflammation. *Nat. Immunol.* 12, 560–567. doi: 10.1038/ni.2027

Drayton, D. L., Liao, S., Mounzer, R. H., and Ruddle, N. H. (2006). Lymphoid organ development: from ontogeny to neogenesis. *Nat. Immunol.* 7, 344–353. doi: 10.1038/ni1330

Dwivedi, G., Fitz, L., Hegen, M., Martin, S. W., Harrold, J., Heatherington, A., et al. (2014). A multiscale model of interleukin-6-mediated immune regulation in Crohn's disease and its application in drug discovery and development. *CPT Pharmacometr. Syst. Pharmacol.* 3:e89. doi: 10.1038/psp.2013.64

Eftimie, R., Bramson, J. L., and Earn, D. J. (2010). Modeling anti-tumor Th1 and Th2 immunity in the rejection of melanoma. *J. Theor. Biol.* 265, 467–480. doi: 10.1016/j.jtbi.2010.04.030

El-Behi, M., Ciric, B., Dai, H., Yan, Y., Cullimore, M., Safavi, F., et al. (2011). The encephalitogenicity of T(H)17 cells is dependent on IL-1- and IL-23-induced production of the cytokine GM-CSF. *Nat. Immunol.* 12, 568–575. doi: 10.1038/ni.2031

Esplugues, E., Huber, S., Gagliani, N., Hauser, A. E., Town, T., Wan, Y. Y., et al. (2011). Control of TH17 cells occurs in the small intestine. *Nature* 475, 514–518. doi: 10.1038/nature10228

Eyerich, S., Eyerich, K., Pennino, D., Carbone, T., Nasorri, F., Pallotta, S., et al. (2009). Th22 cells represent a distinct human T cell subset involved in epidermal immunity and remodeling. *J. Clin. Invest.* 119, 3573–3585. doi: 10.1172/JCI40202

Fishman, M. A., and Perelson, A. S. (1999). Th1/Th2 differentiation and cross-regulation. *Bull. Math. Biol.* 61, 403–436. doi: 10.1006/bulm.1998.0074

Fontenot, J. D., Gavin, M. A., and Rudensky, A. Y. (2003). Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. *Nat. Immunol.* 4, 330–336. doi: 10.1038/ni904

Fujino, S., Andoh, A., Bamba, S., Ogawa, A., Hata, K., Araki, Y., et al. (2003). Increased expression of interleukin 17 in inflammatory bowel disease. *Gut* 52, 65–70. doi: 10.1136/gut.52.1.65

Fujita, H., Nograles, K. E., Kikuchi, T., Gonzalez, J., Carucci, J. A., and Krueger, J. G. (2009). Human Langerhans cells induce distinct IL-22-producing CD4+ T cells lacking IL-17 production. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21795–21800. doi: 10.1073/pnas.0911472106

Gerriets, V. A., and Rathmell, J. C. (2012). Metabolic pathways in T cell fate and function. *Trends Immunol.* 33, 168–173. doi: 10.1016/j.it.2012.01.010

Gross, F., Metzner, G., and Behn, U. (2011). Mathematical modeling of allergy and specific immunotherapy: Th1-Th2-Treg interactions. *J. Theor. Biol.* 269, 70–78. doi: 10.1016/j.jtbi.2010.10.013

Guo, L., Hu-Li, J., and Paul, W. E. (2004). Probabilistic regulation of IL-4 production in Th2 cells: accessibility at the Il4 locus. *Immunity* 20, 193–203. doi: 10.1016/S1074-7613(04)00025-1

Hardtke, S., Ohl, L., and Forster, R. (2005). Balanced expression of CXCR5 and CCR7 on follicular T helper cells determines their transient positioning to lymph node follicles and is essential for efficient B-cell help. *Blood* 106, 1924–1931. doi: 10.1182/blood-2004-11-4494

Hofer, T., Nathansen, H., Lohning, M., Radbruch, A., and Heinrich, R. (2002). GATA-3 transcriptional imprinting in Th2 lymphocytes: a mathematical model. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9364–9368. doi: 10.1073/pnas.142284699

Hong, T., Xing, J., Li, L., and Tyson, J. J. (2011). A mathematical model for the reciprocal differentiation of T helper 17 cells and induced regulatory T cells. *PLoS Comput. Biol.* 7:e1002122. doi: 10.1371/journal.pcbi.1002122

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., et al. (2006). COPASI–a COmplex PAthway SImulator. *Bioinformatics* 22, 3067–3074. doi: 10.1093/bioinformatics/btl485

Hori, S., Nomura, T., and Sakaguchi, S. (2003). Control of regulatory T cell development by the transcription factor Foxp3. *Science* 299, 1057–1061. doi: 10.1126/science.1079490

Huang, Z. (2010). *Multi-Scale Models of T Cell Activation*. Boston, MA: Massachusetts Institute of Technology.

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 representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531. doi: 10.1093/bioinformatics/btg015

Hueber, W., Sands, B. E., Lewitzky, S., Vandemeulebroecke, M., Reinisch, W., Higgins, P. D., et al. (2012). Secukinumab in Crohn's Disease Study, Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: unexpected results of a randomised, double-blind placebo-controlled trial. *Gut* 61, 1693–1700. doi: 10.1136/gutjnl-2011-301668

Imanishi, M., Okada, N., Konishi, Y., Morikawa, T., Maeda, I., Kitabayashi, C., et al. (2013). Angiotensin II receptor blockade reduces salt sensitivity of blood pressure through restoration of renal nitric oxide synthesis in patients with diabetic nephropathy. *J. Renin Angiotensin Aldosterone Syst.* 14, 67–73. doi: 10.1177/1470320312454764

Islam, S. A., and Luster, A. D. (2012). T cell homing to epithelial barriers in allergic disease. *Nat. Med.* 18, 705–715. doi: 10.1038/nm.2760

Ito, R., Kita, M., Shin-Ya, M., Kishida, T., Urano, A., Takada, R., et al. (2008). Involvement of IL-17A in the pathogenesis of DSS-induced colitis in mice. *Biochem. Biophys. Res. Commun.* 377, 12–16. doi: 10.1016/j.bbrc.2008.09.019

Ivanov, I. I., McKenzie, B. S., Zhou, L., Tadokoro, C. E., Lepelley, A., Lafaille, J. J., et al. (2006). The orphan nuclear receptor RORgammat directs the differentiation program of proinflammatory IL-17+ T helper cells. *Cell* 126, 1121–1133. doi: 10.1016/j.cell.2006.07.035

Jacobs, S. R., Herman, C. E., Maciver, N. J., Wofford, J. A., Wieman, H. L., Hammen, J. J., et al. (2008). Glucose uptake is limiting in T cell activation and requires CD28-mediated Akt-dependent and independent pathways. *J. Immunol.* 180, 4476–4486. doi: 10.4049/jimmunol.180.7.4476

Jagannathan-Bogdan, M., McDonnell, M. E., Shin, H., Rehman, Q., Hasturk, H., Apovian, C. M., et al. (2011). Elevated proinflammatory cytokine production by a skewed T cell compartment requires monocytes and promotes inflammation in type 2 diabetes. *J. Immunol.* 186, 1162–1172. doi: 10.4049/jimmunol.1002615

Klein, L., Hinterberger, M., Wirnsberger, G., and Kyewski, B. (2009). Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat. Rev. Immunol.* 9, 833–844. doi: 10.1038/nri2669

Klinke, D. J. 2nd. (2007). A multi-scale model of dendritic cell education and trafficking in the lung: implications for T cell polarization. *Ann. Biomed. Eng.* 35, 937–955. doi: 10.1007/s10439-007-9318-6

Krinner, A., Roeder, I., Loeffler, M., and Scholz, M. (2013). Merging concepts-coupling an agent-based model of hematopoietic stem cells with an ODE model of granulopoiesis. *BMC Syst. Biol.* 7:117. doi: 10.1186/1752-0509-7-117

Kurschus, F. C., Croxford, A. L., Heinen, A. P., Wortge, S., Ielo, D., and Waisman, A. (2010). Genetic proof for the transient nature of the Th17 phenotype. *Eur. J. Immunol.* 40, 3336–3346. doi: 10.1002/eji.201040755

Langrish, C. L., Chen, Y., Blumenschein, W. M., Mattson, J., Basham, B., Sedgwick, J. D., et al. (2005). IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. *J. Exp. Med.* 201, 233–240. doi: 10.1084/jem.20041257

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science* 343, 1203–1205. doi: 10.1126/science.1248506

Lee, Y. K., Turner, H., Maynard, C. L., Oliver, J. R., Chen, D., Elson, C. O., et al. (2009). Late developmental plasticity in the T helper 17 lineage. *Immunity* 30, 92–107. doi: 10.1016/j.immuni.2008.11.005

Liao, W., Lin, J. X., Wang, L., Li, P., and Leonard, W. J. (2011). Modulation of cytokine receptors by IL-2 broadly regulates differentiation into helper T cell lineages. *Nat. Immunol.* 12, 551–559. doi: 10.1038/ni.2030

Lin, S., Yang, X., Liang, D., and Zheng, S. G. (2014). Treg cells: a potential regulator for IL-22 expression? *Int. J. Clin. Exp. Pathol.* 7, 474–480.

Lochner, M., Peduto, L., Cherrier, M., Sawa, S., Langa, F., Varona, R., et al. (2008). In vivo equilibrium of proinflammatory IL-17+ and regulatory IL-10+ Foxp3+ RORgamma t+ T cells. *J. Exp. Med.* 205, 1381–1393. doi: 10.1084/jem.20080034

Luckheeram, R. V., Zhou, R., Verma, A. D., and Xia, B. (2012). CD4(+)T cells: differentiation and functions. *Clin. Dev. Immunol.* 2012:925135. doi: 10.1155/2012/925135

Ma, C. S., Tangye, S. G., and Deenick, E. K. (2010). Human Th9 cells: inflammatory cytokines modulate IL-9 production through the induction of IL-21. *Immunol. Cell Biol.* 88, 621–623. doi: 10.1038/icb.2010.73

Maciver, N. J., Jacobs, S. R., Wieman, H. L., Wofford, J. A., Coloff, J. L., and Rathmell, J. C. (2008). Glucose metabolism in lymphocytes is a regulated process with significant effects on immune cell function and survival. *J. Leukoc. Biol.* 84, 949–957. doi: 10.1189/jlb.0108024

Magombedze, G., Eda, S., and Ganusov, V. V. (2014). Competition for antigen between Th1 and Th2 responses determines the timing of the immune response switch during Mycobaterium avium subspecies paratuberulosis infection in ruminants. *PLoS Comput. Biol.* 10:e1003414. doi: 10.1371/journal.pcbi.1003414

Magombedze, G., Reddy, P. B., Eda, S., and Ganusov, V. V. (2013). Cellular and population plasticity of helper CD4(+) T cell responses. *Front. Physiol.* 4:206. doi: 10.3389/fphys.2013.00206

Manninen, T., Linne, M. L., and Ruohonen, K. (2006). Developing Ito stochastic differential equation models for neuronal signal transduction pathways. *Comput. Biol. Chem.* 30, 280–291. doi: 10.1016/j.compbiolchem.2006.04.002

Mariani, L., Lohning, M., Radbruch, A., and Hofer, T. (2004). Transcriptional control networks of cell differentiation: insights from helper T lymphocytes. *Prog. Biophys. Mol. Biol.* 86, 45–76. doi: 10.1016/j.pbiomolbio.2004.02.007

Mariani, L., Schulz, E. G., Lexberg, M. H., Helmstetter, C., Radbruch, A., Lohning, M., et al. (2010). Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression. *Mol. Syst. Biol.* 6, 359. doi: 10.1038/msb.2010.13

Mathur, A. N., Chang, H. C., Zisoulis, D. G., Kapur, R., Belladonna, M. L., Kansas, G. S., et al. (2006). T-bet is a critical determinant in the instability of the IL-17-secreting T-helper phenotype. *Blood* 108, 1595–1601. doi: 10.1182/blood-2006-04-015016

Mei, Y., Carbo, A., Hontecillas, R., and Bassaganya-Riera, J. (2013a). "ENISI SDE: a novel web-based stochastic modeling tool for computational biology," in *2013 IEEE International Conference on Bioinformatics and Biomedicine* (Shanghai), 392–397. doi: 10.1109/BIBM.2013.6732524

Mei, Y., Hontecillas, R., Zhang, X., Bisset, K. R., Eubank, S., Hoops, S., et al. (2012). "ENISI visual, an agent-based simulator for modeling gut immunity," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Philadelphia, PA), 1–5. doi: 10.1109/BIBM.2012.6392624

Mei, Y., Hontecillas, R., Zhang, X., Carbo, A., and Bassaganya-Riera, J. (2013b). "Neural network models for classifying immune cell subsets," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Shanghai), 5–11.

Mendoza, L. (2006). A network model for the control of the differentiation process in Th cells. *BioSystems* 84, 101–114. doi: 10.1016/j.biosystems.2005.10.004

Mendoza, L. (2013). A virtual culture of CD4+ T lymphocytes. *Bull. Math. Biol.* 75, 1012–1029. doi: 10.1007/s11538-013-9814-9

Mendoza, L., and Pardo, F. (2010). A robust model to describe the differentiation of T-helper cells. *Theory Biosci.* 129, 283–293. doi: 10.1007/s12064-010-0112-x

Michalek, R. D., Gerriets, V. A., Jacobs, S. R., Macintyre, A. N., MacIver, N. J., Mason, E. F., et al. (2011). Cutting edge: distinct glycolytic and lipid oxidative metabolic programs are essential for effector and regulatory CD4+ T cell subsets. *J. Immunol.* 186, 3299–3303. doi: 10.4049/jimmunol.1003613

Miskov-Zivanov, N., Turner, M. S., Kane, L. P., Morel, P. A., and Faeder, J. R. (2013). The duration of T cell stimulation is a critical determinant of cell fate and plasticity. *Sci. Signal.* 6:ra97. doi: 10.1126/scisignal.2004217

Monsonego, A., Nemirovsky, A., and Harpaz, I. (2013). CD4 T cells in immunity and immunotherapy of Alzheimer's disease. *Immunology* 139, 438–446. doi: 10.1111/imm.12103

Mosmann, T. R., and Coffman, R. L. (1989). TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* 7, 145–173. doi: 10.1146/annurev.iy.07.040189.001045

Nakayamada, S., Kanno, Y., Takahashi, H., Jankovic, D., Lu, K. T., Johnson, T. A., et al. (2011). Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* 35, 919–931. doi: 10.1016/j.immuni.2011.11.012

Naldi, A., Carneiro, J., Chaouiya, C., and Thieffry, D. (2010). Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Comput. Biol.* 6:e1000912. doi: 10.1371/journal.pcbi.1000912

O'Connor, W. Jr., Kamanaka, M., Booth, C. J., Town, T., Nakae, S., Iwakura, Y., et al. (2009). A protective function for interleukin 17A in T cell-mediated intestinal inflammation. *Nat. Immunol.* 10, 603–609. doi: 10.1038/ni.1736

O'Connor, W. Jr., Zenewicz, L. A., and Flavell, R. A. (2010). The dual nature of T(H)17 cells: shifting the focus to function. *Nat. Immunol.* 11, 471–476. doi: 10.1038/ni.1882

Oestreich, K. J., Mohn, S. E., and Weinmann, A. S. (2012). Molecular mechanisms that control the expression and activity of Bcl-6 in TH1 cells to regulate flexibility with a TFH-like gene profile. *Nat. Immunol.* 13, 405–411. doi: 10.1038/ni.2242

Ogawa, A., Andoh, A., Araki, Y., Bamba, T., and Fujiyama, Y. (2004). Neutralization of interleukin-17 aggravates dextran sulfate sodium-induced colitis in mice. *Clin. Immunol.* 110, 55–62. doi: 10.1016/j.clim.2003.09.013

Ohshima, K., Mogi, M., Jing, F., Iwanami, J., Tsukuda, K., Min, L. J., et al. (2012). Roles of interleukin 17 in angiotensin II type 1 receptor-mediated insulin resistance. *Hypertension* 59, 493–499. doi: 10.1161/HYPERTENSIONAHA.111.183178

Osorio, F., LeibundGut-Landmann, S., Lochner, M., Lahl, K., Sparwasser, T., Eberl, G., et al. (2008). DC activated via dectin-1 convert Treg into IL-17 producers. *Eur. J. Immunol.* 38, 3274–3281. doi: 10.1002/eji.200838950

Pearce, E. L. (2010). Metabolism in T cell activation and differentiation. *Curr. Opin. Immunol.* 22, 314–320. doi: 10.1016/j.coi.2010.01.018

Pedicini, M., Barrenas, F., Clancy, T., Castiglione, F., Hovig, E., Kanduri, K., et al. (2010). Combining network modeling and gene expression microarray analysis to explore the dynamics of Th1 and Th2 cell regulation. *PLoS Comput. Biol.* 6:e1001032. doi: 10.1371/journal.pcbi.1001032

Pot, C., Apetoh, L., and Kuchroo, V. K. (2011). Type 1 regulatory T cells (Tr1) in autoimmunity. *Semin. Immunol.* 23, 202–208. doi: 10.1016/j.smim.2011.07.005

Pua, H. H., Dzhagalov, I., Chuck, M., Mizushima, N., and He, Y. W. (2007). A critical role for the autophagy gene Atg5 in T cell survival and proliferation. *J. Exp. Med.* 204, 25–31. doi: 10.1084/jem.20061303

Pua, H. H., Guo, J., Komatsu, M., and He, Y. W. (2009). Autophagy is essential for mitochondrial clearance in mature T lymphocytes. *J. Immunol.* 182, 4046–4055. doi: 10.4049/jimmunol.0801143

Ramirez, J. M., Brembilla, N. C., Sorg, O., Chicheportiche, R., Matthes, T., Dayer, J. M., et al. (2010). Activation of the aryl hydrocarbon receptor reveals distinct requirements for IL-22 and IL-17 production by human T helper cells. *Eur. J. Immunol.* 40, 2450–2459. doi: 10.1002/eji.201040461

Riviere, I., Sunshine, M. J., and Littman, D. R. (1998). Regulation of IL-4 expression by activation of individual alleles. *Immunity* 9, 217–228. doi: 10.1016/S1074-7613(00)80604-4

Santoni, D., Pedicini, M., and Castiglione, F. (2008). Implementation of a regulatory gene network to simulate the TH1/2 differentiation in an agent-based model of hypersensitivity reactions. *Bioinformatics* 24, 1374–1380. doi: 10.1093/bioinformatics/btn135

Schulz, E. G., Mariani, L., Radbruch, A., and Hofer, T. (2009). Sequential polarization and imprinting of type 1 T helper lymphocytes by interferon-gamma and interleukin-12. *Immunity* 30, 673–683. doi: 10.1016/j.immuni.2009.03.013

Sloot, P. M., and Hoekstra, A. G. (2010). Multi-scale modelling in computational biomedicine. *Brief. Bioinformatics* 11, 142–152. doi: 10.1093/bib/bbp038

Sonnenberg, G. F., Fouser, L. A., and Artis, D. (2011). Border patrol: regulation of immunity, inflammation and tissue homeostasis at barrier surfaces by IL-22. *Nat. Immunol.* 12, 383–390. doi: 10.1038/ni.2025

Stamatakis, M., and Zygourakis, K. (2010). A mathematical and computational approach for integrating the major sources of cell population heterogeneity. *J. Theor. Biol.* 266, 41–61. doi: 10.1016/j.jtbi.2010.06.002

Trifari, S., and Spits, H. (2010). IL-22-producing CD4+ T cells: middle-men between the immune system and its environment. *Eur. J. Immunol.* 40, 2369–2371. doi: 10.1002/eji.201040848

van den Ham, H. J., and de Boer, R. J. (2008). From the two-dimensional Th1 and Th2 phenotypes to high-dimensional models for gene regulation. *Int. Immunol.* 20, 1269–1277. doi: 10.1093/intimm/dxn093

Wendelsdorf, K. V., Alam, M., Bassaganya-Riera, J., Bisset, K., Eubank, S., Hontecillas, R., et al. (2012). ENteric Immunity SImulator: a tool for in silico study of gastroenteric infections. *IEEE Trans. Nanobioscience* 11, 273–288. doi: 10.1109/TNB.2012.2211891

Yang, X. O., Chang, S. H., Park, H., Nurieva, R., Shah, B., Acero, L., et al. (2008). Regulation of inflammatory responses by IL-17F. *J. Exp. Med.* 205, 1063–1075. doi: 10.1084/jem.20071978

Yates, A., Bergmann, C., Van Hemmen, J. L., Stark, J., and Callard, R. (2000). Cytokine-modulated regulation of helper T cell populations. *J. Theor. Biol.* 206, 539–560. doi: 10.1006/jtbi.2000.2147

Yates, A., Callard, R., and Stark, J. (2004). Combining cytokine signalling with T-bet and GATA-3 regulation in Th1 and Th2 differentiation: a model for cellular decision-making. *J. Theor. Biol.* 231, 181–196. doi: 10.1016/j.jtbi.2004.06.013

Yeghiazarian, L., Cumberland, W. G., and Yang, O. O. (2013). A stochastic multiscale model of HIV-1 transmission for decision-making: application to a MSM population. *PLoS ONE* 8:e70578. doi: 10.1371/journal.pone.0070578

Yosef, N., Shalek, A. K., Gaublomme, J. T., Jin, H., Lee, Y., Awasthi, A., et al. (2013). Dynamic regulatory network controlling TH17 cell differentiation. *Nature* 496, 461–468. doi: 10.1038/nature11981

Zeng, C., Shi, X., Zhang, B., Liu, H., Zhang, L., Ding, W., et al. (2012). The imbalance of Th17/Th1/Tregs in patients with type 2 diabetes: relationship with metabolic factors and complications. *J. Mol. Med.* 90, 175–186. doi: 10.1007/s00109-011-0816-5

Zhang, Z., Zheng, M., Bindas, J., Schwarzenberger, P., and Kolls, J. K. (2006). Critical role of IL-17 receptor signaling in acute TNBS-induced colitis. *Inflamm. Bowel Dis.* 12, 382–388. doi: 10.1097/01.MIB.0000218764.06959.91

Zhou, L., Lopes, J. E., Chong, M. M., Ivanov, I. I., Min, R., Victora, G. D., et al. (2008). TGF-beta-induced Foxp3 inhibits T(H)17 cell differentiation by antagonizing RORgammat function. *Nature* 453, 236–240. doi: 10.1038/nature06878

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

Received: 17 May 2014; accepted: 10 July 2014; published online: 29 July 2014.
Citation: Carbo A, Hontecillas R, Andrew T, Eden K, Mei Y, Hoops S and Bassaganya-Riera J (2014) Computational modeling of heterogeneity and function of CD4+ T cells. Front. Cell Dev. Biol. ***2****:31. doi: 10.3389/fcell.2014.00031*
This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

PERSPECTIVE ARTICLE
published: 17 September 2014
doi: 10.3389/fcell.2014.00051

Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems

Claudia Angelini[1,2]* and Valerio Costa[2,3]

[1] Istituto per le Applicazioni del Calcolo "M. Picone" - CNR, Napoli, Italy
[2] Computational and Biology Open Laboratory (ComBOlab), Napoli, Italy
[3] Institute of Genetics and Biophysics "A. Buzzati-Traverso" - CNR, Napoli, Italy

Edited by:
Christine Nardini, Partner Institute for Computational Biology, China

Reviewed by:
Hao Wu, Emory University, USA
Reina Luco, CNRS, France

****Correspondence:***
Claudia Angelini, Istituto per le Applicazioni del Calcolo "M. Picone" - Consiglio Nazionale delle Ricerche, Via Pietro Castellinol, 111 80131 Napoli, Italy
e-mail: claudia.angelini@cnr.it

The availability of omic data produced from international consortia, as well as from worldwide laboratories, is offering the possibility both to answer long-standing questions in biomedicine/molecular biology and to formulate novel hypotheses to test. However, the impact of such data is not fully exploited due to a limited availability of multi-omic data integration tools and methods. In this paper, we discuss the interplay between gene expression and epigenetic markers/transcription factors. We show how integrating ChIP-seq and RNA-seq data can help to elucidate gene regulatory mechanisms. In particular, we discuss the two following questions: (i) Can transcription factor occupancies or histone modification data predict gene expression? (ii) Can ChIP-seq and RNA-seq data be used to infer gene regulatory networks? We propose potential directions for statistical data integration. We discuss the importance of incorporating underestimated aspects (such as alternative splicing and long-range chromatin interactions). We also highlight the lack of data benchmarks and the need to develop tools for data integration from a statistical viewpoint, designed in the spirit of reproducible research.

Keywords: ChIP-seq, data integration, gene regulatory mechanisms, RNA-seq, statistics

INTRODUCTION

High-throughput technologies have made the collection of genome-wide data in cells, tissues and model organisms easier and cheaper. These data allow one to investigate biological aspects of cell functionality and to better understand previously unexplored disease etiologies. Nowadays, RNA-seq and ChIP-seq are widely used to measure gene expression and to obtain genome-wide maps of transcription factor (TF) occupancies and epigenetic signatures (Park, 2009; Wang et al., 2009; Costa et al., 2010; Ozsolak and Milos, 2011; Furey, 2012). Several computational tools have been developed to independently analyze these data, both for single sample characterization and differential analysis (Pepke et al., 2009; Garber et al., 2011; Bailey et al., 2013). The interplay between transcriptomics and epigenomics has been widely demonstrated. Chromatin accessibility to the transcription machinery regulates gene expression and, *viceversa*, some non-coding RNAs can affect local chromatin states (Wang et al., 2011b). Such interplay has significant biomedical implications in physiological processes and pathologic states (Feng et al., 2014). Therefore, integrating ChIP-seq and RNA-seq data is a compelling need to predict gene expression during cell differentiation and development (Comes et al., 2013; Lesch et al., 2013; Malouf et al., 2013; Jiang et al., 2014; Kadaja et al., 2014) and to study human diseases, including cancer (Portela and Esteller, 2010).

The seminal work of Hawkins et al. (2010) explained why integrative omic data analysis can provide unprecedented opportunities to address some long-standing questions about genome functions and diseases. To date, large-scale data produced by ENCODE/GENCODE (ENCODE Project Consortium., 2012; Harrow et al., 2012), Cancer Genome Atlas (http://cancergenome.nih.gov/), Roadmap Epigenomics (http://www.roadmapepigenomics.org) offer the possibility to answer specific questions, as well as to raise, formulate and test novel hypotheses and questions in life science. However, despite the pros, multi-omic data integration is still one of the most challenging problems in modern science (Gomez-Cabrero et al., 2014).

In this paper we discuss the following questions: (i) how to explain and predict gene expression (and differential expression) and (ii) how to define gene regulatory network (GRN) in humans or model organisms using epigenetic data (**Figure 1**). Section Gene regulation and its impact in biology and medicine describes the biological context. Section An overview on ChIP-seq and RNA-seq data integration approaches and tools contains an overview of data visualization and integration tools. Section Statistical solutions to some biological questions illustrates the most recent statistical advances for ChIP-seq and RNA-seq data integration. Finally, Section Open biological questions and future perspectives enlightens our perspective view on the open biological questions and the tools that need to be developed in the next years. Section Conclusions reports our conclusions.

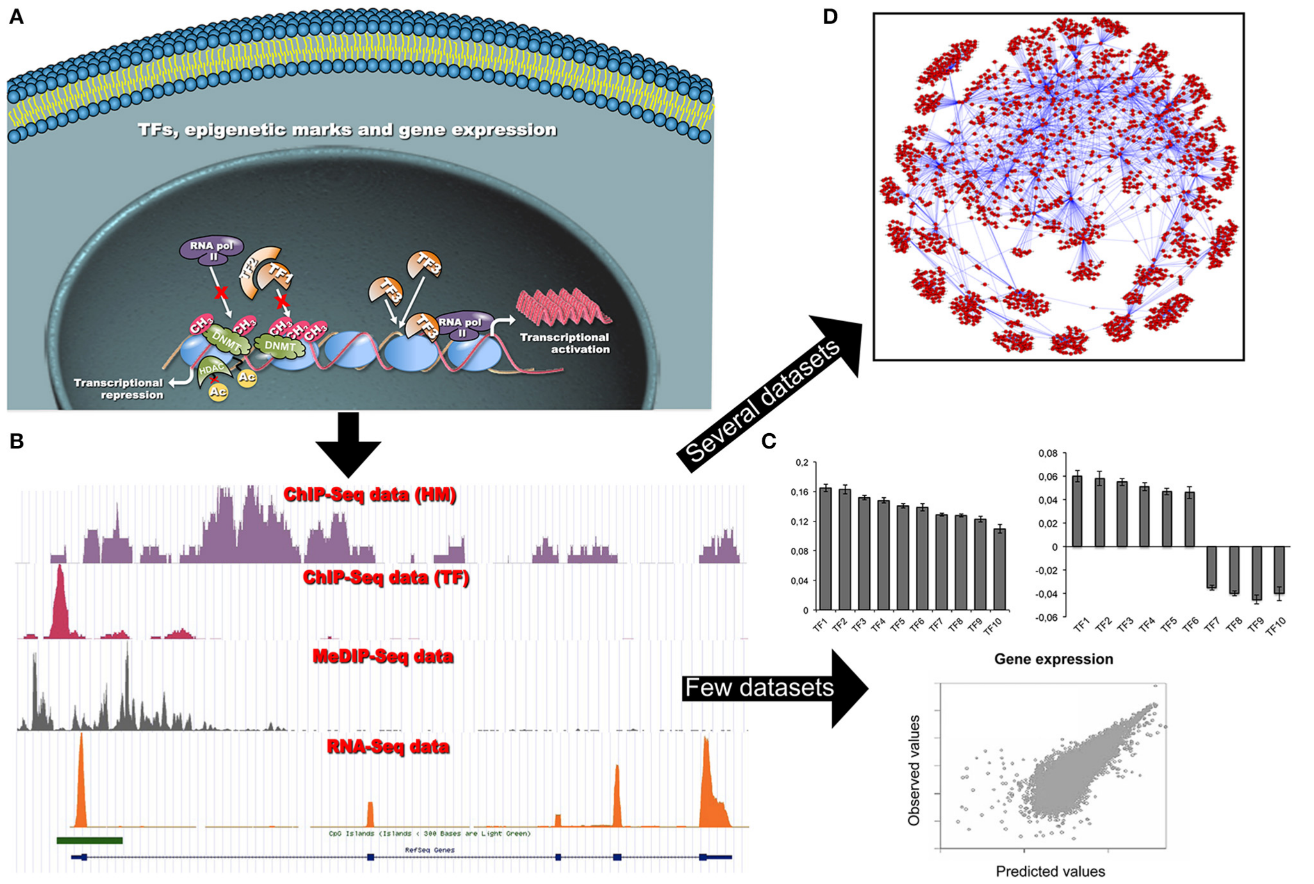


FIGURE 1 | (A) Schematic representation of the dynamic interactions among chromatin modifications and TFs, and their impact on gene transcription in a cell. Different cells share the same TF binding sites despite of differences in functionality, shape and differentiation state. Transcriptional patterns are controlled by differential TF bindings and other factors, such as local chromatin states and epigenetic modifications. These factors can limit, or promote, TF occupancies at specific *loci*, and regulate gene transcription. **(B)** Each NGS coverage data track (bedgraph format) is representative of the result of a single omic data analysis (i.e., ChIP-seq or RNA-seq experiments). The visualization of several tracks allows qualitatively studying a specific gene locus. The computational analysis of single omics allows investigating (on a genome-wide scale) different epigenetic modifications (TFs, HMs, CpG methylation, chromatin accessibility) and measuring gene expression. **(C)** When a limited amount of ChIP-seq (TF binding and/or HMs) and RNA-seq datasets are available, simple predictive models based on PCA and log-linear or support vector regression are used to predict gene expression and to reveal the most relevant epigenetic signatures able to explain the gene expression. By plotting loading factors it is possible to reveal that epigenetic signatures can act either as activators or repressors of transcription at different loci (see Section Can TF occupancies or histone modification data predict gene expression?). **(D)** In the presence of a large number of gene expression datasets more sophisticated models can be used to infer complex GRNs. This network allows visualizing TF-gene relations. In particular, it is possible to show that a given TF can control several genes and that genes are strongly interconnected (see Section Can ChIP-seq and RNA-seq data be used to infer gene regulatory networks?).

GENE REGULATION AND ITS IMPACT IN BIOLOGY AND MEDICINE

The sole nucleotide sequence of a gene does not explain its functions nor its regulation. Gene transcription is specified by DNA structure and by its accessibility to the basal transcription machinery. A physical interaction of TFs, chromatin-modifying enzymes (histone acetyl/methyltransferases and deacetylases/demethylases) and other accessory proteins with DNA is needed to modulate transcription dynamics, determining cell fate (Atkinson and Halfon, 2014). Local chromatin states and epigenetic modifications can limit, or promote, TF occupancies at specific *loci*. Several diseases can result from the alteration of chromatin remodeling and gene transcription (Portela and Esteller, 2010). Thus, understanding—and controlling—such processes may help to define potential therapies, as well as to drive cell differentiation toward specific directions.

Many efforts have been made to measure transcript levels, to detect differential expression and to identify novel alternatively spliced transcripts in various conditions (reviewed in Costa et al., 2010, 2013; Steijger et al., 2013; Angelini et al., 2014). However, regardless of the technology, a challenge is to explain and to predict gene expression by means of the coordinated binding of TFs, epigenetic marks and long-range interactions among distant chromatin domains. Recent studies demonstrate that the binding of specific TFs and some histone modifications (HMs) can be used to predict gene expression *in vitro* and to identify relevant epigenetic actors (Ouyang et al., 2009; Karlić et al., 2010; Cheng et al., 2011a, 2012; McLeay et al., 2012). Analogously, gene expression changes have been correlated to modification of TF bindings and chromatin marks (Althammer et al., 2012; Klein et al., 2014).

In general, gene expression can be predicted using a limited number of samples (in specific conditions). On the opposite, inferring large GRNs can be reached only using several high-throughput datasets, as in Gerstein et al. (2012). However, some networks can be less complicated than expected and can rely on a low number of factors and interactions. Dunn et al. (2014) recently identified a minimal set of components (12 TFs and 16 interactions) sufficient to explain the self-renewal of ES cells.

In terms of potential impact on human genetics, we highlight the following considerations. Cell differentiation is accompanied by global—and local—chromatin changes, leading to the silencing of pluripotency genes and lineage-specific gene activation (Chen and Dent, 2014). In this regard, multi-omic integration and single-cell omics can be used to explain and to potentially control differentiation and to explore heterogeneity of cells in development and disease (Comes et al., 2013; Macaulay and Voet, 2014).

Understanding such mechanisms will significantly improve the treatment of human genetic diseases, particularly of cancer. Indeed, epigenetic—unlike genetic—modifications are reversible, and modulating epi-marks through up/down-regulation of histone methyltransferases can affect gene expression and tissue-specific alternative splicing (Luco et al., 2010, 2011). By correcting the aberrant distribution of epi-marks, we may in turn control pathologic changes in gene expression (Schenk et al., 2012). In this regard, the proper identification of aberrant epigenetic regulators in tumors is of major interest. The final objective is to identify new therapeutic targets and to develop novel molecules (*epi-drugs*, inhibitors or activators of histone acetyl/methyltransferases and deacetylases/demethylases) that are able to correct or prevent aberrant epi-marks (Mai and Altucci, 2009). These interesting compounds promise to define more efficient cancer treatment strategies.

AN OVERVIEW ON ChIP-seq AND RNA-seq DATA INTEGRATION APPROACHES AND TOOLS

Data integration can be achieved with different methodologies. Genome browsers and other multidimensional visualization tools (Schroeder et al., 2013) provide integrated environments to navigate and visualize heterogeneous experimental data. Multi-omic data visualization in few loci of interest helps to formulate novel functional hypotheses. However, this is not sufficient to fully benefit from the genome-wide information that next-generation sequencing (NGS) data can provide. Naive approaches, so far used to integrate epigenetic signatures with gene expression, annotate (by proximity) either peaks or enriched regions with genes. The epigenetic profiles are displayed on the top of the gene structures. Then enriched regions are associated to pathways and gene ontologies by means of gene names (McLean et al., 2010; Statham et al., 2010; Zhu et al., 2010; Lawrence et al., 2013).

Nowadays, public repositories represent a relevant data source. Few web-based resources provide integrated information at both epigenetic and transcriptional levels, e.g., ChIP-Array (Qin et al., 2011), EpiRegNet (Wang et al., 2011a), ISMARA (Balwierz et al., 2014), and GeneProf (Halbritter et al., 2011, 2014). In particular, the latter allows one retrieving data and results of already processed ChIP-seq and RNA-seq studies; each result is connected to the workflow used to generate it. Therefore, previous results can be easily integrated with user data. Other computational platforms, such as Galaxy (Goecks et al., 2010), constitute a general framework for omic data integration.

All these approaches are very useful to summarize and visualize global information or to identify associations among different data types. However, they do not provide mathematical models for explanatory and predictive inference, as methods described in Section Statistical solutions to some biological questions.

STATISTICAL SOLUTIONS TO SOME BIOLOGICAL QUESTIONS

The questions posed in Section Introduction and illustrated in **Figure 1** are discussed in the next subsections.

CAN TF OCCUPANCIES OR HISTONE MODIFICATION DATA PREDICT GENE EXPRESSION?

The work of Ouyang et al. (2009) represents one of the first attempts to address the question using ChIP-seq and RNA-seq data and log-linear regression. In this framework, gene expression is regarded as a response variable and different TF-related features as predictors. The authors build the TF association strength matrix $\mathbf{X}$ as a weighted sum of intensities of peaks surrounding the genes of interest (**Figure 2**). They found that a remarkably high proportion of gene expression variation can be explained by the binding of 12 specific TFs. Principal component analysis (PCA) revealed that these TFs may have a dual effect. They can activate a subset of genes and repress other ones. Similarly, a simple model selection regression strategy shows that gene expression can be accurately predicted using only a small number of HMs (Karlić et al., 2010). The combined usage of different epigenetic features and chromatin accessibility data (DNase I hypersensitive sites from DNase-seq), within a log-linear regression and PCA further improves gene expression prediction (McLeay et al., 2012). More interestingly, McLeay and colleagues demonstrated that *in silico* TF binding prediction could be used as surrogate information, in absence of *in vivo* binding data.

Differently, Cheng and co-authors (Cheng et al., 2011a, 2012; Cheng and Gerstein, 2012; Dong et al., 2012) mapped each epigenetic feature into a vector of several components, measured both at the transcription starting sites (TSSs) and at the transcription termination sites (TTSs). They showed that TF binding achieves the highest predictive power in a small region centered at the TSS, whereas HMs have high predictive power in wider regions across genes. Their approach differs both for the building of the feature matrix and for the use of support vector regression. The latter does not assume a linear relationship between gene expression and signals for TFs or HMs, allowing one to capture more complex relationships. Other supervised and unsupervised statistical methods have been proposed in Xu et al. (2010); Hebenstreit et al. (2011); Park and Nakai (2011); Gagliardi and Angelini (2013). The advantage of the above-described statistical approaches is that they allow carrying out both explanatory and predictive inference.

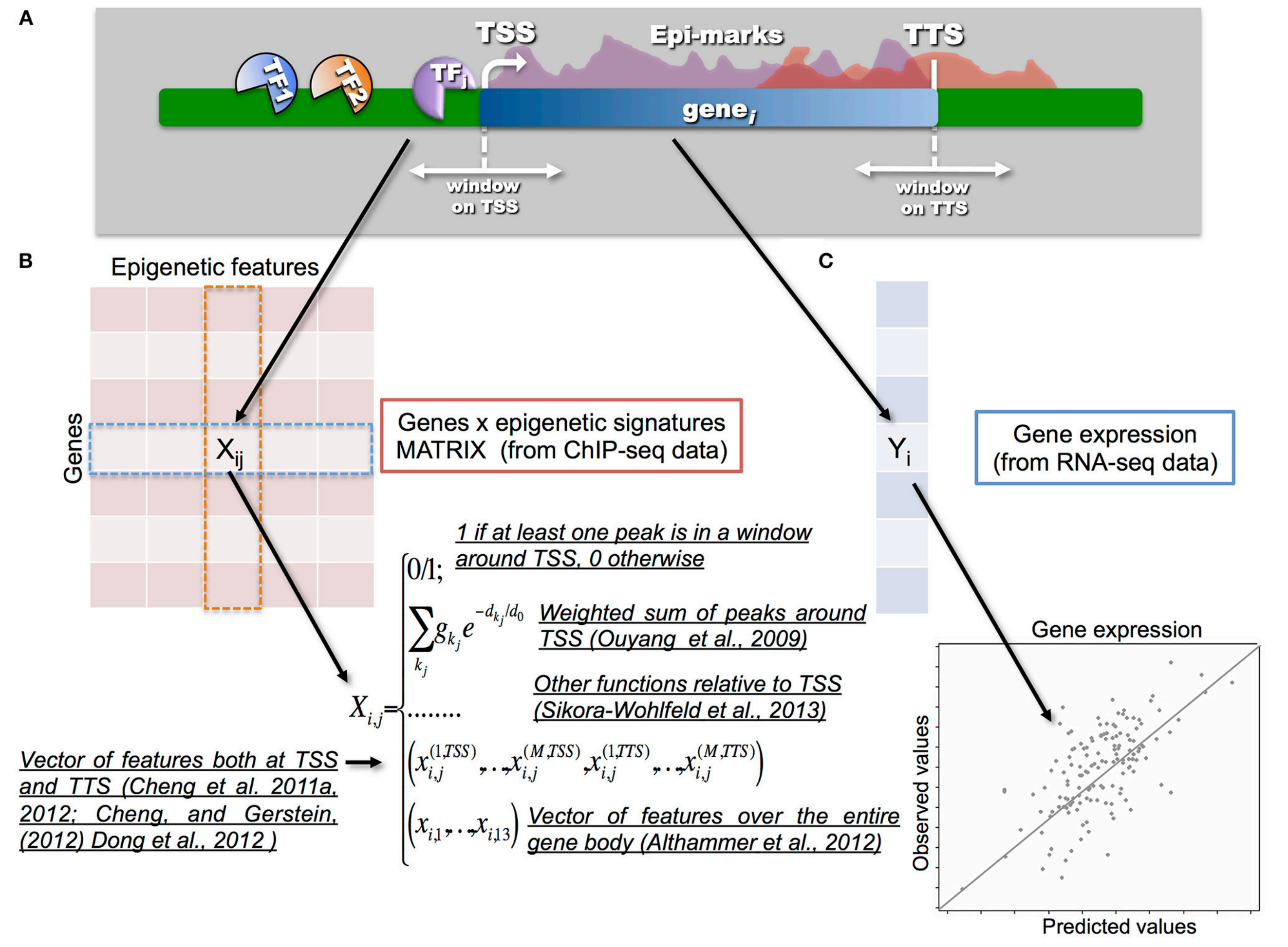

A
TSS
Epi-marks
TTS
TF1
TF2
TF$_j$
gene$_i$
window on TSS
window on TTS
B
Epigenetic features
Genes
X$_{ij}$
Genes x epigenetic signatures MATRIX (from ChIP-seq data)
C
Y$_i$
Gene expression (from RNA-seq data)
$X_{i,j}=$
0/1; 1 if at least one peak is in a window around TSS, 0 otherwise
$\sum_{k_j} g_{k_j} e^{-d_{k_j}/d_0}$ Weighted sum of peaks around TSS (Ouyang et al., 2009)
.... Other functions relative to TSS (Sikora-Wohlfeld et al., 2013)
$\left(x_{i,j}^{(1,TSS)},..,x_{i,j}^{(M,TSS)},x_{i,j}^{(1,TTS)},..,x_{i,j}^{(M,TTS)}\right)$
Vector of features both at TSS and TTS (Cheng et al. 2011a, 2012; Cheng, and Gerstein, (2012) Dong et al., 2012)
$\left(x_{i,1},..,x_{i,13}\right)$ Vector of features over the entire gene body (Althammer et al., 2012)
Gene expression
Observed values
Predicted values

FIGURE 2 | One of the key-points in the integration process is the way in which the epigenetic and transcriptional signals are transformed into a statistical model that relates a response vector Y (i.e., gene expression) with a set of predictors, represented by a matrix X (i.e., epigenetic signatures). (A) A scheme showing gene transcription, and the molecular factors involved (TFs and HMs), is illustrated in the upper part. **(B)** Different models have been proposed to build the so-called gene to epigenetic signature matrix **X**. Naive models proposed to use a binary matrix to integrate epigenetic signatures with gene expression. Therefore, 0/1 values were used to annotate and associate a given TF or HM to a specific gene according to a proximity measure between the peak and/or the enriched region and TSS of the corresponding gene. More advanced models, such as the one from Ouyang et al. (2009), proposed to use a weighed sum of peaks around the TSS. In this way it is possible to tune the strength of the binding and the distance from the TSS in a continuous way. Along the same direction, Sikora-Wohlfeld et al. (2013) compared several other measures to build **X**. All such approaches share the idea that matrix **X** is built with respect to the position of the TSSs (or using reads in a window around the TSSs) by collapsing each epigenetic feature into a single value per gene. A slightly different, and more sophisticated, approach consists in mapping each epigenetic feature into a vector of several components measured (in several bins) both at the TSSs and TTSs, as proposed in the series of papers by Cheng and colleagues. In this way, they showed that the best predictive power for TFs is indeed achieved at TSSs, however for HMs the information available at TTSs can provide further improvement. Finally, a set of 13 features for each epigenetic mark is used in Althammer et al. (2012) to classify genes as up-regulated; down regulated and no-change between two experimental conditions. The features are evaluated over the gene body, on its upstream and downstream regions (including promoters, TSSs, first exons, first introns, etc). **(C)** Gene expression Y (usually measured in terms of Fragment per kilobase of exon per million fragments mapped, FPKM) is obtained from RNA-seq data.

Previous methods focused on single biological systems for which both RNA-seq and ChIP-seq data are available. In principle, the same methods could be applied to correlate gene expression variations and changes in epigenetic mark densities between two conditions. In this context, Althammer et al. (2012) used 13 features for each epigenetic mark and a machine learning approach (based on random forest) to classify genes as up-, down-regulated or no-change when comparing two conditions. The vectors of features are extracted from TFs and HMs, and also DNase-seq and DNA methylation data. More recently, approaches based on Bayesian mixture models have been used to detect genes with differential expression and variations in the HM profiles between two experimental conditions (Klein et al., 2014).

Despite the differences in the statistical models, all the above-mentioned approaches revealed that it is possible to predict gene expression using genome-wide TF occupancies or HM data.

CAN ChIP-seq AND RNA-seq DATA BE USED TO INFER GENE REGULATORY NETWORKS?

The availability of several gene expression datasets generated from knock-out cells for one or few TFs has made possible to infer GRNs. Reconstructing GRNs using gene expression data has been one of the most widely studied problems in the last decade (Wang and Huang, 2014). However, the integration of TF occupancies data and mRNA expression values, as well as data from other transcriptional and post-transcriptional regulators, can improve methods for inferring GRNs. This task still constitutes a challenge in system biology especially for complex organisms.

ChIP-seq data were first used to determine target genes and miRNAs using data from modENCODE (Cheng et al., 2011b). Then, a regulatory network was obtained by using the correlation between TF binding and gene expression. A more comprehensive study, involving hundreds of TFs from ENCODE disclosed several structural properties of human regulatory networks (Gerstein et al., 2012). Both studies are mainly descriptive (i.e., analysis of how regulatory information is organized) and do not fully benefit from the amount of information available in terms of improving inferential approaches.

Under the assumption that network sparseness is higher in complex than in small genomes, GRN inference can be turned into a sparse optimization problem (LpRGNI, Qin et al., 2014). The identification of a small TF set that controls the network is obtained by solving a regularized lasso-type problem. The integration of ChIP-seq data improves the inference performance. As an alternative, as proposed in CMGRN (Guan et al., 2014), Bayesian network models can be first used to infer causal interrelationship among TFs and HMs (i.e., to understand how several regulators influence or associate with each other) by analyzing the sequences of regulators based on ChIP-seq read counts on the promoter of target genes. Then, Bayesian hierarchical Gibbs sampling allows integrating ChIP-based regulatory signals of TFs and HMs, microRNA binding targets with differential expression profile of genes, to construct GRN at different levels (epigenetic, transcriptional and post-transcriptional).

In general, we are far from inferring realistic quantitative models of genome-wide regulatory networks. However, it is possible to reveal the main interactions and the most relevant players. Then, computational methods can refine sub-networks for specific functions. In this spirit, Dunn et al. (2014) first generated all possible networks that could explain stem cell self-renewal. Then, by using formal verification procedures and Boolean network formalisms, they selected a core network of only 12 TFs and 16 interactions, showing that ES self-renewal relies on a relatively low number of factors and interactions.

OPEN BIOLOGICAL QUESTIONS AND FUTURE PERSPECTIVES

From a biological perspective, data integration is not *an end* to answer fundamental questions, but *a means* to generate new hypotheses. In this regard, genome-wide omic data are fundamental to drive researchers into a deeper understanding of many biological aspects (Hawkins et al., 2010).

To date, there is a limited use of multi-omic data. The association between epigenetic features and genes is still mainly done according to their proximity with respect to TSSs (with few exceptions, Althammer et al., 2012) and the existing approaches only account for local interactions. Moreover, genome-wide maps (by ChIA-PET and Hi-C) of long-range chromatin interactions and of chromatin nuclear organization have not been fully integrated in the previously described inferential models. Regression approaches in Section Can TF occupancies or histone modification data predict gene expression? are based on assumption of independence between genes, whereas the physical proximity of genes in the chromosomes in the nucleus is evidence of physical interaction. Therefore, we suggest that future computational methods for multi-omic data integration include information from genome-wide long-range interaction studies. To this aim, we propose the use of locus-by-locus interaction matrix, as a kind of correlation matrix within a regression model.

Similarly, chromatin accessibility data (Thurman et al., 2012) such as DNase-seq data, DNA regions associated with regulatory activity (FAIRE-seq), and DNA methylation data (MeDip-seq and BS-seq) should be used to better model DNA-binding background and reduce the number of false positive relations (as also suggested by Cheng et al., 2012). In such cases, we believe that the approaches described by Althammer et al. (2012) could be useful. However, the choice of the initial set of features has to be tuned according to the specific omic data at hand. Then, feature selection strategies have to be applied.

In absence of *in vivo* data, surrogate data (based on computational predictions or data from closely related cell lines or conditions) could be used to decrease experimental costs. McLeay et al. (2012) and Liò et al. (2012) showed in two different contexts that such strategy is feasible and can improve the results. Further studies should be devoted to investigate pros and cons of such approaches.

Another interesting consideration comes from the evidence that relatively few factors (TFs and/or HMs) are sufficient to explain gene expression quite accurately. Such an apparent redundancy for HMs (Cheng and Gerstein, 2012) opens the question whether such factors have a causal function or only constitute a regulatory code. Notably, such redundancy has been described only with regard to gene expression levels, without taking into account alternative splicing and differential isoform abundance. We hypothesize that the observed redundancy could partially account for a different layer of complexity, poorly explored till now. Many recent evidences indicate that some epi-marks are associated to tissue-specific alternative splicing (Luco et al., 2010, 2011; Ye et al., 2014). In this regard, the works from Chen and Dent (2014) have tried to partially overcome this issue by achieving higher predictive accuracy. Although this approach led to a higher predictive accuracy, it was not able to capture the differential expression of transcripts sharing the same TSS. We believe that a more sophisticated analysis may reveal that different combinations of epigenetic patterns can tune isoform switching (e.g., controlling the type of alternative splicing) and determine their relative abundance. The answer to such a complex question is still a challenge.

We want to underline that, despite the possibility to predict gene expression using few epigenetic features, no causal relationships can be directly inferred from such methods. The possibility of determining whether causal relationships exist or markers only constitute a code (Henikoff and Shilatifard, 2011; Cheng and Gerstein, 2012) requires developing causal inference that till now received only limited attention (Yu et al., 2008; Guan et al., 2014). In this regard, we propose Bayesian models to carry on causal inference.

Finally, while there exist several tools for data visualization (as described in Section An overview on ChIP-seq and RNA-seq data integration approaches and tools), only few tools implementing the statistical algorithms (Section Statistical solutions to some biological questions) are available. In addition, there are not general tools that allow comparing the developed methods for gene expression prediction and GRN on the same benchmarks. In light of these considerations, it is now very difficult for biologists to carry on data integration. Therefore, to facilitate biologists in such a task we strongly emphasize the need to develop new and intuitive explorative tools for the integration of ChIP-seq and RNA-seq data from a statistical viewpoint. Moreover, we firmly believe such tools should be designed in the spirit of reproducible research (Goecks et al., 2010; Russo and Angelini, 2014) to allow reproducibility and transparent verification of published results and to improve transfer of knowledge.

CONCLUSIONS

The diffusion of high-throughput technologies has offered the possibility to answer new questions, but has also posed new challenges to old problems in life science, such as data integration (Gomez-Cabrero et al., 2014). Indeed, data integration is gradually losing the merely descriptive function (as representation of data from different sources) and it is quickly acquiring inferential role. In this *scenario*, statistical methods can be used not only to analyze specific types of omic data, but also to integrate them within explanatory and predictive models. Such models can be used for further inference and to simulate the effect of specific changes *in silico*. However, to fully exploit the data available from international consortia, novel statistical methods and tools are required. In this paper, we discussed the work carried out in the last few years, and we provided our perspective about future developments.

ACKNOWLEDGMENT

Work supported by the Italian Flagship Project Epigenomics and Italian Flagship Project InterOmics.

REFERENCES

Althammer, S., Pagès, A., and Eyras, E. (2012). Predictive models of gene regulation from high-throughput epigenomics data. *Comp. Funct. Genomics* 2012:284786. doi: 10.1155/2012/284786

Angelini, C., De Canditiis, D., and De Feis, I. (2014). Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics* 15:135. doi: 10.1186/1471-2105-15-135

Atkinson, T., and Halfon, M. S. (2014). Regulation of gene expression in the genomic context. *Comput. Struct. Biotechnol. J.* 9:e201401001. doi: 10.5936/csbj.201401001

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., et al. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.* 9:e1003326. doi: 10.1371/journal.pcbi.1003326

Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M., and van Nimwegen, E. (2014). ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* 24, 869–884. doi: 10.1101/gr.169508.113

Chen, T., and Dent, S. Y. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.* 15, 93–106. doi: 10.1038/nrg3607

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* 22, 1658–1667. doi: 10.1101/gr.136838.111

Cheng, C., and Gerstein, M. (2012). Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* 40, 553–568. doi: 10.1093/nar/gkr752

Cheng, C., Yan, K. K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., et al. (2011b). Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.* 7:e1002190. doi: 10.1371/journal.pcbi.1002190

Cheng, C., Yan, K. K., Yip, K. Y., Rozowsky, J., Alexander, R., Shou, C., et al. (2011a). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.* 12:R15. doi: 10.1186/gb-2011-12-2-r15

Comes, S., Gagliardi, M., Laprano, N., Fico, A., Cimmino, A., Palamidessi, A., et al. (2013). L-proline induces a mesenchymal-like invasive program in embryonic stem cells by remodeling H3K9 and H3K36 methylation. *Stem Cell Rep.* 1, 307–321. doi: 10.1016/j.stemcr.2013.09.001

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.* 2010:853916. doi: 10.1155/2010/853916

Costa, V., Aprile, M., Esposito, R., and Ciccodicola, A. (2013). RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.* 21, 134–142. doi: 10.1038/ejhg.2012.129

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., et al. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 13:R53. doi: 10.1186/gb-2012-13-9-r53

Dunn, S. J., Martello, G., Yordanov, B., Emmott, S., and Smith, A. G. (2014). Defining an essential transcription factor program for naïve pluripotency. *Science* 344, 1156–1160. doi: 10.1126/science.1248882

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

Feng, J., Wilkinson, M., Liu, X., Purushothaman, I., Ferguson, D., Vialou, V., et al. (2014). Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome Biol.* 15:R65. doi: 10.1186/gb-2014-15-4-r65

Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* 2012, 840–852. doi: 10.1038/nrg3306

Gagliardi, F., and Angelini, C. (2013). Discovering typical transcription-factors patterns in gene expression levels of mouse embryonic stem cells by instance-based classifiers. *Lect. Notes Comp. Sci.* 8158, 381–388. doi: 10.1007/978-3-642-411 90-8_41

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100. doi: 10.1038/nature11245

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86. doi: 10.1186/gb-2010-11-8-r86

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges *BMC Syst. Biol.* 8(Suppl. 2):I1 doi: 10.1186/1752-0509-8-S2-I1

Guan, D., Shao, J., Deng, Y., Wang, P., Zhao, Z., Liang, Y., et al. (2014). CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. *Bioinformatics* 30, 1190–1192. doi: 10.1093/bioinformatics/btt76

Halbritter, F., Kousa, A. I., and Tomlinson, S. R. (2014). GeneProf data: a resource of curated, integrated and reusable high-throughput genomics experiments. *Nucleic Acids Res.* 42, D851–D858. doi: 10.1093/nar/gkt966

Halbritter, F., Vaidya, H. J., and Tomlinson, S. R. (2011). GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods* 9, 7–8. doi: 10.1038/nmeth.1809

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.1353 50.111

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11, 476–486. doi: 10.1038/nrg2795

Hebenstreit, D., Gu, M., Haider, S., Turner, D. J., Liò, P., and Teichmann, S. A. (2011). EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic Acids Res.* 39, e27. doi: 10.1093/nar/gkq1226

Henikoff, S., and Shilatifard, A. (2011). Histone modification: cause or cog? *Trends Genet.* 27, 389–396. doi: 10.1016/j.tig.2011.06.006

Jiang, L., Wallerman, O., Younis, S., Rubin, C. J., Gilbert, E. R., Sundström, E., et al. (2014). ZBED6 modulates the transcription of myogenic genes in mouse myoblast cells. *PLoS ONE* 9:e94187. doi: 10.1371/journal.pone.00 94187

Kadaja, M., Keyes, B. E., Lin, M., Pasolli, H. A., Genander, M., Polak, L., et al. (2014). SOX9: a stem cell transcriptional regulator of secreted niche signaling factors. *Genes Dev.* 28, 328–341. doi: 10.1101/gad.233247.113

Karlić, R., Chung, H. R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2926–2931. doi: 10.1073/pnas.0909344107

Klein, H. U., Schäfer, M., Porse, B. T., Hasemann, M. S., Ickstadt, K., and Dugas, M. (2014). Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics* 30, 1154–1162. doi: 10.1093/bioinformatics/btu003

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol.* 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Lesch, B. J., Dokshin, G. A., Young, R. A., McCarrey, J. R., and Page, D. C. (2013). A set of genes critical to development is epigenetically poised in mouse germ cells from fetal stages through completion of meiosis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16061–16066. doi: 10.1073/pnas.1315204110

Liò, P., Angelini, C., De Feis, I., and Nguyen, V. A. (2012). Statistical approaches to use a model organism for regulatory sequences annotation of newly sequenced species. *PLoS ONE* 7:e42489. doi: 10.1371/journal.pone.0042489

Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. *Cell* 144, 16–26. doi: 10.1016/j.cell.2010.11.056

Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science* 327, 996–1000. doi: 10.1126/science.1184208

Macaulay, I. C., and Voet, T. (2014). Single cell genomics: advances and future perspectives. *PLoS Genet.* 10:e1004126. doi: 10.1371/journal.pgen. 1004126

Mai, A., and Altucci, L. (2009). Epi-drugs to fight cancer: from chemistry to cancer treatment, the road ahead. *Int. J. Biochem. Cell Biol.* 41, 199–213. doi: 10.1016/j.biocel.2008.08.020

Malouf, G. G., Taube, J. H., Lu, Y., Roysarkar, T., Panjarian, S., Estecio, M. R., et al. (2013). Architecture of epigenetic reprogramming following Twist1-mediated epithelial-mesenchymal transition. *Genome Biol.* 14:R144. doi: 10.1186/gb-2013-14-12-r144

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., et al. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28, 495–501. doi: 10.1038/nbt.1630

McLeay, R. C., Lesluyes, T., Cuellar-Partida, G., and Bailey, T. L. (2012). Genome-wide *in silico* prediction of gene expression. *Bioinformatics* 28, 2789–2796. doi: 10.1093/bioinformatics/bts529

Ouyang, Z., Zhou, Q., and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21521–21526. doi: 10.1073/pnas.0904863106

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi: 10.1038/nrg2934

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680. doi: 10.1038/nrg2641

Park, S. J., and Nakai, K. A. (2011). A regression analysis of gene expression in ES cells reveals two gene classes that are significantly different in epigenetic patterns. *BMC Bioinformatics* 12(Suppl. 1):S50. doi: 10.1186/1471-2105-12-S1-S50

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* 6(Suppl. 11), S22–S32. doi: 10.1038/ nmeth.1371

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. *Nat. Biotechnol.* 28, 1057–1068. doi: 10.1038/nbt.1685

Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K., and Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* 67, 294–303. doi: 10.1016/j.ymeth.2014.03.006

Qin, J., Li, M. J., Wang, P., Zhang, M. Q., and Wang, J. (2011). ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res.* 39, W430–W436. doi: 10.1093/nar/gkr332

Russo, F., and Angelini, C. (2014). RNASeqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics* 30, 2514–2516. doi: 10.1093/bioinformatics/btu308

Schenk, T., Chen, W. C., Göllner, S., Howell, L., Jin, L., Hebestreit, K., et al. (2012). Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. *Nat. Med.* 18, 605–611. doi: 10.1038/nm.266

Schroeder, M. P., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). Visualizing multidimensional cancer genomics data. *Genome Med.* 5, 9. doi: 10.1186/gm413. doi: 10.1186/gm413

Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K., and Beyer, A. (2013). Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLoS Comput. Biol.* 9:e1003342. doi: 10.1371/journal.pcbi.1003342

Statham, A. L., Strbenac, D., Coolen, M. W., Stirzaker, C., Clark, S. J., and Robinson, M. D. (2010). Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics* 26, 1662–1663. doi: 10.1093/bioinformatics/btq247

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., RGASP Consortium, Abril, J. F., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. 10, 1177–1184. doi: 10.1038/nmeth.2714

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82. doi: 10.1038/nature11232

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., et al. (2011b). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 7, 120–124. doi: 10.1038/nature 09819

Wang, L. Y., Wang, P., Li, M. J., Qin, J., Wang, X., Zhang, M. Q., et al. (2011a). EpiRegNet: constructing epigenetic regulatory network from high throughput gene expression data for humans. *Epigenetics* 6, 1505–1512. doi: 10.4161/epi.6.12.18176

Wang, Y. X., and Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. *J. Theor. Biol.* doi: 10.1016/j.jtbi.2014.03.040. [Epub ahead of print].

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

Xu, X., Hoang, S., Mayo, M. W., and Bekiranov, S. (2010). Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. *BMC Bioinformatics* 11:396. doi: 10.1186/1471-2105-11-396

Ye, Z., Chen, Z., Lan, X., Hara, S., Sunkel, B., Huang, T. H., et al. (2014). Computational analysis reveals a correlation of exon-skipping events with splicing, transcription and epigenetic factors. *Nucleic Acids Res.* 42, 2856–2869. doi: 10.1093/nar/gkt1338

Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J. D. (2008). Inferring causal relationships among different histone modifications and gene expression. *Genome Res.* 18, 1314–1324. doi: 10.1101/gr.073080.107

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., et al. (2010). ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 11:237. doi: 10.1186/1471-2105-11-237

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

Received: 29 May 2014; accepted: 01 September 2014; published online: 17 September 2014.

Citation: Angelini C and Costa V (2014) Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. Front. Cell Dev. Biol. ***2****:51. doi: 10.3389/fcell.2014.00051*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

REVIEW ARTICLE
published: 19 November 2014
doi: 10.3389/fcell.2014.00070

Integrative workflows for metagenomic analysis

Efthymios Ladoukakis*[1]*, Fragiskos N. Kolisis*[1] *and Aristotelis A. Chatziioannou*[2]***

[1] Laboratory of Biotechnology, Department of Chemical Engineering, School of Chemical Engineering, National Technical University of Athens, Athens, Greece
[2] Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece

Edited by:
Paolo Tieri, Consiglio Nazionale delle Ricerche, Italy

Reviewed by:
Debashis Sahoo, Stanford University, USA
Zahraa Naji Sabra, American University of Beirut, Lebanon

****Correspondence:***
Aristotelis A. Chatziioannou, Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, 48 Vassileos Constantinou Avenue, Athens 111635, Greece
e-mail: achatzi@eie.gr

The rapid evolution of all sequencing technologies, described by the term Next Generation Sequencing (NGS), have revolutionized metagenomic analysis. They constitute a combination of high-throughput analytical protocols, coupled to delicate measuring techniques, in order to potentially discover, properly assemble and map allelic sequences to the correct genomes, achieving particularly high yields for only a fraction of the cost of traditional processes (i.e., Sanger). From a bioinformatic perspective, this boils down to many GB of data being generated from each single sequencing experiment, rendering the management or even the storage, critical bottlenecks with respect to the overall analytical endeavor. The enormous complexity is even more aggravated by the versatility of the processing steps available, represented by the numerous bioinformatic tools that are essential, for each analytical task, in order to fully unveil the genetic content of a metagenomic dataset. These disparate tasks range from simple, nonetheless non-trivial, quality control of raw data to exceptionally complex protein annotation procedures, requesting a high level of expertise for their proper application or the neat implementation of the whole workflow. Furthermore, a bioinformatic analysis of such scale, requires grand computational resources, imposing as the sole realistic solution, the utilization of cloud computing infrastructures. In this review article we discuss different, integrative, bioinformatic solutions available, which address the aforementioned issues, by performing a critical assessment of the available automated pipelines for data management, quality control, and annotation of metagenomic data, embracing various, major sequencing technologies and applications.

Keywords: metagenomics, bioinformatics, distributed computing, cloud computing, workflow engines

INTRODUCTION

Metagenomics refers to the exhaustive study of a collection of genetic material, encompassing various genomes from a mixed community of organisms as defined from the National Human Genome Research Institute (*Talking Glossary of Genetic Terms*[1]). The definition embraces the cases where either the sampling is conducted, in an environmental habitat, or the material is collected from the tissue of a particular host organism, aiming to unravel the complexity of the microbial species, which are adapted to cooperate through symbiotic modes. The scrupulous study of a metagenome (Handelsman et al., 1998) offers insight concerning not only the phylogenetic properties of the environmental niche itself, but also of its exceptionally abundant arsenal of enzymes while, at the same time, provides us with a "recipe" to recreate or even redesign them *in vitro*, for the sake of various biotechnological applications. Genomic information acquired from metagenomic sampling, has become a fundamental step for the elucidation of the taxonomic composition of the niche together with each organism's potent enzymatic capabilities and is derived through the proper analysis of the chunks of DNA sequences, i.e., the full documentation of the nucleotide sequences that constitute the metagenome that are generated from a metagenomic sequencing experiment. Sequencing techniques have greatly evolved (Metzker, 2010b) the last decade and exploiting a variety of high-throughput protocols, so as to achieve exceptionally high yields for only a fraction of the cost of traditional processes (i.e., Sanger sequencing, Sanger et al., 1977). This evolution has resulted in a massive outbreak of data that are becoming increasingly hard to process due to their size and the numerous different tools essential for each step of the analytical endeavor. A thorough analysis of a metagenomic sample requests certain successive bioinformatic tasks that comprise (i) quality control, (ii) assembly, (iii) gene detection, (iv) gene annotation, (v) taxonomic analysis, and (vi) comparative analysis, whilst storing the generated results under a database-structured computational repository enabling advanced data management, processing, mining, and meta-mining capabilities (**Figure 1**). Each stage in this succession of bioinformatic tasks necessitates substantive expertise concerning the apposite utilization of the given software tool or algorithm, something that concerns either the mathematical concepts underlying the operation of a tool, or knowledge about programming aspects of its implementation and performance. The complexity of these tasks augments radically, with

[1] http://www.genome.gov/glossary/index.cfm?id=503 [Accessed].

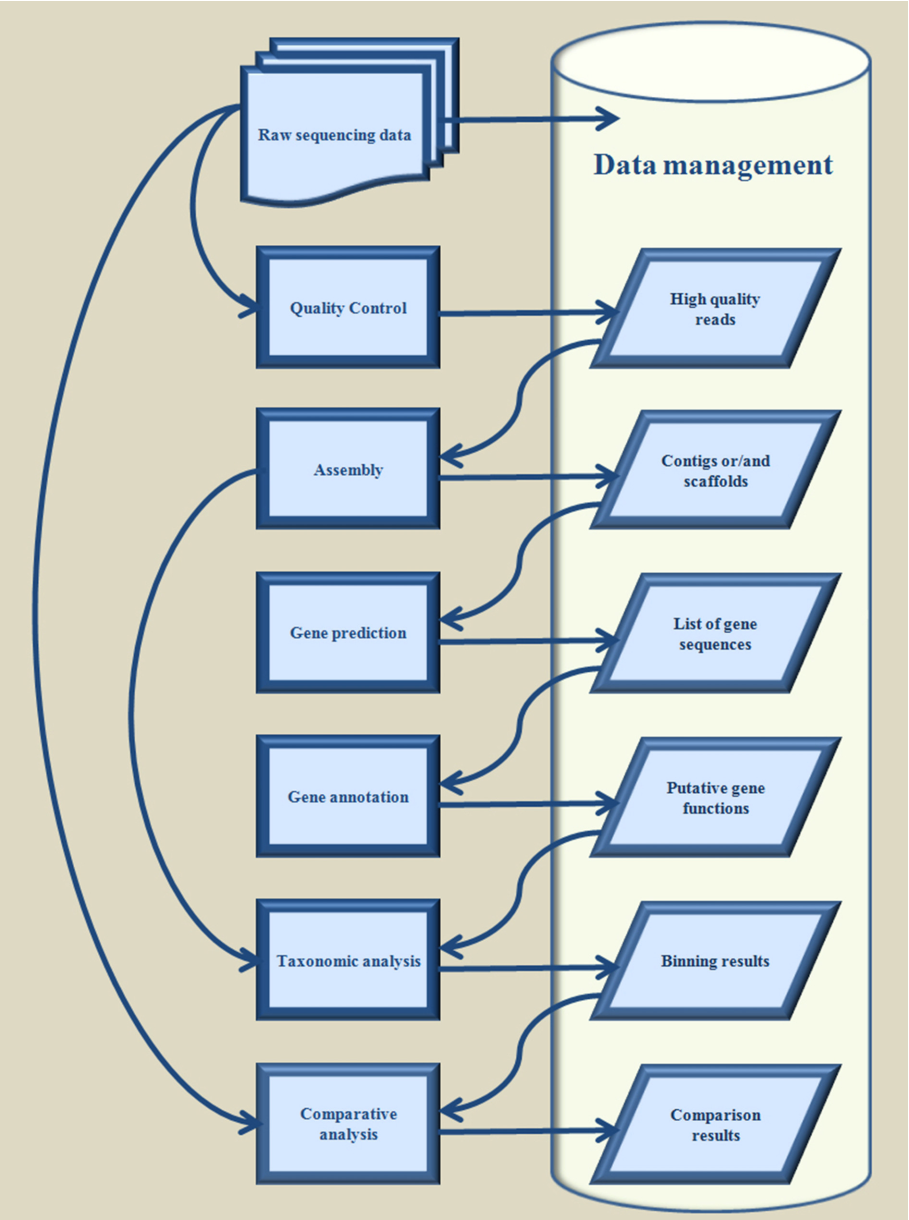
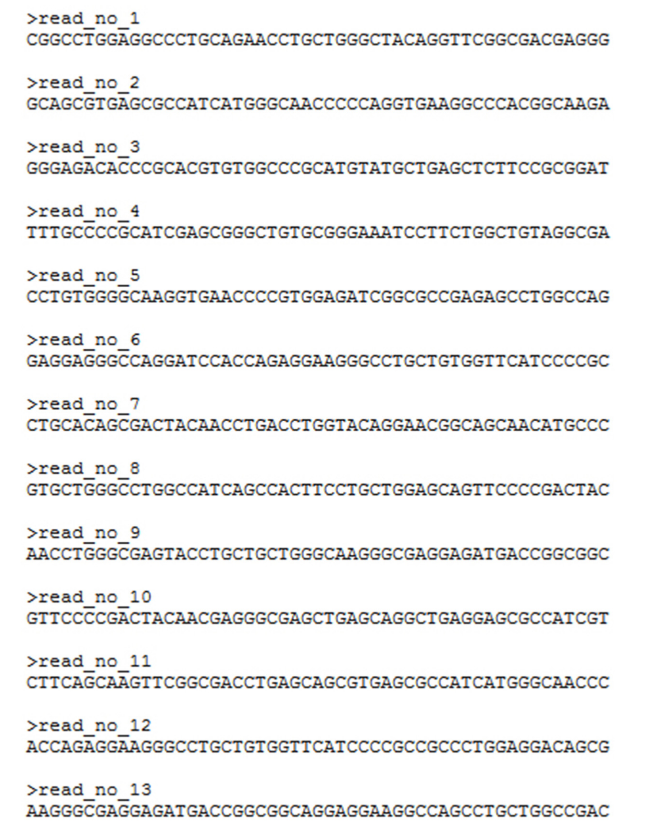


FIGURE 1 | Typical workflow for analysis of metagenomic sequencing data.

an increasing number of analyses. Recently, many bioinformatic pipelines have emerged that aim to address these issues through the provision of automated workflows and user friendly interfaces, in an effort to simplify the analytical procedure as much as possible, and minimize the entry barrier concerning the familiarization of the user with advanced programming or computational techniques. Each of these integrative analysis pipelines encapsulates a plethora of bioinformatic algorithms, seamlessly embedded into a multi-tasking framework that can address all aspects of a complete metagenomic analysis in an automated fashion. In this review we perform an appraisal of the available solutions of this kind for metagenomic purposes, by describing their configuration and their particular operational features, together with an assessment of their pros and cons, while we propose the most appropriate ones for particular analytical tasks.

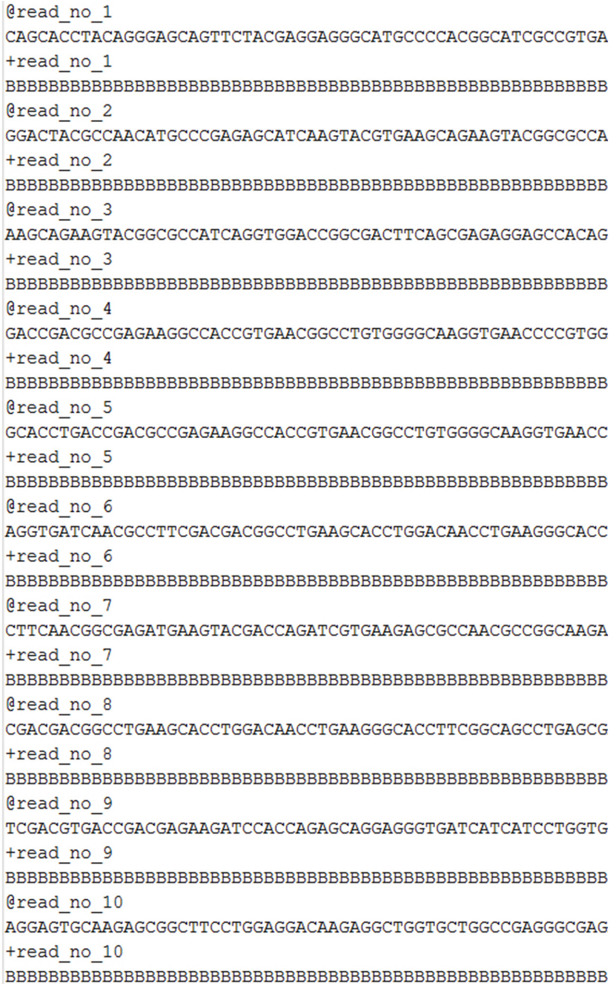
DATA ACQUISITION

There are numerous protocols available for environmental sample collection, metagenomic DNA extraction and amplification with several commercial kits available on the market. The sequencing of the acquired metagenomic DNA either with traditional sequencing techniques (Sanger sequencing) or with Next Generation Sequencing (NGS) (Metzker, 2010a) methodologies provides data in the form of small nucleotide sequences (reads) that correspond to different amplified strands of the same DNA molecule(s) each of which is randomly sheared into smaller pieces



```
>read_no_1
CGGCCTGGAGGCCCTGCAGAACCTGCTGGGCTACAGGTTCGGCGACGAGGG

>read_no_2
GCAGCGTGAGCGCCATCATGGGCAACCCCCAGGTGAAGGCCCACGGCAAGA

>read_no_3
GGGAGACACCCGCACGTGTGGCCCGCATGTATGCTGAGCTCTTCCGCGGAT

>read_no_4
TTTGCCCCGCATCGAGCGGGCTGTGCGGGAAATCCTTCTGGCTGTAGGCGA

>read_no_5
CCTGTGGGGCAAGGTGAACCCCGTGGAGATCGGCGCCGAGAGCCTGGCCAG

>read_no_6
GAGGAGGGCCAGGATCCACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGC

>read_no_7
CTGCACAGCGACTACAACCTGACCTGGTACAGGAACGGCAGCAACATGCCC

>read_no_8
GTGCTGGGCCTGGCCATCAGCCACTTCCTGCTGGAGCAGTTCCCCGACTAC

>read_no_9
AACCTGGGCGAGTACCTGCTGCTGGGCAAGGGCGAGGAGATGACCGGCGGC

>read_no_10
GTTCCCCGACTACAACGAGGGCGAGCTGAGCAGGCTGAGGAGCGCCATCGT

>read_no_11
CTTCAGCAAGTTCGGCGACCTGAGCAGCGTGAGCGCCATCATGGGCAACCC

>read_no_12
ACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGCCGCCCTGGAGGACAGCG

>read_no_13
AAGGGCGAGGAGATGACCGGCGGCAGGAGGAAGGCCAGCCTGCTGGCCGAC
```

FIGURE 2 | Raw sequence reads in FASTA format.



```
@read_no_1
CAGCACCTACAGGGAGCAGTTCTACGAGGAGGGCATGCCCCACGGCATCGCCGTGA
+read_no_1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_2
GGACTACGCCAACATGCCCGAGAGCATCAAGTACGTGAAGCAGAAGTACGGCGCCA
+read_no_2
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_3
AAGCAGAAGTACGGCGCCATCAGGTGGACCGGCGACTTCAGCGAGAGGAGCCACAG
+read_no_3
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_4
GACCGACGCCGAGAAGGCCACCGTGAACGGCCTGTGGGGCAAGGTGAACCCCGTGG
+read_no_4
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_5
GCACCTGACCGACGCCGAGAAGGCCACCGTGAACGGCCTGTGGGGCAAGGTGAACC
+read_no_5
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_6
AGGTGATCAACGCCTTCGACGACGGCCTGAAGCACCTGGACAACCTGAAGGGCACC
+read_no_6
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_7
CTTCAACGGCGAGATGAAGTACGACCAGATCGTGAAGAGCGCCAACGCCGGCAAGA
+read_no_7
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_8
CGACGACGGCCTGAAGCACCTGGACAACCTGAAGGGCACCTTCGGCAGCCTGAGCG
+read_no_8
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_9
TCGACGTGACCGACGAGAAGATCCACCAGAGCAGGAGGGTGATCATCATCCTGGTG
+read_no_9
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@read_no_10
AGGAGTGCAAGAGCGGCTTCCTGGAGGACAAGAGGCTGGTGCTGGCCGAGGGCGAG
+read_no_10
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

FIGURE 3 | Raw sequence reads in FASTQ format.

(shotgun sequencing). The generated datasets consist of text files in FASTA (**Figure 2**) or FASTQ (**Figure 3**) format containing, in the case of a typical experiment, millions of such reads, which are used for the assembly (partial or complete) of the DNA strand from which they originated. These datasets correspond to data files, which size can level, according to the depth of the sequencing analysis and quality of the instrumentation, up to several GB, thus rendering their proper processing, an elaborate, intensive task.

DEVELOPMENT OF ANALYTICAL WORKFLOWS

Despite the fact that the experimental implementation of a NGS experiment comprises a painstaking and arduous procedure, its output, namely the volumes of short sequence reads, in digitized format, represents just the initial step, for the whole analytical process, setting a point where the plethora of available data are totally illegible and non-comprehensible. In order to dig out the information hidden in these datasets, one needs to define elaborate, multi-step, bioinformatic analytical workflows that can be performed either serially or in parallel with each other. As such processing tasks are so profoundly versatile and complicated in their logical structure and programmatic development, that even an experienced team of programmers can only develop a handful of them. In this respect, the intricate nature of the various processing steps that need to be assembled together, in order to form computational workflows appropriate for different analytical tasks, strongly supports the formation of federated computational infrastructures, representing repositories of software services, that can be transparently, (namely without any knowledge about their internal architecture), integrated in the available workflows, or can compile new ones. The vision for the creation of a suitable collaborative, environment, for a long list of genomic sequence analysis tasks, representing an analog of a virtual laboratory, relies on the extent of automation, easiness in integration, transparency, and functional versatility it provides. Beneath, follows a rough account of the main processing modules, incorporated in the workflows developed for metagenomic analysis.

Quality control

The genomic (DNA) material, isolated from a metagenomic sample, is transformed through the complicated experimental DNA sequencing protocols into short sequence reads of variable length, according to the protocols and instrumentation applied (Mardis, 2008; Shendure and Ji, 2008). This base calling procedure, is susceptible to bias depending on a number of factors (Clark and

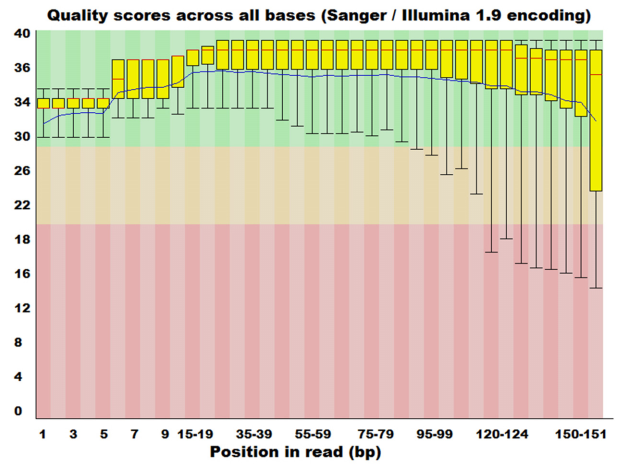


FIGURE 4 | Distribution of quality scores of raw sequence reads from FASTQC software. Taxonomic sorting of sequencing reads from MEGAN software (rank level: "species").

Whittam, 1992) such as G+C content and the actual location of the base in the sequence. This bias is quantified by measuring the probability of a base call to be false, providing an index of overall quality of the sequencing task. The computation of a quality score (Phred) (Cox et al., 2010; Schmieder and Edwards, 2011) for each sequenced base is now possible with this type of information being handily accommodated in the FASTQ file format, which represents a highly popular solution for genomic sequencing data exchange and storage, bearing both sequence and corresponding quality information (Cock et al., 2010). Several tools (Patel and Jain, 2012; Davis et al., 2013; Yang et al., 2013) have been developed that can utilize these scores and provide error probability distributions (**Figure 4**) as well as utilize appropriate filtering algorithms to trim sequences in a way that maintains only high quality genomic sequences.

Assembly

The next data processing step is the utilization of reads to assemble larger coherent sequence constructs (contigs) and, if possible, constructs that contain multiple contigs (scaffolds) with reliable connections between them. Each of these constructs originates from a different DNA sequence, that can be part of or a genome by itself and can be later investigated for the detection of open reading frames (ORFs), that is genomic areas, containing gene encoding sequences. The assembly task is so far, from the aspect of computational load, the bottleneck for any sequencing project whether the data corresponds to single cell genomes or metagenomic samples. The assembly of reads to contigs (and scaffolds) is a very laborious task, requesting avidly memory processing power resources, setting an important challenge, for which numerous algorithms (Miller et al., 2010) have been developed to address various performance issues stemming from it. Whereas there are numerous algorithms (Miller et al., 2010) dedicated to the assembly of NGS raw data, we can distinguish two discreet computational approaches; mapping reads to a template genome and *de novo* assembly. Assembly via mapping to a known genome as reference can provide very reliable results for sequencing projects dealing with single-cell samples as it can bypass performance issues originating from sequence repeats, short length of reads, low coverage of sequencing, etc. (Scheibye-Alsing et al., 2009). It is mainly driven by the choice of the reference genome which has to be as phylogenetically related to the sequenced sample as possible. *De novo* assembly is by far the most computationally intensive task (Scheibye-Alsing et al., 2009) as it requires algorithms that perform all possible comparisons between the millions of reads in order to detect any overlaps between them; a method referred to as overlay-layout-consensus (OLC). Although the *de novo* assembly endeavor has been simplified by novel algorithms abandoning the OLC method and exploiting mathematical concepts such as de Bruijn graphs (Zerbino and Birney, 2008; Peng et al., 2011), it still heavily depends on the quality of the sequencing protocol (read length, sequencing depth, etc.). Nevertheless, because of the immense diversity of the genomic content in a metagenomic sample, utilization of a reference genome is ruled out, making thus the computationally intensive task of *de novo* assembly the sole practical alternative, at least at the first steps of an analytical effort, when there is no prior knowledge about the sequences pertaining the sample.

Open reading frame/gene detection

The functional patterns, which form the response of all living organisms in an environmental niche as well as their symbiotic or competitive interactions, are encapsulated their genetic code, where all necessary information for functions such as nutrition, chemotaxis, adaptation to hostile environments and proliferation, is encoded in the form of genes. In this sense, the identification of genes within a genome, through apt mapping of each gene to its sequence or sequences, is an indispensable step, for its proper functional annotation and the decipherment of the underlying regulatory mechanisms. Computationally, the detection of genes inside a genome starts with the detection of ORFs, after their evaluation whether they can be translated into functional proteins (so that the respective nucleotide sequences may be considered as candidate gene encoding ones). The algorithms (Yok and Rosen, 2010) that perform this assessment, use various methodologies for gene prediction either from the area of machine-learning (Hoff et al., 2009; Zhu et al., 2010) or not (Noguchi et al., 2008), whereas their underlying operational features, are critically modified according to whether the gene prediction targets prokaryotic or eukaryotic organisms.

Gene annotation

Even if all gene sequences of a metagenomic population are distinguished successfully, the abundance of information they contain cannot be exploited without a proper annotation of their function. The most widespread method of annotating a gene sequence is by measuring its homology (Altschul et al., 1990; Kent, 2002) to already known genes taken from public databases (Apweiler et al., 2004; Pruitt et al., 2005; Parasuraman, 2012; Benson et al., 2014). However, as more than 99% of bacterial species cannot be cultured in the lab (Rappe and Giovannoni, 2003; Sharon and Banfield, 2013) and the quantity of metagenomic data that is generated each year continuously expands, these methods are no

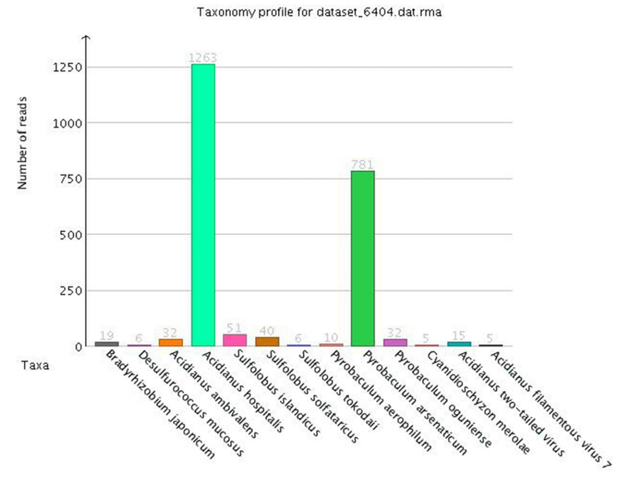


FIGURE 5 | Taxonomic sorting of sequencing reads from MEGAN software (rank level: "species").

longer sufficient to predict the function of novel genes. Instead new predictive approaches have emerged, becoming the standard practice for this sort of analysis, such as Hidden Markov Models techniques (Finn et al., 2011) and machine learning methodologies (Tian et al., 2004) that assess sequence similarity, exploiting the whole area of the sequence, seeking profiles (Claudel-Renard et al., 2003) or motives for any known gene with a given functionality, i.e., belong to the same Enzyme Commission (EC) number, rather than prioritizing serial homology.

Taxonomic analysis (binning)

An environmental niche is composed by a broad range of different microorganisms being constantly under evolutionary pressure, which have developed biological interrelations between them, as a means of symbiotic adaptation to the extreme conditions they face. As the DNA extraction from a metagenomic sample gets extracted as a whole, there is no way to separate and segregate beforehand the collected DNA, according to the organism it originated from. Nonetheless this challenge may be addressed computationally, sorting raw sequencing reads taxonomically (**Figure 5**) and phylogenetically (Weisburg et al., 1991; Retief, 2000; Darling et al., 2014) and thus yield conclusive information about the population of the niche, which can be extended subsequently to the assembled contigs and genes. This process is called taxonomic binning (Droge and Mchardy, 2012) and there are numerous tools (Mohammed et al., 2011; Pati et al., 2011; Luo et al., 2014; Wang et al., 2014) that rely on homology based or composition based approaches (Rosen and Essinger, 2010).

Comparative integrative analysis

When different metagenomic datasets are brought together, their overall diversity, which reflects the diversity in the corresponding environmental niches, can be examined computationally. The available tools (Huson et al., 2007; Markowitz et al., 2008; Meyer et al., 2008) for this task incorporate algorithms that compare the functional and taxonomical content of the different datasets and examine if the detected differences are statistically significant.

Data management

Following the massive advances of NGS technologies, the generated data from each sequencing analytical job can now reach the order of several gigabytes (GB) or even terabytes (TB) in size(Richter and Sexton, 2009). Moreover if elaborate analytical workflows like the aforementioned are applied, they yield similarly voluminous chunks of processed metadata (in some cases even at a higher order of size e.g., gene annotation). Thus, it is imperative for computational infrastructures, in the form of repositories, to integrate in a single environment, numerous algorithmic workflows that addressing versatile processing tasks together with advanced relational database management functionalities, in order to ensure easy data access, iterative comparative processing and integration of similar information from other datasets. Such infrastructures are now feasible by exploiting the potential of cloud computing (Schatz et al., 2010; Stein, 2010) and provide not only the necessary disk space for large data management but also the appropriate processing capacity for heavy duty bioinformatic tasks.

CURRENT SOLUTIONS

Each of the aforementioned tasks not only requests high processing power and storage capacity but also an in depth knowledge of regarding the proper application of computational methodologies from a broad spectrum of fields (information theory, signal processing, systems theory, statistics, programming) along with a yearlong experience in order to produce reliable results. This is why, there is an earnest need for metagenomic analysis platforms introducing automated, workflows for various processing goals, integrating tools in the form of services, operative inside processing pipelines. This has resulted into the development of various pipelines (Almeida et al., 2004; Harrington et al., 2010; Angiuoli et al., 2011) dedicated to the analysis of single organism genomic data. However, the exploitation of NGS technologies in metagenomic analysis has set off the limitations of similar solutions developed for single organism data, for the sake of metagenomic projects. Therefore, for the purposes of this review we will skip the reference to any single-genome tool and will only appraise the most recent pipelines (i.e., frameworks that incorporate two or more tools in consecutive running order) developed for the analysis of metagenomic sequencing datasets. We will also omit pipelines (Schloss et al., 2009; Caporaso et al., 2010) dedicated solely to the analysis of 16s rDNA datasets as these are targeting only phylogenetic studies (Weisburg et al., 1991; Woo et al., 2008), or CAMERA (Seshadri et al., 2007) pipeline as it is no longer supported starting from 1st of July 2014. We also exclude MEGAN (Huson et al., 2007) because despite the fact that it targets metagenomic data, it lacks critical tasks (BLASTX, taxonomic and functional analysis) as part of an automated pipeline.

The current bioinformatic arsenal of pipelines able to take up the challenge of analyzing a metagenomic sequencing dataset comprises the following tools (in alphabetical order): (i) CloVR-metagenomics (Angiuoli et al., 2011), (ii) Galaxy platform (metagenomics pipeline) (Giardine et al., 2005; Kosakovsky Pond et al., 2009), (iii) IMG/M (Markowitz et al., 2008, 2014), (iv) MetAMOS (Treangen et al., 2013), (v) MG-RAST (Aziz et al., 2008; Meyer et al., 2008), (vi) RAMMCAP (Li, 2009), and (vii) SmashCommunity (Arumugam et al., 2010).

CloVR-METAGENOMICS

CloVR-metagenomics (CloVR: Cloud Virtual Resource) is a desktop application for automated sequence analysis, which requires two different inputs; a set of fasta-formatted files (raw sequencing data), and a tab-delimited metadata file which provides sample-associated information for comparative analysis. Local installation requires a Virtual Machine (VM) player in order to boot the appropriate VM image available by their website. For a cloud-based instance, users can use the Amazon Cloud where they find an available Amazon Machine Image (AMI) from the Request Instances Wizard. The pipeline initiates by clustering redundant sequence reads with UCLUST (Edgar, 2010) and uses BLAST (Altschul et al., 1990) homology searches against COG (Tatusov et al., 2000) and RefSeq (Pruitt et al., 2005) databases for functional and taxonomic annotation respectively. The resulting data from the two different analyses are transferred as input to the integrated Metastats program for detection of differentially abundant features (White et al., 2009). Finally integrated custom scripts in R language (*The R Project for Statistical Computing*[2]) are utilized in order to normalize taxonomic or functional counts for clustering and for visualization purposes. The main advantage of CloVR's setup is that it provides the user with the option of using local resources or to access a cloud provider for additional computational capacity. A potential downside of the platform is the lack of quality control, assembly and gene detection tools (which are available only in the single-genome and 16S-rRNA versions of the software) making it highly dependent on the read length of the sequencing datasets.

GALAXY PLATFORM (METAGENOMICS PIPELINE)

Galaxy is an open-source, generic framework for the integration of computational tools and databases into a cohesive collaborative workspace, being developed primarily for data intensive biomedical research. A free Galaxy public server (*Galaxy*[3]) is available but a user can download and install an instance on his/her server for exploitation of local resources, tools and databases in order to create custom workflows. Local installation requires only the downloading of the latest release and the initiation of the local instance can be done by running the appropriate BASH *(BASH—The GNU Bourne-Again SHell*[4]*)* script (run.sh) included in the downloaded directory. A Galaxy workflow for metagenomic datasets was published (Kosakovsky Pond et al., 2009) that requires as input a single dataset of raw sequencing reads and performs an automated series of analyses exploiting specific integrated tools. Those analyses include: (i) quality control and filtering of the reads (custom tool), (ii) text editing and data format converting (custom tools), (iii) homology search against NCBI-nt database (Megablast, Altschul et al., 1990), (iv) taxonomic analysis (custom tools), and (v) visualization of results (custom tools). The biggest advantage of this platform is besides the rich collection of workflows it provides, the capability it offers, via its local installation, to each user to build customized workflows integrating any customized tools of his/her choice (third party or proprietary) that can handle a very wide range of analytical tasks, while simultaneously providing a very friendly user interface. However, in order that a full local installation is achieved, sophisticated, far from trivial, programming expertise rendering the solution inappropriate for other than proficient users. Nevertheless, as the platform becomes more and more popular, many scientific groups develop their own tools and integrate them into new workflows (Pilalis et al., 2012), rendering them available to the relevant communities of users. These workflows provide automated metagenomic analyses that cover from sequence assembly to protein annotation even enzymatic functional classification via machine learning methodologies (Koutsandreas et al., 2013).

IMG/M

IMG/M is an experimental metagenome data management and analysis system that provides a genome database from bacterial, archaeal and selected eukaryotic organisms and a suite of tools for data exploration and comparative data analysis. The data exploration tools facilitate advanced search queries in assembled sequence data for genes, for the contigs and scaffolds where they originated from as well as their associated functional characterizations (COG, Pfam, Finn et al., 2014, etc.). The comparative data analysis suite contains (i) profile-based selection tools, (ii) gene neighborhood analysis tools, and (iii) multiple sequence alignment tools that can elucidate the gene content and phylogenetic profile of any metagenomic sample. This platform constitutes a very robust and user friendly system for publishing and managing a user's (meta) genome via its web server's graphical user interface (GUI) as well as performing further functional annotation on it, while exploiting their cloud infrastructure. Nevertheless, the burden of quality control of the raw reads as well as the assembly task still befalls on the user. IMG/M is designed for assembled metagenomes only with no supporting tools for the tasks up to assembly. Local installation is not available and all users need to have an IMG Account which can be requested from IMG website.

MetAMOS

MetAMOS is a metagenomic assembly and analysis pipeline that accepts either raw sequence reads as input or already assembled contigs. Installation requires downloading the latest version and running a Python script (INSTALL.py) included in the release, which automatically handles the whole process. The modules of this pipeline make up a complete analytical workflow that includes: (i) quality control using two different tools (*FASTX-Toolkit*[5], *Babraham Bioinformatics - FastQC*[6]), (ii) sequence assembly to contigs with eight different assembly methods exploiting four different assembly tools (Zerbino and Birney, 2008; Peng et al., 2011; Treangen et al., 2011; Xie et al., 2014) and to scaffolds with Bambus 2 (Koren et al., 2011), (iii) assembly assessment using a short read aligner tool (Langmead and Salzberg, 2012) and a sequence repeats detection

[2] http://www.r-project.org/ [Accessed].

[3] https://usegalaxy.org/ [Accessed].

[4] http://tiswww.case.edu/php/chet/bash/bashtop.html [Accessed].

[5] http://hannonlab.cshl.edu/fastx_toolkit/index.html [Accessed].

[6] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed].

tool (Treangen et al., 2009), (iv) ORF/gene detection with three different available tools (Rho et al., 2010; Zhu et al., 2010; Kelley et al., 2012), (v) gene annotation with seven different available tools (Altschul et al., 1990; Bo et al., 2010; Brady and Salzberg, 2011; Finn et al., 2011; Parks et al., 2011; Darling et al., 2014), and (vi) result visualization using Krona (Ondov et al., 2011). MetAMOS's main strength is the large variety of tools that can be integrated into the workflows, in order to enable a complete automated analysis of any sort of metagenomic dataset, either it constitutes raw sequencing reads or assembled contigs and scaffolds. However, the access to its rich collection of tools is seriously hindered by the lack of a user friendly interface as all tasks must be executed from the linux command line shell, whereas their parameterization requests invocation of appropriate scripts.

MG-RAST

This pipeline supports both raw sequence reads datasets or already assembled contigs, as input. Local installation is not available as it is offered as an online service for which the user must register in order to upload metagenome datasets and to create jobs. The modules of the automated pipeline comprise four main tasks: (i) normalization of the data, (ii) parallel screening of the sequences against public databases (Maidak et al., 2001; Wuyts et al., 2002; Leplae et al., 2004; Overbeek et al., 2005; Desantis et al., 2006; Meyer et al., 2009), with predetermined default search parameters, for potential protein encoding genes and coding elements, (iii) computation of the resulting data in order to assign functional annotations and taxonomic assignments, and (iv) visualization of results via the integrated SEED Viewer. During the implementation of the pipeline, all job-relevant resulting data are incrementally stored in flat file and SQLite (*SQLite*[7]) format for optimal data management based on relational database technology. The results from the previous steps can be utilized for comparative metagenomic analysis of the original dataset against other metagenomes or complete genomes derived from the SEED environment. What makes this platform attractive to the user is that similar to IMG/M, it provides a user friendly GUI behind a web server that makes the handling of the data and its analysis as intuitive as possible. It also provides numerous tools both for functional analysis and for comparative genomics and it can handle both assembled and not assembled sequence data. The only thing missing from the pipeline are the appropriate modules for raw read quality control and assembly tasks but either than that it constitutes an easy to use and well established functional and taxonomic annotation system that fully exploits the potential of public sequence databases.

RAMMCAP

RAMMCAP (RAMMCAP: Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline) is a metagenomic platform, which workflows enable a complete metagenomic analysis, emphasizing in the programmatic optimization so that the computational cost of the various processing tasks, is minimized. Installation requires downloading the latest version of the package which includes all the essential programs, scripts, and databases. Each of the required programs of the pipeline must then be compiled and installed separately in order to be able to be called upon by the automated pipeline. This pipeline, works with raw read datasets from one or more metagenomic samples, whose sequences are clustered together using CD-HIT (Fu et al., 2012) algorithm. Parallel to clustering the reads, an ORF detection task is implemented, on the raw reads, using a local algorithm (ORF_finder) followed by yet another clustering of the resulting protein sequences. For the clustered and original amino-acid sequences, two parallel workflows are run for similarity detection against Pfam, Tigrfam, (Haft et al., 2001) (HMMER tool) and COG (RPS-BLAST tool) databases generating the subsequent annotation. The final results from (i) clustered raw reads, (ii) database results from clustered protein sequences, and (iii) database results from unclustered protein sequences are examined for statistical comparison of the metagenomes and visualization of their differences. The RAMMCAP pipeline was available as a web service via the CAMERA framework but since the latter has been discontinued it is now only available as a standalone tool for local installation. As is the case with MetAMOS, RAMMCAP's potential gets thwarted by the lack of user friendliness toward the inexperienced user. There is no GUI for the pipeline and its installation and run require a user somewhat more inclined to (bio)informatics. The lack of an integrated assembler also renders it highly dependent to the sequencing read length when it comes to the ORF detection tasks. Besides that it is considered a highly optimized solution in regards to CPU processing and memory demands for comparative metagenomic analysis.

SMASHCOMMUNITY

SmashCommunity can be considered as the metagenomic version of its predecessor SmashCell (Harrington et al., 2010), a software designed for the analysis of high-throughput single cell-amplified microbial genomes. Installing SmashCommunity requires the user to download the latest version of the package and to compile/install it using the usual BASH commands (configure, make, make install). Before installing the pipeline the user must also install a list of prerequisite programs and databases that are essential to the various modules of the workflow. This is facilitated by running the BASH scripts (e.g., install_dependencies.ubuntu.sh) included in the release. The required input for this pipeline is raw read datasets from 454 or Sanger sequencing technologies (i.e., long read sequence data). The automated workflow includes integrated tools for: (i) sequence assembly (Myers et al., 2000), (ii) gene detection (Noguchi et al., 2008), (iii) phylogenetic annotation of raw reads (Altschul et al., 1990; Wang et al., 2007; Finn et al., 2011), (iv) functional annotation of detected genes (Altschul et al., 1990; Powell et al., 2014), and (v)comparative analysis (Retief, 2000). Each tool of this workflow is integrated in the automated pipeline via a wrapper script written in Perl[8] (Stajich et al., 2002) language for facilitating the input/output (I/O) of data between different tasks. SmashCommunity can be considered an "all-inclusive" bioinformatic package but as with similar packages its greatest strength is

[7] http://www.sqlite.org/ [Accessed].

[8] http://www.perl.org/ [Accessed].

also its greatest weakness. The numerous prerequisite tools that make up the complete analytical pipeline need to be manually installed beforehand by the user adding to the complexity of the command-line only package. Plus the assembler's restrictions are passed through the rest of the pipeline making its performance optimum only with long read sequencing data (an issue that will soon be obsolete as even Illumina machines are increasing their read length output with each new sequencer release). Despite that, the most advanced user will find that it is a great solution for the conduct of complete and fully automated metagenomic analyses on a local server with dedicated resources.

DISCUSSION

In order to assess the potential of each metagenomic pipeline we take into account the range of features each pipeline introduces in order to offer an all-inclusive analysis, as well as the level of complexity of its installation. The main features of a full metagenomic analytical workflow should include: (i) sequencing quality control (ii) metagenomic assembly, (iii) ORF/gene detection, (iv) functional annotation, (v) taxonomic analysis, (vi) comparative analysis, and (vii) data management capabilities. From the pipelines we examined, only MetAmos and SmashCommunity included analytical tools for raw sequencing data whereas the rest mainly focused on detecting and annotating putative gene coding regions, as well as providing taxonomic characterization for the generated metagenome. Assessing the complexity of an installation is a fairly subjective matter, yet as "easy" we consider the installation, where the user doesn't have to perform arduous compilation and dependencies' installation tasks, since those usually require a higher level of informatics expertise. For example we consider complex for the inexperienced user, that of RAMMCAP, as it requires a manual installation of each of the integrated tools of the pipeline contrary of the installation of MetAmos, which is handled automatically through the execution of a Python script. The number of features that constitute each of the above-mentioned pipelines are summarized in **Table 1**.

CONCLUSIONS

The field of Metagenomics holds the promise for the elucidation of the genomic and taxonomic diversity of environmental niches. The rapid advances in sequencing technologies and in the development of algorithms for massive functional annotation of the analyzed genomic content intensify the capabilities of metagenomic analysis, rendering it feasible for an ever-growing number of projects. Powerful, fully automated bioinformatic pipelines lower the entry barrier to the field, through the compilation of numerous workflows, incorporating state-of-the-art algorithms optimized for specific analytical tasks, adjusted also for integration of various datasets, by resolving compatibility issues between them. There are pipelines focusing more on functional and taxonomic analysis, omitting the data-crunchy assembly part while others offer complete solutions where the user simply inputs the data from the sequencer machine and gets a fully annotated genomic report. As expected from other areas of computer science, a trade-off between user-friendliness and efficiency or flexibility of performance is observed here too. The highest the quality and the performance superiority of the workflows, the more profound knowledge they request for their impeccable installation and operation, thus minimizing their accessibility by different scientific communities, short of these skills. On the contrary, pipelines dedicated in resolving smaller, more specific processing tasks, have matured so as to provide very intuitive GUI-based solutions, often via a web server, accessible through the Internet. The broad range of integrative analysis platforms encompasses various pipelines, addressing the pressing need for disparate, versatile, complex, processing tasks. The adopted strategy for the development of efficient workflows, adjustable to varying, yet very specific every time, processing needs, posits on the modularity and transparency of the integrated code, that is the autonomous character of these modules, together with their easiness in integration and user-friendliness in their utilization. Moreover, in order to optimize the computational cost of such processing tasks, parallel processing designs are put forward, aiming to maximally exploit, multi-processor configurations. Among the examined suites of tools (**Table 1**), we believe, based in our experience for a wide range of metagenomic analysis tasks, that SmashCommunity and MetAmos represent very reliable pipelines, in terms of quality of results, reliability of operation and versatility of tools, for the most experienced users. For those who are analyzing already assembled data for the task of the functional analysis of their metagenome(s), we consider MG-RAST and IMG/M as two very robust and intuitive pipelines.

Table 1 | Display of features of current bioinformatic pipelines for metagenomic data analysis.

Pipeline Tasks	Quality control	Assembly	Gene detection	Functional annotation	Taxonomic analysis	Comparative analysis	Data management
CloVR-metagenomics	✗	✗	✓	✓	✓	✓	✓
Galaxy platform*	✓	✗	✗	✗	✓	✓	✗
IMG/M	✗	✗	✓	✓	✓	✓	✓
MetAMOS	✓	✓	✓	✓	✓	✓	✓
MG-RAST	✗	✗	✓	✓	✓	✓	✓
RAMMCAP	✗	✗	✓	✓	✓	✓	✗
SmashCommunity	✓	✓	✓	✓	✓	✓	✓

**Refers to the metagenomic pipeline of Galaxy.*

These two aforementioned workflows not only provide tools for a full analysis of any assembled metagenome, but also efficient ways for dissemination of the generated results to the scientific community through a secure database setup.

ACKNOWLEDGMENTS

The authors would like to thank Professor Anders Krogh, Dr. Peter Menzel and Hanne Munkholm from University of Copenhagen for providing access to their university server for evaluation of numerous bioinformatic tools and platforms. The work in this article, has been funded by the European EU/FP7/KBBE-2010.3.5-04 Microbial diversity and metagenomic mining for biotechnological innovation, "Systematic screening for novel hydrolases from hot environments" project HotZyme.

REFERENCES

Almeida, L. G., Paixao, R., Souza, R. C., Costa, G. C., Barrientos, F. J., Santos, M. T., et al. (2004). A System for Automated Bacterial (genome) Integrated Annotation–SABIA. *Bioinformatics* 20, 2832–2833. doi: 10.1093/bioinformatics/bth273

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Angiuoli, S. V., Matalka, M., Gussman, A., Galens, K., Vangala, M., Riley, D. R., et al. (2011). CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 12:356. doi: 10.1186/1471-2105-12-356

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119. doi: 10.1093/nar/gkh131

Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J., and Bork, P. (2010). SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* 26, 2977–2978. doi: 10.1093/bioinformatics/btq536

Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2014). GenBank. *Nucleic Acids Res.* 42, D32–D37. doi: 10.1093/nar/gkt1030

Bo, L., Gibbons, T., Ghodsi, M., and Pop, M. (2010). "MetaPhyler: taxonomic profiling for metagenomic sequences," in *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference* (Hong Kong), 95–100.

Brady, A., and Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods* 8, 367. doi: 10.1038/nmeth0511-367

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Clark, A. G., and Whittam, T. S. (1992). Sequencing errors and molecular evolutionary analysis. *Mol. Biol. Evol.* 9, 744–752.

Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* 31, 6633–6639. doi: 10.1093/nar/gkg847

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771. doi: 10.1093/Nar/Gkp1137

Cox, M. P., Peterson, D. A., and Biggs, P. J. (2010). SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485. doi: 10.1186/1471-2105-11-485

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., Bik, H. M., and Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243. doi: 10.7717/peerj.243

Davis, M. P. A., Van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* 63, 41–49. doi: 10.1016/j.ymeth.2013.06.027

Desantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Droge, J., and Mchardy, A. C. (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform.* 13, 646–655. doi: 10.1093/bib/bbs031

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

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–D230. doi: 10.1093/nar/gkt1223

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15, 1451–1455. doi: 10.1101/gr.4086505

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., et al. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29, 41–43. doi: 10.1093/nar/29.1.41

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249. doi: 10.1016/S1074-5521(98)90108-9

Harrington, E. D., Arumugam, M., Raes, J., Bork, P., and Relman, D. A. (2010). SmashCell: a software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics* 26, 2979–2980. doi: 10.1093/bioinformatics/btq564

Hoff, K. J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 37, W101–W105. doi: 10.1093/nar/gkp327

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/Gr.5969107

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 40, e9. doi: 10.1093/nar/gkr1067

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Koren, S., Treangen, T. J., and Pop, M. (2011). Bambus 2: scaffolding metagenomes. *Bioinformatics* 27, 2964–2971. doi: 10.1093/bioinformatics/btr520

Kosakovsky Pond, S., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W. Y., Taylor, J., et al. (2009). Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Res.* 19, 2144–2153. doi: 10.1101/gr.094508.109

Koutsandreas, T. G., Pilalis, E. D., and Chatziioannou, A. A. (2013). "Prediction of enzymatic activity of proteins based on structural and functional domains," in *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference* (Chania), 1–3.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leplae, R., Hebrant, A., Wodak, S. J., and Toussaint, A. (2004). ACLAME: a classification of mobile genetic elements. *Nucleic Acids Res.* 32, D45–D49. doi: 10.1093/nar/gkh084

Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 10:359. doi: 10.1186/1471-2105-10-359

Luo, C., Rodriguez, R. L., and Konstantinidis, K. T. (2014). MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* 42:e73. doi: 10.1093/nar/gku169

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T. Jr., Saxman, P. R., Farris, R. J., et al. (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.* 29, 173–174. doi: 10.1093/nar/29.1.173

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Markowitz, V. M., Chen, I. M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., et al. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* 42, D568–D573. doi: 10.1093/Nar/Gkt919

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., et al. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 36, D534–D538. doi: 10.1093/Nar/Gkm869

Metzker, M. L. (2010a). Applications of next-generation sequencing sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/Nrg2626

Metzker, M. L. (2010b). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Meyer, F., Overbeek, R., and Rodriguez, A. (2009). FIGfams: yet another set of protein families. *Nucleic Acids Res.* 37, 6643–6654. doi: 10.1093/nar/gkp698

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95, 315–327. doi: 10.1016/j.ygeno.2010.03.001

Mohammed, M. H., Ghosh, T. S., Singh, N. K., and Mande, S. S. (2011). SPHINX–an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* 27, 22–30. doi: 10.1093/bioinformatics/btq608

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000). A whole-genome assembly of Drosophila. *Science* 287, 2196–2204. doi: 10.1126/science.287.5461.2196

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396. doi: 10.1093/dnares/dsn027

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702. doi: 10.1093/nar/gki866

Parasuraman, S. (2012). Protein data bank. *J. Pharmacol. Pharmacother.* 3, 351–352. doi: 10.4103/0976-500X.103704

Parks, D. H., Macdonald, N. J., and Beiko, R. G. (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics* 12:328. doi: 10.1186/1471-2105-12-328

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* 7:e30619. doi: 10.1371/journal.pone.0030619

Pati, A., Heath, L. S., Kyrpides, N. C., and Ivanova, N. (2011). ClaMS: a Classifier for Metagenomic Sequences. *Stand. Genomic Sci.* 5, 248–253. doi: 10.4056/sigs.2075298

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2011). Meta-IDBA: a *de Novo* assembler for metagenomic data. *Bioinformatics* 27, i94–i101. doi: 10.1093/bioinformatics/btr216

Pilalis, E., Ladoukakis, E., Kolisis, F. N., and Chatziioannou, A. (2012). "A galaxy workflow for the functional annotation of metagenomic samples," in *Proceedings of the 7th Hellenic Conference on Artificial Intelligence: Theories and Applications* (Lamia: Springer-Verlag).

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., et al. (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* 42, D231–D239. doi: 10.1093/nar/gkt1253

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501–D504. doi: 10.1093/nar/gki025

Rappe, M. S., and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annu. Rev. Microbiol.* 57, 369–394. doi: 10.1146/annurev.micro.57.030502.090759

Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* 132, 243–258. doi: 10.1385/1-59259-192-2:243

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Richter, B. G., and Sexton, D. P. (2009). Managing and analyzing next-generation sequence data. *PLoS Comput. Biol.* 5:e1000369. doi: 10.1371/journal.pcbi.1000369

Rosen, G. L., and Essinger, S. D. (2010). Comparison of statistical methods to classify environmental genomic fragments. *IEEE Trans. Nanobioscience* 9, 310–316. doi: 10.1109/Tnb.2010.2081375

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467. doi: 10.1073/pnas.74.12.5463

Schatz, M. C., Langmead, B., and Salzberg, S. L. (2010). Cloud computing and the DNA data race. *Nat. Biotechnol.* 28, 691–693. doi: 10.1038/Nbt0710-691

Scheibye-Alsing, K., Hoffmann, S., Frankel, A., Jensen, P., Stadler, P. F., Mang, Y., et al. (2009). Sequence assembly. *Comput. Biol. Chem.* 33, 121–136. doi: 10.1016/j.compbiolchem.2008.11.003

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol.* 5:e75. doi: 10.1371/journal.pbio.0050075

Sharon, I., and Banfield, J. F. (2013). Genomes from metagenomics. *Science* 342, 1057–1058. doi: 10.1126/science.1247023

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi: 10.1038/nbt1486

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611–1618. doi: 10.1101/gr.361602

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* 11:207. doi: 10.1186/Gb-2010-11-5-207

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Tian, W., Arakaki, A. K., and Skolnick, J. (2004). EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.* 32, 6226–6239. doi: 10.1093/nar/gkh956

Treangen, T. J., Darling, A. E., Achaz, G., Ragan, M. A., Messeguer, X., and Rocha, E. P. (2009). A novel heuristic for local multiple alignment of interspersed DNA repeats. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6, 180–189. doi: 10.1109/TCBB.2009.9

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011). Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* Chapter 11, Unit 11.8. doi: 10.1002/0471250953.bi1108s33

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wang, Y., Leung, H., Yiu, S., and Chin, F. (2014). MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genomics* 15(Suppl. 1):S12. doi: 10.1186/1471-2164-15-S1-S12

Weisburg, W. G., Barns, S. M., Pelletier, D. A., and Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* 173, 697–703.

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Woo, P. C., Lau, S. K., Teng, J. L., Tse, H., and Yuen, K. Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin. Microbiol. Infect.* 14, 908–934. doi: 10.1111/j.1469-0691.2008.02070.x

Wuyts, J., Van De Peer, Y., Winkelmans, T., and De Wachter, R. (2002). The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* 30, 183–185. doi: 10.1093/nar/30.1.183

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30, 1660–1666. doi: 10.1093/bioinformatics/btu077

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., et al. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* 14:33. doi: 10.1186/1471-2105-14-33

Yok, N., and Rosen, G. (2010). Benchmarking of gene prediction programs for metagenomic data. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2010, 6190–6193. doi: 10.1109/IEMBS.2010.5627744

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132. doi: 10.1093/nar/gkq275

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

Received: 01 September 2014; accepted: 05 November 2014; published online: 19 November 2014.

Citation: Ladoukakis E, Kolisis FN and Chatziioannou AA (2014) Integrative workflows for metagenomic analysis. Front. Cell Dev. Biol. ***2****:70. doi: 10.3389/fcell.2014.00070*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 October 2014
doi: 10.3389/fcell.2014.00062

Integrative analysis of multiple diverse omics datasets by sparse group multitask regression

Dongdong Lin[1,2], Jigang Zhang[2,3], Jingyao Li[1,2], Hao He[2,3], Hong-Wen Deng[2,3] and Yu-Ping Wang[1,2,3]*

[1] Biomedical Engineering Department, Tulane University, New Orleans, LA, USA
[2] Center for Bioinformatics and Genomics, Tulane University, New Orleans, LA, USA
[3] Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, USA

Edited by:
Jennifer Elizabeth Dent, Quintiles, UK

Reviewed by:
Dougu Nam, Ulsan National Institute of Science and Technology, South Korea
Ayse Meric Ovacik, Merck, USA

***Correspondence:**
Yu-Ping Wang, Department of Biomedical Engineering, Biostatistics and Bioinformatics, Tulane University, New Orleans, LA 70118, USA
e-mail: wyp@tulane.edu

A variety of high throughput genome-wide assays enable the exploration of genetic risk factors underlying complex traits. Although these studies have remarkable impact on identifying susceptible biomarkers, they suffer from issues such as limited sample size and low reproducibility. Combining individual studies of different genetic levels/platforms has the promise to improve the power and consistency of biomarker identification. In this paper, we propose a novel integrative method, namely sparse group multitask regression, for integrating diverse omics datasets, platforms, and populations to identify risk genes/factors of complex diseases. This method combines multitask learning with sparse group regularization, which will: (1) treat the biomarker identification in each single study as a task and then combine them by multitask learning; (2) group variables from all studies for identifying significant genes; (3) enforce sparse constraint on groups of variables to overcome the "small sample, but large variables" problem. We introduce two sparse group penalties: sparse group lasso and sparse group ridge in our multitask model, and provide an effective algorithm for each model. In addition, we propose a significance test for the identification of potential risk genes. Two simulation studies are performed to evaluate the performance of our integrative method by comparing it with conventional meta-analysis method. The results show that our sparse group multitask method outperforms meta-analysis method significantly. In an application to our osteoporosis studies, 7 genes are identified as significant genes by our method and are found to have significant effects in other three independent studies for validation. The most significant gene SOD2 has been identified in our previous osteoporosis study involving the same expression dataset. Several other genes such as TREML2, HTR1E, and GLO1 are shown to be novel susceptible genes for osteoporosis, as confirmed from other studies.

Keywords: sparse regression, multitask learning, group lasso, significant test, osteoporosis

INTRODUCTION

Increasing amounts of high-throughput biological data have been collected to investigate the genetic mechanism underlying complex traits at different levels, e.g., genomics, transcriptomics, proteomics, and metabolomics. However, there are usually two bottlenecks for these genetic studies. One is availability of limited sample size due to the experimental cost. Small sample size can lead to the loss of detection power and the reduction of confidence on identified biomarkers. To analyze data with small sample size but large variables is still a challenging statistical problem (Hamid et al., 2009). The other is that biomarkers identified from these different studies often suffer from poor reproducibility. This issue could be caused by many factors such as differences on profiling techniques, demographic, and ancestral information of subjects, sample sizes, and quality control in these datasets (Phan et al., 2012; Song et al., 2012). To increase the power and consistency of biomarker identification, integrating the information of diverse biological datasets from different levels and platforms shows great promise and is highly demanded.

Methods for integration of diverse biological datasets include conventional meta-analysis and a variety of integrative approaches recently developed (Huttenhower et al., 2006; Liu et al., 2013). Meta-analysis is a statistical method to summarize the p-values or statistics (e.g., z-score) from each individual dataset (Evangelou and Ioannidis, 2013). There are a dozen of approaches for combing multiple p-values or statistics such as Fisher method. Meta-analysis is usually used to find common features across multiple datasets with different sample sizes and platforms but under the same hypothesis (Rhodes and Chinnaiyan, 2005). Recently, a number of integrative approaches have been developed, which are based on machine learning and statistical methods (Zhang et al., 2010; Kirk et al., 2012; Xiong et al., 2012). They can analyze multiple datasets from: (1) different platforms and levels but for the same subjects; (2) same platforms but different levels and subjects; (3) different platforms but for the same levels and subjects. They have been successfully used for various applications such as a single or a set of biomarker identification (Chen et al., 2013), gene-gene interaction prediction

(Troyanskaya et al., 2003), and genetic network construction (Balbin et al., 2013). The results in these studies demonstrate the advantage of integrating multiple diverse datasets over analyzing them individually.

In this work, we propose a novel method for integrating multiple datasets from different platforms, levels, and samples to identify common biomarkers (e.g., genes). The method was based on multitask regression model enforced with sparse group regularization, which can overcome the "small sample size, but large number of variables" problem. Multitask learning method has been successfully applied to medical imaging data fusion, where multiple types of images (e.g., CT, MRI) were combined for identifying susceptible brain regions and improving disease classification (Zhang and Shen, 2012). Among various sparse regularization terms, the use of sparse group penalty has been shown to outperform other penalties such as lasso in our previous study of pair-wise genomic data integration (Lin et al., 2013). In this study, we enforce two sparse group penalties [i.e., sparse group lasso (Friedman et al., 2010) and sparse group ridge (Chen et al., 2010a)] into the multitask regression model for data integration. We assume a regression model for each dataset as a task, and then multiple regression models will be considered as multiple tasks. Variables from all datasets will be grouped by specific units (e.g., genes). A sparse group penalty is introduced with the aims to (1) reduce dimensionality, i.e., removing a number of irrelevant genes; (2) perform group-wise feature selection, i.e., removing SNPs or expression measurements from the same gene. An effective algorithm based on alternative direction method (ADM) is proposed to solve the model. Based on the estimation of the model, a statistical test is constructed for the identification of potentially causal genes. We perform two simulation studies with both fixed and dynamic genetic effects to evaluate our sparse regression methods, which shows that our sparse group multitask regression model can increase the power of detecting risk genes by integrating multiple diverse datasets effectively. Real data analysis on four osteoporosis studies identifies some significant genes with highly susceptible to bone mineral density and osteoporosis.

MATERIALS AND METHODS

In this section, we will first introduce the sparse group multitask regression model and then propose an effective algorithm based on ADM to solve the model. Finally, a gene based statistical test is constructed to give the level of significance for each selected gene.

SPARSE GROUP MULTITASK REGRESSION MODEL

We assume T independent datasets collected from K levels of genomic data (e.g., SNP, mRNA) with $P_k (k = 1, \ldots, K)$ platforms (e.g., Affymetrix, Illumina) for each level, and thus $T = \sum_{k=1}^{K} P_k$. The number of observations in each dataset is denoted by n_i, $i = 1,\ldots T$. Sample sizes could also be different due to the diversity of protocols in each experiment. The measurement matrix of each experiment is denoted by $X^{(i)} \in R^{n_i \times d_i}$, $i = 1, \ldots, T$, where $\mathrm{d_i}$ is the dimension of features in the i-th dataset, and usually $d_i >> n_i$. These features (e.g., SNPs and mRNA expression probes) are annotated to the genes and we assume that the genes in different datasets are the same, denoted by $G = \{G_i | i = 1, \ldots Q\}$. For example, all SNPs and mRNA expressions are tested for the same set of genes G. To reduce scale differences among different levels and platforms, the features in $X^{(i)}$s will be normalized to have zero mean and unit standard deviation. The phenotypic response in each dataset is $Y^{(i)} \in R^{n_i}$, $i = 1, \ldots T$, which can be binary or quantitative trait. The study is to identify biomarkers shared by different experiments for the same phenotype. The coefficient matrix for the regression model is denoted by $\boldsymbol{C} = \left[C^{(1)'}, C^{(2)'}, \ldots, C^{(T)'}\right]'$, where $C^{(i)} \in R^{d_i}$ is the coefficient vector of the i-th model $Link\left(Y^{(i)}\right) = X^{(i)} C^{(i)}$, and $Link(.)$ is the known link function.

Multitask learning is adopted in this study for identifying the shared biomarkers across a set of distinct but correlated tasks for better accuracy. In this context, each regression model for an experiment under different level and/or platform is considered as a task. For the sake of simplicity, we assume a linear regression model for each experiment with quantitative trait (i.e., link function will be the identity matrix). The loss function for each model $L^{(i)}\left(X^{(i)}, C^{(i)}\right)$ can be derived from the negative log likelihood function and thus the total loss function for the multitask regression model is $L(X, C) = \sum_{i=1}^{T} L^{(i)}\left(X^{(i)},\ C^{(i)}\right)$.

Many conventional regression methods become ineffective for processing the large scale biological data, which usually have small sample sizes and large number of features. This issue can be addressed by introducing sparse penalty in the model. We propose a sparse multitask regression model as follows:

$$min_{\boldsymbol{C}} L(\boldsymbol{X}, \boldsymbol{C}) + \Phi(\boldsymbol{C}) \qquad (1)$$

where $\Phi(\boldsymbol{C})$ is the sparse penalty function. Two popular penalties are used: sparse group lasso and sparse group ridge, and the corresponding models are denoted by multitask-sglasso and multitask-sgridge, respectively. For multitask-sglasso, $\Phi(\boldsymbol{C}) = \lambda_1 \sum_{q=1}^{Q} \left\|\boldsymbol{C}_{\{k \in G_q\}}\right\|_2 + \lambda_2 \|\boldsymbol{C}\|_{1,\ 1}$, where $\boldsymbol{C}_{\{k \in G_q\}}$ indicates a subset of vector $\boldsymbol{C}$ corresponding to the set of features annotated to gene G_q from Ttypes of datasets and $\|C\|_{1,\ 1} = \sum_{i=1}^{T} \sum_{k=1}^{d_i} \left|C^{(i,k)}\right|$ is the l-1 norm on $\boldsymbol{C}$. This sparse group lasso penalty groups features from all datasets based on genes to perform gene level selection. The l-1 norm penalty on $\boldsymbol{C}$ can further remove those irrelevant features from each gene. This bi-level feature selection penalty has been proven to outperform several other single level sparse penalties such as lasso, group lasso, and elastic net for feature identification. For multitask-sgridge, a composite sparse penalty, i.e., group ridge penalty $\Phi(\boldsymbol{C}) = \sum_{q=1}^{Q} \left\|\boldsymbol{C}_{\{k \in G_q\}}\right\|_1^2$, is imposed on $\boldsymbol{C}$ to perform bi-level feature selection, where the features are also grouped by genes. The penalty uses the inner l-1 norm penalty on $\boldsymbol{C}_{\{k \in G_q\}}$ to achieve the sparsity within each gene while the outer ridge penalty to perform ridge regression at the gene level. This group ridge penalty has also been found to give higher power in identifying causal genes in high dimensional genomic dataset than other single level sparse penalties (Chen et al., 2010a).

In this study, we adopt these two bi-level penalties in our multitask regression models to integrate multiple diverse genomic datasets for gene-based test. Specifically, these two sparse group multitask regression models are formulated as follows:

$$\text{Multitask-sglasso: } min_{\boldsymbol{C}} \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_2 + \lambda_2 \left\| \boldsymbol{C} \right\|_{1,\,1} \quad (2)$$

$$\text{Multitask-sgridge: } min_{\boldsymbol{C}} \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_1^2 \quad (3)$$

where ω_is are the weights for the loss function of different levels of datasets, and δ_js are the weights accounting for the sample size differences among the experiments of the same type of datasets. To be more specific, ω_is reflect the prior knowledge on the importance of different levels of measurements, e.g., SNP, gene expression, and proteomics. We choose $\omega_i = 1$, $i = 1, 2, l \ldots K$ in this work, assuming that all levels of measurements contain the same important genetic information. Larger sample size is expected to provide more reliable significance test on biomarkers; therefore, the weight for the experiment under the j-th platform to measure the i-th level of genomic data is given by $\delta_j = \frac{n_j}{\sum_{j=1}^{P_i} n_j}$, $j \in P_i$, where λ_1, λ_2, and λ are the tuning parameters to control the sparsity of genes and the number of features in the models.

It could be noted that our sparse multitask regression model can be taken as the generalization of those existing sparse regression models to the representation of multiple datasets from different levels and/or platforms. For example, when $K = 1$, $P = 1$, it is sparse regression model for single dataset as used in Chen et al. (2010a) and Simon et al. (2013); when $K = 1$, $P > 1$, it can be reduced to sparse model on multiple datasets at the same level but from different platforms, similar to the work in Ma et al. (2011); when $K > 1$, $P = 1$, it can work for multiple datasets at different levels.

SOLUTION ALGORITHM BY ALTERNATIVE DIRECT METHOD (ADM)

Although both (2) and (3) are convex optimization problem with global solutions, the non-smoothness and the composite norms still cause difficulties in solving the optimization. Several algorithms have been studied to address such an issue for single task regression models, e.g., second-order cone programming (SOCP) algorithm (Candes and Romberg, 2005), spectral projected gradient method (SPGL1) (van den Berg et al., 2008), accelerated gradient method (SLEP) (Liu et al., 2009), block-coordinate descent algorithm and SpaRSA (Wright et al., 2009). In sparse multitask regression model, since the loss function is separable, these algorithms are still applicable but expensive in computations. In this study, we apply ADM to solve sparse multitask regression model. ADM uses the splitting strategy to decompose the optimization problem into several easily solvable ones and updates the variable in each subproblem iteratively until the convergence is achieved. It has been successfully applied to solve many convex or non-convex optimization problems, such as lasso (Yang and Zhang, 2011), total variation regularization (Esser, 2009), matrix decomposition and our recent work on sparse low rank decomposition (Dongdong et al., 2013). Deng et al. compared ADM with several other algorithms and found that ADM outperformed others with more robustness and faster computation (Deng et al., 2013).

Taking the model in (2) for example, we use ADM to split the penalties and transform (2) into the following optimization:

$$min_{\boldsymbol{C}} \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{V}_{\boldsymbol{1}\{k \in G_q\}} \right\|_2 + \lambda_2 \left\| \boldsymbol{V_2} \right\|_{1,1} \quad (4)$$

$$s.t.\ \boldsymbol{C} = \boldsymbol{V_1}, \boldsymbol{C} = \boldsymbol{V_2}$$

where $\boldsymbol{V_1}, \boldsymbol{V_2}$ are two variables making the loss function separable. The augmented Lagrange function can be derived as

$$L\left(\boldsymbol{C}, \boldsymbol{V_1}, \boldsymbol{V_2}, \boldsymbol{D_1}, \boldsymbol{D_2}, \lambda_1, \lambda_2, \mu, \rho\right) = \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{V}_{\boldsymbol{1}\{k \in G_q\}} \right\|_2 + \lambda_2 \left\| \boldsymbol{V_2} \right\|_{1,1} + \frac{\rho}{2} \left\| \boldsymbol{C} - \boldsymbol{V_1} - \boldsymbol{D_1} \right\|_2^2 + \frac{\rho}{2} \left\| \boldsymbol{C} - \boldsymbol{V_2} - \boldsymbol{D_2} \right\|_2^2 \quad (5)$$

where ρ is augmentedLagrangian parameter which can be updated iteratively; $\boldsymbol{D_1}$,$\boldsymbol{D_2}$ are the Lagrange multipliers to approximate the residuals between $\boldsymbol{C}$ and $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, respectively. Since the objective function and constraints are both separable and convex, ADM method is effective to solve $\{\boldsymbol{C}, \boldsymbol{V_1}, \boldsymbol{V_2}, \boldsymbol{D_1}, \boldsymbol{D_2}\}$ sequentially. We present the algorithm for solving multitask-sglasso by ADM in **Table 1**.

Remark 1. We decouple (2) into several small convex optimization problems. Step 3 is a regular least square estimation on matrix $\boldsymbol{C}$, where an analytical solution can be derived. Step 4 is a classical sparse group lasso minimization, which can be solved efficiently by block coordinate decent in Sprechmann et al. (2011). Step 5 is a simple lasso problem, which can also be solved by soft-thresholding. The division of complex optimization into

Table 1 | Algorithm of solving multitask-sglasso by ADM.

```
1   Initialization: $k = 0$, choose $\lambda_1, \lambda_2, \mu, \rho, > 0, \boldsymbol{V}_1^0, \boldsymbol{V}_2^0, \boldsymbol{D}_1^0, \boldsymbol{D}_2^0$
2   Repeat:
3   $\boldsymbol{C}^{k+1} \leftarrow argmin_A L\left(\boldsymbol{C}, \boldsymbol{V}_1^{\mathrm{k}}, \boldsymbol{V}_2^{\mathrm{k}}, \boldsymbol{D}_1^{\mathrm{k}}, \boldsymbol{D}_2^{\mathrm{k}}\right)$
4   $\boldsymbol{V}_1^{k+1} \leftarrow argmin_{\boldsymbol{V}_1} \mathrm{L}\left(\boldsymbol{C}^{k+1}, \boldsymbol{V}_1, \boldsymbol{V}_2^{\mathrm{k}}, \boldsymbol{D}_1^{\mathrm{k}}, \boldsymbol{D}_2^{\mathrm{k}}\right)$
        $= \mathrm{argmin}_{\boldsymbol{V}_1} \frac{\rho}{2} \left\| \boldsymbol{C}^{k+1} - \boldsymbol{V}_1 - \boldsymbol{D}_1^{\mathrm{k}} \right\|_2^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{V}_{\boldsymbol{1}\{k \in G_q\}} \right\|_2$
5   $\boldsymbol{V}_2^{k+1} \leftarrow argmin_{\boldsymbol{V}_2} L\left(\boldsymbol{C}^{k+1}, \boldsymbol{V}_1^{\mathrm{k+1}}, \boldsymbol{V}_2, \boldsymbol{D}_1^{\mathrm{k}}, \boldsymbol{D}_2^{\mathrm{k}}\right)$
        $= argmin_{\boldsymbol{V}_2} \frac{\rho}{2} \left\| \boldsymbol{C}^{k+1} - \boldsymbol{V}_2 - \boldsymbol{D}_2^{\mathrm{k}} \right\|_2^2 + \lambda_2 \left\| \boldsymbol{V}_2 \right\|_{1,\,1}$
6   Update Lagrange multipliers
    $\boldsymbol{D}_1^{k+1} \leftarrow \boldsymbol{D}_1^{\mathrm{k}} - \boldsymbol{C}^{k+1} + \boldsymbol{V}_1^{k+1}$
    $\boldsymbol{D}_2^{k+1} \leftarrow \boldsymbol{D}_2^{\mathrm{k}} - \boldsymbol{C}^{k+1} + \boldsymbol{V}_2^{k+1}$
7   Update iteration $k \leftarrow k + 1$
8   Until some stopping criterion is satisfied
```

several simple sub-optimizations will improve the efficiency of computation.

Remark 2. We adopt the stopping criterion as suggested by Boyd et al. (2010) that both primal res$_{\text{pri}}$ and dual res$_{\text{dual}}$ residuals must be small, i.e., $res_{pri} \leq \varepsilon_{pri}$, $res_{dual} \leq \varepsilon_{dual}$, where primal residual indicates the difference between $\boldsymbol{C}$ and $\boldsymbol{V_1}$ ($\boldsymbol{V_2}$) while dual residual measures the difference between $\boldsymbol{V_1}$ ($\boldsymbol{V_2}$) and the values at the last iteration.

Remark 3. The convergence rate depends on the choice of Lagrangian parameter ρ. Some studies adjust ρ based on primal and dual variables iteratively to speed up the convergence. In this work, we update ρ by keeping the ratio between primal and dual residual norms within a given interval, say [0.1, 10] until they both converge to zeros.

For optimization (3), it can similarly be transformed into ADM formulation where only one splitting variable (i.e., $\boldsymbol{V_1}$) is needed to separate (3) into two subproblems. The estimation of $\boldsymbol{V_1}$ at Step 4 can be replaced by:

$$\boldsymbol{V_1^{k+1}} \leftarrow argmin_{\boldsymbol{V_1}} \frac{\rho}{2} \left\| \boldsymbol{C^{k+1}} - \boldsymbol{V_1} - \boldsymbol{D_1^{k}} \right\|_2^2 + \lambda \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_1^2 \tag{6}$$

where soft-threshold can be used to get the solution.

STATISTICAL TEST

λ_1, λ_2, and λ are tuning parameters used to control the number of genes and features within a gene. The K-fold cross validation is widely used to select optimal values of these parameters. Briefly, the subjects are divided into k groups, where k−1 groups of subjects are used for estimating the coefficient matrix $\boldsymbol{C}$ and the rest group of subjects is used to calculate the prediction error by the estimated $\boldsymbol{C}$. We set λ_1, λ_2, and λ to $[10^{0.1}, 10^{0.2}, \ldots, 10^3]$ with 30 values. We search the 30 × 30 grid to find an optimal combination of $(\lambda_1^*, \lambda_2^*)$ for multitask-sglasso and similarly optimal value of λ^* for multitask-sgridge by 5-fold cross validation. Finally, the estimate of $\boldsymbol{C}$ can be calculated by the derived optimal parameters.

To test the significance of identified biomarkers with non-zeros coefficients at $\boldsymbol{C}$, we construct a gene based statistical test to measure the strength and significance of the association between genes and phenotype across experiments from different platforms and levels. For the i-th gene G_i, $\left\{\widehat{\mathbf{C}}_i^{(j)} | j = 1,\ 2, \ldots,\ T\right\}$ indicates the corresponding coefficient vector estimated from the j-th experiment, denoted by $\left[\hat{C}_{i,\,1}^{(j)}, \hat{C}_{i,\,2}^{(j)}, \ldots, \hat{C}_{i,\,m_i}^{(j)}\right]$, where m_i is the number of features annotated to gene $\mathrm{G_i}$ in the j-th experimental dataset. The null hypothesis is there is no association between the i-th gene and phenotype in all T experiments, denoted by $H_0 : \left[\hat{C}_i^{(1)'}, \hat{C}_i^{(2)'}, \ldots, \hat{C}_i^{(T)'}\right]' = 0$, vs. the alternative hypothesis $H_A : \hat{C}_i^{(k)} \neq 0,\ k = 1,\ 2, \ldots, T$ for some k. To test the hypothesis, we summarize the coefficients of the i-th gene on all datasets as follows.

$$\hat{S}_i = \sqrt{\sum_{j=1}^{T} \left\| \hat{C}_i^{(j)} \right\|_2^2} \tag{7}$$

where $\hat{S}_i$, $i = 1, 2, \ldots, Q$ is the statistical value on all Q genes. Due to different number of features included in different genes, an adjustment for gene size is necessary. A permutation based approach is used to reduce the potential bias due to varying gene size. The standardized gene level statistic is given by

$$\tilde{S}_i = \frac{\hat{S}_i - \hat{S}_i^0}{\widehat{\sigma}_i} \tag{8}$$

where $\hat{S}_i^0$ and $\widehat{\sigma}_i$ are the mean and standard deviation of the i-th gene under the null hypothesis. Samples are permuted B times to construct null distribution of $\hat{S}_i$, denoted by $\widehat{\Gamma}_i^0 = \{\hat{S}_{i,j}^0 | j = 1, 2, \ldots, B\}$. $\hat{S}_i^0$ and $\widehat{\sigma}_i$ are then estimated based on permutation data. Since all $\hat{S}_{i,j}^0$ have been normalized, we could pool all $\widehat{\Gamma}_i^0$ into a set $\Gamma^0 = \{\widehat{\Gamma}_i^0 | i = 1,\ 2, \ldots,\ Q\}$ as the estimated null distribution. Therefore, the gene-level p-value of the i-th gene can be calculated by

$$p_i = \frac{\#\ of\ \{\Gamma^0 \geq \hat{S}_i\}}{\#\ of\ \{\Gamma^0\}} \tag{9}$$

SIMULATION

To evaluate the performance of our proposed integrative method for identifying biomarkers, we simulated two levels of measurements: SNP and gene expression, and assigned different sample size for each dataset.

For each simulation, we generated 3 SNP datasets and 3 gene expression datasets. The sample sizes were 600, 400, and 200 for SNP data and 70, 50, and 30 for gene expression, respectively. 200 genes were simulated in each dataset. To mimic the linkage disequilibrium (LD) structure among SNPs, we chose a chromosome, chromosome 22, from HapMap CEU panel with phase III data and sample subjects by software HAPGEN2 (Su et al., 2011). Those SNPs were kept after the following filters were applied: (1) Minor allele frequency (MAF) at least 5%; and (2) Hardy-Weinberg Equilibrium (HWE) with significant level less than 0.001. We generated a dataset consisting of 15,235 SNPs which were assigned to 576 genes as the gene pool. Assuming an additive genetic model, each SNP was recorded as the count of minor allele (denoted as A) at that locus and thereby was valued by 0 (homozygote of major allele, aa), 1 (heterozygote, Aa) and 2 (homozygote of minor allele, AA). 200 genes including more than 10 SNPs were randomly selected from the pool, of which 20 genes were chosen as causal genes and 2 SNPs with MAF from uniform distribution (Unif) (0.15, 0.25) from each causal gene were further used to induce causal genetic effects on gene expression. The number of SNPs from 200 selected genes was randomly set from Unif(10,100) and those non-causal SNPs in each gene were selected from pooled SNPs.

We used SNP data to generate gene expression and phenotype data, referring to the similar method in Huang et al. (2014). Three SNP datasets with 70, 50, and 30 subjects were first simulated, as described in the method section. For each causal gene, e.g., gene i, the expression value G_i was derived from the causal SNPs in this gene by

$$G_i = \sum_{j=1}^{n} SNP_{causal}^{j}\beta_j + \varepsilon \tag{10}$$

where n was the number of causal SNPs included in G_i; and β_j indicated the effect of the j-th causal SNP(SNP_{causal}^{j}) on G_i. We set β value from Unif(1, 1.2) and noise ε from normal distribution $N(0, 1)$. The other non-causal gene expression values were generated by multivariate normal distribution $N(0,\ \Sigma)$, where Σ was the covariance matrix of gene expressions, and the expressions of gene i and j have correlation coefficient $0.3^{|i-j|}$. Based on the simulated gene expression, the phenotype was generated by the following formula:

$$logit\{Pr(Y_i = 1)\} = \sum_{j=1}^{m} G_{causal}^{j}\tau_j + \varepsilon^{'} \tag{11}$$

where m was the number of causal genes, i.e., $m = 20$ in this study; G_{causal}^{j} was gene expression for the j-th causal gene and τ_j was the corresponding effects on the outcome. The logit function was used to generate binary outcome. The identity function can be used if the quantitative phenotype was used. $\varepsilon^{'}$ was non-genetic variable, which was assumed to follow normal distribution $N(0, 1)$.

RESULTS

SYNTHETIC DATA

We assessed the performance of the two proposed sparse multitask models- multitask sglasso and multitask sgridge-on each single dataset and all datasets, respectively, and also compared them with widely used meta-analysis on three SNP datasets (meta-SNP) and three gene expression datasets (meta-EXP). Meta-analysis was implemented by the software MetaL (Willer et al., 2010).

Simulation 1: Fixed effect of causal genes in diverse dataset

In this simulation, we studied the scenario that the effects of causal genes across diverse datasets were fixed, i.e., $\tau_j^1 = \tau_j^2 = \cdots = \tau_j^6, i = 1, 2, \ldots, m$, which indicated a causal gene had the same effect on all datasets. For m casual genes, first, we set a baseline vector $\boldsymbol{\eta} \in R^m$ from Unif(0.2, 2) and Unif(−2, −0.2). Next, to evaluate the performance of different methods on identifying casual genes under different levels of effects, a factor $\delta = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0$ was multiplied by $\boldsymbol{\eta}$ to have the final value of gene effects $\boldsymbol{\tau} = \boldsymbol{\eta} \times \delta$. 50 replicates were performed and $B = 500$ permutations in each replicate were implemented to calculate empirical p-value of sparse multitask models. Finally, we compared the results of the following eight cases: multitask-sglasso on three expression datasets, three SNP datasets, and all six datasets; multitask-sgridge on three expression datasets, three SNP datasets and all six datasets; meta-analysis on three SNP datasets and three expression datasets.

Figure 1 shows the comparison result of a set of methods under different values of δ, i.e., [0, 0.2, 0.4, 0.6, 0.8, 1.0]. The ROC curves were plotted using the false positive rate against true positive rate by varying the p-value threshold from 10^{-4} to 1. It could be seen that all methods had similar performance when there were no effective causal genes in all datasets (i.e., $\delta = 0$). When the effects of causal genes (i.e., δ) increase, i.e., more variability of phenotypes could be explained by genetic variants, multitask-sglasso method shows better performance by removing the irrelevant genes with improved signal to noise ratio. When δ was greater than 0.2, multitask-sglasso methods on SNP,

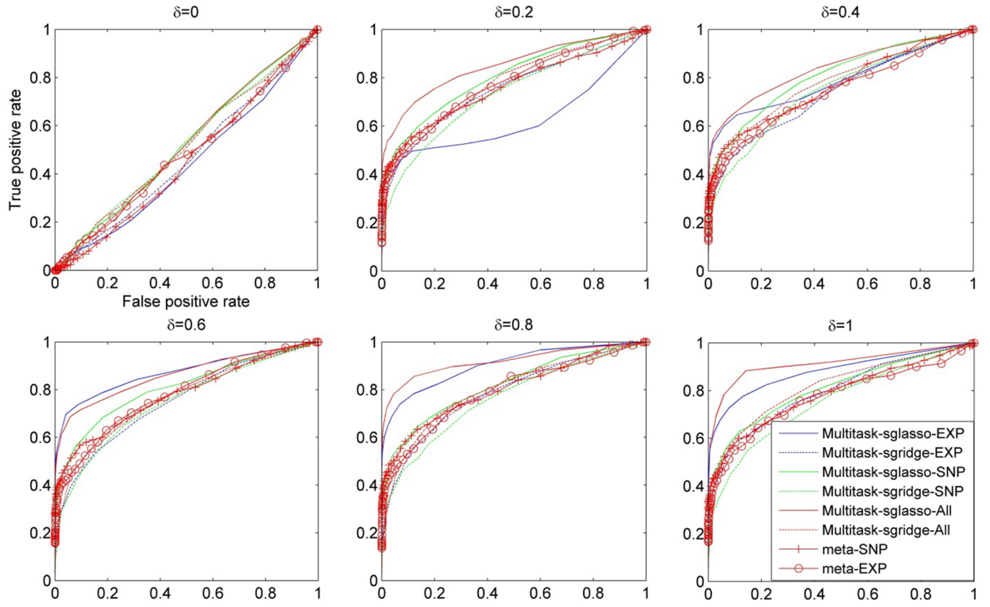


FIGURE 1 | The ROC curves for the comparison of eight cases: sparse multitask-sglasso and multitask-sgridge methods on three SNP datasets, expression datasets and all datasets, and meta-analysis on SNP and expression datasets, respectively.

expression and both datasets significantly outperformed the other methods. This indicates that Multitask-sglasso method showed better performance by integrating all datasets than that of using only one level of data. In addition, when δ was greater than 0.4, multitask-sglasso method using only SNP or expression datasets still gave higher power than meta-analysis method. Multitask-sgridge method had less power than multitask-sglasso method and only showed better performance than meta-analysis method when causal genes have high effect sizes.

Simulation 2: Dynamic effects of causal genes in diverse datasets

In this simulation, we consider the situation that a causal gene has different effects at different levels and platforms. This is more likely to happen for real datasets since multiple datasets are usually generated from different studies with different study protocols, profiling techniques, and experimental platforms, leading to dynamic effect sizes of casual genes. We aimed to compare the performance of our sparse multitask methods with meta-analysis for biomarker identification in this dynamic case. Six datasets were generated with the same sample size and causal genes as those in the first simulation study. We simulated the dynamic effects of causal genes at different datasets by setting $\tau_j \sim N\left(\eta, \sigma^2\right), i = 1, 2, \ldots, 6$, where η was fixed effect as described above, and σ was standard deviation indicating the dynamic effect of genes across datasets. We changed the value of σ from 0 to 1 with the interval of 0.2 to show different extent of heterogeneity of causal genes across diverse datasets. 50 replicates were averaged to draw the ROC curve for comparison.

Figure 2 showed the comparison result of eight cases under dynamic effect models with variance of causal genes varying from 0 to 1. When $\sigma = 0$, the models reduced to the ones with fixed effects. When σ was greater than 0.4, sparse multitask-sglasso method on SNP, expression and both datasets significantly outperformed other methods in identifying casual genes. Except for sparse multitask-sglasso method, we can also see that the performance of sparse multitask-sgridge on all datasets was better than meta-analysis methods, which indicated the advantage of multitask method for integrating diverse datasets.

REAL DATA ANALYSIS

In this study, we took advantage of 3 gene expression datasets and 1 GWAS dataset with bone mineral density (BMD) measurements from our previous studies. The cohort I of gene expression data contained 80 Caucasian females, including 40 high and 40 low hip subjects (Chen et al., 2010b). The cohort II of gene expression data contained 19 Caucasian females, including 10 high and 9 low hip BMD subjects (Liu et al., 2005). The cohort III of gene expression data contained 26 Chinese females, all premenopausal and including 14 high and 12 low hip BMD subjects (Lei et al., 2009). For the GWAS dataset, SNP data were obtained using Affymetrix 500K arrays on 1,000 unrelated homogeneous Caucasians. After a suite of quality control procedures were performed, the SNP set for subsequent analysis contained 379,319 SNPs, yielding an average marker spacing of ~7.9 kb throughout the human genome (Xiong et al., 2009).

We combined gene expression and SNP datasets to identify those risk genes of BMD by our sparse multitask-sglasso integrative method. We chose one chromosome 6 containing the largest number of genes to perform gene-based analysis. 504 genes were included in the chromosome. More details in each dataset were given in **Table 2**.

We applied sparse multitask-sglasso method to SNP, gene expression and both datasets, respectively. To compare with meta-analysis, two gene expression datasets with the same level and experimental platforms, EXP-19 and EXP-80, were used for meta-analysis, denoted by meta-Exp. The most significant expression

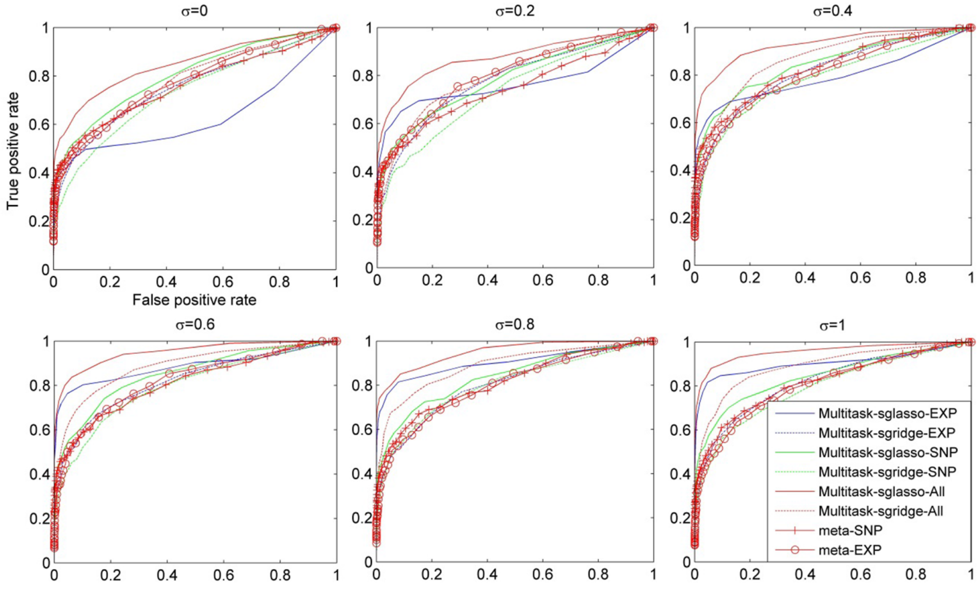


FIGURE 2 | The comparison of eight methods on three SNP and three expression datasets simulated with the dynamic model. The variance of effect size of causal genes is set to normal distribution with variance varying from 0 to 1 at an interval of 0.2.

Table 2 | A summary of four datasets from different levels and platforms used in this analysis.

Data type	Platform	Gene	Genetic variants	Sample
SNP	Affymetrix 500K	504	10685	1000
Gene expression	HG-U133A	504	874	19
Gene expression	HG-U133A	504	1225	26
Gene expression	HG-U133A-Plus_2.0	504	874	80

measurement in each gene was chosen to represent significance level of the gene. **Figure 3** shows the Venn diagram of gene list by three methods: multitask-sglasso on all gene expression datasets, multitask-sglasso on all gene expression and SNP datasets, and meta-analysis on two expression datasets under the significant threshold 0.05. We could see that there were 45 genes shared by meta-Exp and multitask-sglasso on three expression datasets; 10 genes overlapped by meta-Exp and multitask-sglasso on both SNP and expression datasets; and three genes ("GPR116," "HLA-DMB," "PHACTR1") identified by all methods. The small overlapping between multitask-sglasso Exp and multitask sglasso SNP + Exp is due to the use of additional information from large sample size of SNP dataset.

Table 3 lists 7 top significant genes identified and sorted by their *p*-values from sparse multitask-sglasso method on all datasets and the corresponding *p*-values by meta-analysis. Note that the *p*-values of the same gene usually were different in different studies. For example, SOD2 had much lower *p*-values in SNP and EXP-26 datasets than those in other datasets. This difference showed the dynamic effects of genes across diverse datasets with different levels and platforms. There are three genes ("TREML2," " HTR1E," and "GLO1") shared by sparse multitask-sglasso method on all of datasets and meta-Exp. Except for gene TREML2, the *p*-values of genes derived from all datasets were lower than those from the other methods, indicating higher level of significance given by our multitask method. The relatively smaller *p*-values of these genes in SNP data were due to the large sample size of SNP dataset, which will give more confidence on the findings.

To further evaluate the significance of identified genes by multitask-sglasso, we performed gene level meta-analysis on three independent BMD studies for validation, more details were shown in supplementary data. The result (Table S1) listed the *p*-values of 24 identified genes based on single studies and meta-analysis. Most of these genes showed significant effects on BMD ($p < 0.01$), indicating the effectiveness of our sparse multitask regression method in identifying genetic risk factors.

Three shared genes ("TREML2," "HTR1E," and "GLO1") may have important biological functions related to BMD associated with osteoporosis. TREML2 (also known as TLT-2) was located in a gene cluster on chromosome 6 with the single Ig variable (IgV) domain activating receptors TREM1 and TREM2, while these TREM receptor families were found to participate in the process of bone homeostasis by controlling the rate of osteoclastogenesis and regulating the differentiation of osteoclasts (Klesney-Tait et al., 2006; Otero et al., 2012). HTR1E was recently identified to contain SNPs significantly associated with a linear combination of multiple osteoporosis-related phenotypes including BMD (Karasik et al., 2012). GLO1, as a binding protein of methyl-gerfelin (M-GFN), was found to be able to result in the inhibition of osteoclastogenesis (Kawatani et al., 2008). Besides these three common genes, our method was also able to identify other osteoporosis-susceptible genes but was undetectable by meta-analysis. For instance, SOD2 has been identified as the gene susceptible to osteoporosis in our previous integrative analysis of mRNA, SNP, and protein data (Deng et al., 2011). It may play a significant role in BMD variation and pathogenesis of osteoporosis. HDAC2, as a member of histone deacetylases (HDACs), was found to play a critical role in bone development and biology (McGee-Lawrence and Westendorf, 2011). These genes were missed out with meta-analysis but can be detected with our proposed method, showing improved sensitivity.

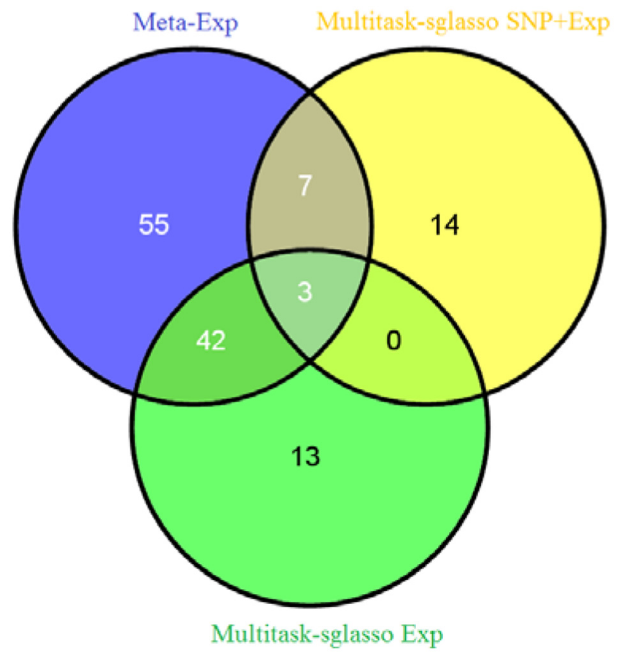


FIGURE 3 | The Venn diagram of identified genes by three methods: meta-analysis on EXP-19 and EXP-80 datasets, multitask-sglasso on all three expression datasets and multitask-sglasso on all gene expression and SNP datasets.

CONCLUSION AND DISCUSSION

In this work, we proposed a multi-omics integration method, i.e., sparse group multitask regression model, which can integrate multiple genomic datasets from different levels, platforms, and subjects for gene based analysis. An efficient computational algorithm based on ADM was provided for its solution. The performance of the model was compared with meta-analysis in simulation datasets. The simulation results showed that our sparse group multitask regression model can increase the power of detecting risk genes by integrating multiple diverse datasets effectively. In particular, multitask-sglasso model outperformed

Table 3 | The top 7 identified genes and their *p*-values by sparse multitask-sglasso method in bone mineral density studies.

Methods Gene ID	SNP	EXP-19	EXP-26	EXP-80	EXP_all	Meta-EXP	SNP + EXP
SOD2	0.0021	0.9136	0.0017	0.9566	0.7152	0.0752	0.0016
TREML2*	0.0014	0.1295	0.5243	0.1648	0.1665	0.0312	0.0018
HTR1E*	0.0030	0.4062	0.3481	0.0963	0.0750	0.0203	0.0023
HDAC2	0.0067	0.0089	0.1118	0.4382	0.4360	0.0553	0.0032
HCRTR2	0.0045	0.1074	0.5972	0.3293	0.3282	0.6297	0.0044
MUT	0.0073	0.2173	0.7665	0.9763	0.9910	0.571	0.0055
GLO1*	0.0084	0.0651	0.6182	0.1012	0.1298	0.0183	0.0073

** Genes identified by both meta-Exp and sparse multitask-sglasso on all datasets.*

meta-analysis method in simulations on genes with both fixed and dynamic effects. Our real data analysis on osteoporosis studies identified significant genes but missed by meta-analysis, and these genes were reported to be highly susceptible to BMD and osteoporosis. Overall, the advantages of our sparse group multitask regression method for biomarker identification from multiple omics datasets include: (1) it can combine diverse and complementary omic datasets without; (2) group the features by gene or gene set to account for the group structures in data (e.g., LD structure, co-expression, and genetic regulatory network); (3) remove irrelevant genes and/or features within a gene simultaneously.

Our proposed sparse multitask regression model provided a general framework for integrative analysis of diverse datasets. To fuse multiple diverse datasets, we considered the regression on each single dataset as a single task and then combined all single tasks into the model. Two sets of parameters were used in the model. ω_is were used to weight object functions (i.e., data fitting term at each level) different levels, while δ_j were used for different platforms. Similar to other works, we set ω to be equal by assuming each level of genetic data contains the same information (Ma et al., 2011). We assign δ_j to the data from different platforms by their sample sizes (Wilson and Lipsey, 2001). Other methods can also be applied to estimating weights such as Kaplan–Meier estimate (Liu et al., 2013) and inverse variance (Wilson and Lipsey, 2001). In order to account for the group effects and reduce the large number of features, we used two group sparse penalties in our multitask regression models, i.e., sparse group lasso and sparse group ridge, respectively. These penalties can perform feature selection at both group level and individual for multiple dataset levels, showing better performance than those of using lasso and group lasso penalties for single dataset analysis. Similar regression models were also recently proposed for using two-level sparse group penalties such as group bridge and group MCP (Huang et al., 2012). Ma et al. has recently applied these penalties in regression model for cancer studies to identify those risk oncology genes by integrating multiple expression level datasets from different cancer studies (Liu et al., 2013). Chen et al. has also compared and found that sparse group ridge outperformed group bridge penalty in single dataset regression model (Chen et al., 2010c). However, no study has been performed to compare them for multiple dataset integration and further work is needed in this direction.

WEB SOURCES

The gene expression datasets from three cohorts can be accessed in GEO database (http://www.ncbi.nlm.nih.gov/geo/) with the following accession numbers: 19 Caucasians BMD study (GSE2208), 26 Chinese study (GSE7158), and 80 Caucasians study (GSE56815).

ACKNOWLEDGMENTS

Yu-Ping Wang was partially supported by NIH grants (R01MH104680, R01GM109068). Hong-Wen Deng was partially supported by NIH grants (P50AR055081, R01AG026564, R01AR059781, R01AR050496, and R01AR057049) and Edward G. Schlieder Endowment. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. The FHS datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession phs000007.v14.p6.

FUNDING SOURCE

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute NHLBI in collaboration with Boston University (Contract No. N01-HC-25195). Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. Funding support for the Framingham Bone Mineral Density datasets was provided by NIH grants R01AR/AG 41398.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fcell.2014.00062/abstract

REFERENCES

Balbin, O. A., Prensner, J. R., Sahu, A., Yocum, A., Shankar, S., Malik, R., et al. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat. Commun.* 4, 2617. doi: 10.1038/ncomms3617

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Candes, E., and Romberg, J. (2005). *l1-Magic: Recovery of Sparse Signals via Convex Programming*. Available online at: http://users.ece.gatech.edu/justin/l1magic/downloads/l1magic.pdf

Chen, L., Hutter, C., Potter, J. D., Liu, Y., Prentice, R. L., and Peters, U. L. H. (2010c). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* 86, 860–871. doi: 10.1016/j.ajhg.2010.04.014

Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., et al. (2010a). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* 86, 860–871. doi: 10.1016/j.ajhg.2010.04.014

Chen, X. D., Xiao, P., Lei, S. F., Liu, Y. Z., Guo, Y. F., Deng, F. Y., et al. (2010b). Gene expression profiling in monocytes and SNP association suggest the importance of the STAT1 gene for osteoporosis in both Chinese and Caucasians. *J. Bone Miner. Res.* 25, 339–355. doi: 10.1359/jbmr.090724

Chen, Y., Wu, X., and Jiang, R. (2013). Integrating human omics data to prioritize candidate genes. *BMC Med. Genomics* 6:57. doi: 10.1186/1755-8794-6-57

Deng, F. Y., Lei, S. F., Chen, X. D., Tan, L. J., Zhu, X. Z., and Deng, H. W. (2011). An integrative study ascertained SOD2 as a susceptibility gene for osteoporosis in Chinese. *J. Bone Miner. Res.* 26, 2695–2701. doi: 10.1002/jbmr.471

Deng, W., Yin, W., and Zhang, Y. (2013). "Group sparse optimization by alternating direction method," in *SPIE Optical Engineering+ Applications: 2013: International Society for Optics and Photonics; 88580R-88580R-88515* (San Diego, CA).

Dongdong, L., Hao, H., Jingyao, L., Hong-Wen, D., Calhoun, V. D., and Yu-Ping, W. (2013). "Network-based investigation of genetic modules associated with functional brain networks in schizophrenia," in *2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Beijing), 9–16.

Esser, E. (2009). *Applications of Lagrangian-Based Alternating Direction Methods and Connections to Split Bregman*. Technical Report 09-31. Berkeley, CA: University of California.

Evangelou, E., and Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* 14, 379–389. doi: 10.1038/nrg3472

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv*:1001.0736.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics* 2009:869093. doi: 10.4061/2009/869093

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Stat. Sci.* 27, 481–499. doi: 10.1214/12-STS392

Huang, Y. T., Vanderweele, T. J., and Lin, X. (2014). Joint analysis of Snp and Gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat.* 8, 352–376. doi: 10.1214/13-AOAS690

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* 22, 2890–2897. doi: 10.1093/bioinformatics/btl492

Karasik, D., Cheung, C. L., Zhou, Y., Cupples, L. A., Kiel, D. P., and Demissie, S. (2012). Genome−wide association of an integrated osteoporosis−related phenotype: is there evidence for pleiotropic genes? *J. Bone Miner. Res.* 27, 319–330. doi: 10.1002/jbmr.563

Kawatani, M., Okumura, H., Honda, K., Kanoh, N., Muroi, M., Dohmae, N., et al. (2008). The identification of an osteoclastogenesis inhibitor through the inhibition of glyoxalase I. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11691–11696. doi: 10.1073/pnas.0712239105

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28, 3290–3297. doi: 10.1093/bioinformatics/bts595

Klesney-Tait, J., Turnbull, I. R., and Colonna, M. (2006). The TREM receptor family and signal integration. *Nat. Immunol.* 7, 1266–1273. doi: 10.1038/ni1411

Lei, S. F., Wu, S., Li, L. M., Deng, F. Y., Xiao, S. M., Jiang, C., et al. (2009). An *in vivo* genome wide gene expression study of circulating monocytes suggested GBP1, STAT1 and CXCL10 as novel risk genes for the differentiation of peak bone mass. *Bone* 44, 1010–1014. doi: 10.1016/j.bone.2008.05.016

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H. W., and Wang, Y. P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics* 14:245. doi: 10.1186/1471-2105-14-245

Liu, J., Huang, J., Xie, Y., and Ma, S. (2013). Sparse group penalized integrative analysis of multiple cancer prognosis datasets. *Genet. Res. (Camb).* 95, 68–77. doi: 10.1017/S0016672313000086

Liu, J., Ji, S., and Ye, J. (2009). *SLEP: Sparse learning with efficient projections. Arizona State University 6*. Available online at: http://www.public.asu.edu/~jye02/Software/SLEP

Liu, Y. Z., Dvornyk, V., Lu, Y., Shen, H., Lappe, J. M., Recker, R. R., et al. (2005). A novel pathophysiological mechanism for osteoporosis suggested by an *in vivo* gene expression study of circulating monocytes. *J. Biol. Chem.* 280, 29011–29016. doi: 10.1074/jbc.M501164200

Ma, S., Huang, J., and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* 12, 763–775. doi: 10.1093/biostatistics/kxr004

McGee-Lawrence, M. E., and Westendorf, J. J. (2011). Histone deacetylases in skeletal development and bone mass maintenance. *Gene* 474, 1–11. doi: 10.1016/j.gene.2010.12.003

Otero, K., Shinohara, M., Zhao, H., Cella, M., Gilfillan, S., Colucci, A., et al. (2012). TREM2 and beta-catenin regulate bone homeostasis by controlling the rate of osteoclastogenesis. *J. Immunol.* 188, 2612–2621. doi: 10.4049/jimmunol.1102836

Phan, J. H., Quo, C. F., Cheng, C., and Wang, M. D. (2012). Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. *IEEE Rev. Biomed. Eng.* 5, 74–87. doi: 10.1109/RBME.2012.2212427

Rhodes, D. R., and Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37(Suppl.), S31–S37. doi: 10.1038/ng1570

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Stat.* 22, 231–245. doi: 10.1080/10618600.2012.681250

Song, R., Huang, J., and Ma, S. (2012). Integrative prescreening in analysis of multiple cancer genomic studies. *BMC Bioinformatics* 13:168. doi: 10.1186/1471-2105-13-168

Sprechmann, P., Ramirez, I., Sapiro, G., and Eldar, Y. C. (2011). C-HiLasso: a collaborative hierarchical sparse modeling framework. *IEEE Trans. Signal Process.* 59, 4183–4198. doi: 10.1109/TSP.2011.2157912

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 27, 2304–2305. doi: 10.1093/bioinformatics/btr341

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. U.S.A.* 100, 8348–8353. doi: 10.1073/pnas.0832373100

van den Berg, E., Schmidt, M., Friedlander, M. P., and Murphy, K. (2008). *Group Sparsity via Linear-Time Projection*. Vancouver, BC: Dept Comput Sci, Univ British Columbia.

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Wilson, D. B., and Lipsey, M. (2001). *Practical Meta-Analysis. Оригинал презен тации* Available online at: http://www.mason.gmu.edu.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. (2009). Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.* 57, 2479–2493. doi: 10.1109/TSP.2009.2016892

Xiong, D. H., Liu, X. G., Guo, Y. F., Tan, L. J., Wang, L., Sha, B. Y., et al. (2009). Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups. *Am. J. Hum. Genet.* 84, 388–398. doi: 10.1016/j.ajhg.2009.01.025

Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S., and Furey, T. S. (2012). Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res.* 22, 386–397. doi: 10.1101/gr.124370.111

Yang, J., and Zhang, Y. (2011). Alternating direction algorithms for \ell_1-problems in compressive sensing. *SIAM J. Sci. Comput.* 33, 250–278. doi: 10.1137/090777761

Zhang, D., and Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59, 895–907. doi: 10.1016/j.neuroimage.2011.09.069

Zhang, K., Gray, J. W., and Parvin, B. (2010). Sparse multitask regression for identifying common mechanism of response to therapeutic targets. *Bioinformatics* 26, i97–i105. doi: 10.1093/bioinformatics/btq181

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

Received: 21 June 2014; accepted: 01 October 2014; published online: 27 October 2014.

Citation: Lin D, Zhang J, Li J, He H, Deng H-W and Wang Y-P (2014) Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. Front. Cell Dev. Biol. ***2****:62. doi: 10.3389/fcell.2014.00062*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.

ADVANTAGES OF PUBLISHING IN FRONTIERS

FAST PUBLICATION

Average 90 days
from submission
to publication

COLLABORATIVE PEER-REVIEW

Designed to be rigorous –
yet also collaborative, fair and
constructive

RESEARCH NETWORK

Our network
increases readership
for your article

OPEN ACCESS

Articles are free to read,
for greatest visibility

TRANSPARENT

Editors and reviewers
acknowledged by name
on published articles

GLOBAL SPREAD

Six million monthly
page views worldwide

COPYRIGHT TO AUTHORS

No limit to
article distribution
and re-use

IMPACT METRICS

Advanced metrics
track your
article's impact

SUPPORT

By our Swiss-based
editorial team

frontiers

EPFL Innovation Park · Building I · 1015 Lausanne · Switzerland
T +41 21 510 17 00 · info@frontiersin.org · frontiersin.org