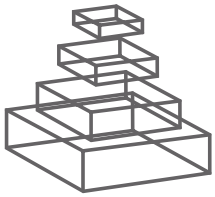
frontiers

RESEARCH TOPICS

COMPREHENSIVE SYSTEMS BIOMEDICINE

Topic Editors
Enrico Capobianco and Pietro Lió

frontiers in GENETICS



frontiers

FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2014
Frontiers Media SA.
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-374-5
DOI 10.3389/978-2-88919-374-5

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

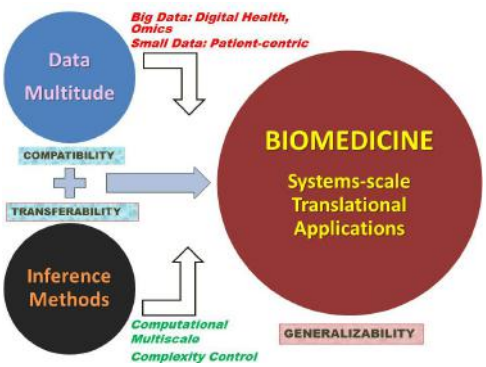
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - 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

COMPREHENSIVE SYSTEMS BIOMEDICINE

Topic Editors:
Enrico Capobianco, University of Miami, USA & LISM – Laboratory of Integrative Systems Medicine, CNR, Pisa, Italy
Pietro Lió, University of Cambridge, Computer Lab, United Kingdom



Systems Biomedicine Axes.
Many data types require ad hoc inference methods to enable a translational systems approach to biomedicine.

Systems Biomedicine is a field in perpetual development. By definition a translational discipline, it emphasizes the role of quantitative systems approaches in biomedicine and aims to offer solutions to many emerging problems characterized by levels and types of complexity and uncertainty unmet before. Many factors, including technological and societal ones, need to be considered.

In particular, new technologies are providing researchers with the data deluge whose management and exploitation requires a reinvention of cross-disciplinary team efforts. The advent of "omics" and high-content imaging are examples of advances de facto establishing the necessity of systems approaches. Hypothesis-driven models and in silico validation tools in support to all the varieties of experimental applications call for a profound revision. The focus on phases like mining and assimilating the data has substantially increased so to allow for interpretable knowledge to be inferred. Notably, to be able to tackle the newly generated data dimensionality, heterogeneity and complexity, model-free and data-driven intensive applications are increasingly shaping the computational pipelines and architectures that quant specialists set aside of the high-throughput genomics, transcriptomics, proteomics platforms.

As for the societal aspects, in many advanced societies health care needs now more than in the past to address the problem of managing ageing populations and their complex morbidity patterns. In parallel, there is a growing research interest on the impact that cross-disciplinary clinical, epidemiological and quantitative modelling studies can have in relation to outcomes potentially affecting the quality of life of many people.

Complex systems, including those characterizing biomedicine, are assessed in both their functionality and stability, and also relatively to the capacity of generating information from diversity, variation, and complexity.

Due to the combined interactions and effects, such systems embed prediction power available for instance in both target identification or marker discovery, or more generally for conducting inference about patients' pathological states, i.e. normal versus disease, diagnostic or prognostic analysis, and preventive assessment (e.g., risk evaluation). The ultimate goal, personalized medicine, will be achieved based on the confluence of the system's predictive power to patient-specific profiling.

Table of Contents

frontiers in GENETICS

EDITORIAL
published: 14 August 2014
doi: 10.3389/fgene.2014.00273

Advances in translational biomedicine from systems approaches

***Enrico Capobianco*[1,2]* and *Pietro Lió*[3]**

[1] *Center for Computational Science, University of Miami, Miami, FL, USA*
[2] *Laboratory of Integrative Systems Medicine, Institute of Clinical Physiology, CNR, Pisa, Italy*
[3] *Computer Laboratory, Cambridge University, Cambridge, UK*
**Correspondence: ecapobianco@med.miami.edu*

Edited and reviewed by:
Raina Robeva, Sweet Briar College, USA

Keywords: systems biomedicine, translational science, big data, inference, paradigm shift

Systems Biomedicine (see for instance Antony et al., 2012) is a field in perpetual development. By definition a translational discipline almost holistically centered on the patient, it emphasizes in light of its multifaceted characterization the need of assessing its constitutive components as a system, whose dynamics occur across multiple and hierarchical scales (organs, tissues, cells, molecules).

A principal role in systems approaches is played by quantitative inference methods resolving problems of high complexity and uncertainty levels. Not surprisingly, it is expected that complex systems may generate information from heterogeneity of sources and diversity of components. Researchers can use this information to look for data patterns with a signal-to-noise ratio which help explaining variation and interdependent phenomena (gene expression and methylation, pervasive transcription, alternative splicing etc.).

Next-Gen technologies and Electronic Medical/Health Records are providing researchers with data resources correctly classified as "Big Data" (Pathak et al., 2013). The management of such resources implies that a cross-disciplinary approach must be put in place, involving team work targeting multiplexed research topics (clinical, experimental, omics, high-content imaging, etc.) whose separate analysis would not be as informative as their synergistic fusion. In parallel, the growing impact of integration of medical records, epidemiological studies and quantitative measures referred to patients is increasingly expanding the frontier of personalized or individualized medicine by leveraging on a multi-evidenced mosaic of information designed to improve patient-specific profiling.

While in principle it clearly appears from the most recent literature what systems biomedicine is aiming to achieve, and the attention is now on what instruments are needed, a main question to pose is: *How fast and effectively are we moving into this translation?* Given the current speed at which the translation is taking place, there are cultural, technical and methodological bottlenecks that need be solved. The proposed Special Topic on "Comprehensive Systems Biomedicine" overviews the path of progression of the field along three main axes:

(1) Data: once the accessibility is guaranteed and the dimensionality is managed, these will require novel generation analytics to discriminate between signal and noise and thus reveal with accuracy the inherent verifiability, relevance, completeness, prediction power making of the data optimal candidate for integrative inference approaches. The non-coding RNA role is being increasingly revealed by high-throughput studies (The ENCODE Project Consortium, 2004; Harrow et al., 2012) in both healthy and diseased conditions, but refers also to the possibility of re-using data from previous technologies, i.e., microarray, as shown by the contributed work on *neuropathic pain*. Then, this role is destined to have a strong impact in *pluripotency and neural differentiation of hESCs and ihPSCs* (following Li et al., 2011). Also, data integration is currently a major topic, in particular with reference to *profiling and pathway annotation of large-scale cancer cell lines*.

(2) Methods: when modularly designed and semi-parametric, methods guarantee wide-spectrum applicability. Hybrid pipelines can take advantage of different quantitative approaches (statistics, machine learning, optimization, control, graph theory) combining multiple platform outcomes, with analyzers and optimizers outflowing into metadata and visual frameworks. *Molecular interaction network approaches in pharmacology* are reviewed in a contributed study, while in another study *magnetic resonance techniques* are discussed with regard to morphological and physiological characterizations of cancer tissue *in vivo*. Finally, a study is presented for *pathway, network, and multiplex methods in the context of brain data*.

(3) Systems: an organized functionally interactive aggregate of entities operating under coordinated and harmonic rules in normal conditions, should be comparatively evaluated against altered (disordered, dysregulated, etc.) conditions to assess phenotypic variations determining the systems characteristics preventively or prospectively, at disease onset and pre/post intervention. In one example, the integration of cytokines, lipoproteins, tissue proteins, and histology indexes cast within a statistical model to study plaque growth opens for new possible interpretations of the *atherogenesis inflammatory disorder*. In another study, the *modeling of metabolism* is considered and an algorithm proposed to detect functional groups from existing databases and to identify metabolites,

and from the perspective of *pandemic studies*. Then, a study introduces an *information system for precision oncology* designed for the integration of data and real-time processing of samples with the computational analysis of genomic alterations and mutations observed in the molecular profiles.

The three axes—Data, Methods, Systems—can be naturally integrated through key properties (such as compatibility, transferability, generalizability), characteristic features and state-of-the-art tendencies.

The communication across the axes is established on the basis of the specific application domains. The final impacts (clinical, societal, etc.) depend on both axis prioritization and solutions that are selected to optimize the key properties.

While much work is on the way for empowering systems approaches to enable a change in biomedical research, we hope that the newly presented studies in this Special Topic can offer opportunities to appreciate the current endeavors and prospective potential in this field.

ACKNOWLEDGMENTS

The authors thank all the people contributing to the Special Topic, and the Frontiers in Genetics Editorial Office for their work, especially Dr. Rossana Mirabella for her precious support.

REFERENCES

Antony, P. M., Balling, R., and Vlassis, N. (2012). From systems biology to systems biomedicine. *Curr. Opin. Biotechnol.* 2012, 604–608. doi: 10.1016/j.copbio.2011.11.009

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

Li, Z., Yang, C. S., Nakashima, K., and Rana, T. M. (2011). Small RNA-mediated regulation of iPS cell generation. *EMBO J.* 30, 823–834. doi: 10.1038/emboj.2011.2

Pathak, J., Kho, A. N., and Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J. Am. Med. Inform. Assoc.* 20, 206–211. doi: 10.1136/amiajnl-2013-002428

The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640. doi: 10.1126/science.1105136

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 July 2014; accepted: 25 July 2014; published online: 14 August 2014.
*Citation: Capobianco E and Lió P (2014) Advances in translational biomedicine from systems approaches. Front. Genet. **5**:273. doi: 10.3389/fgene.2014.00273*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.

Copyright © 2014 Capobianco and Lió. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
GENETICS

REVIEW ARTICLE
published: 13 January 2014
doi: 10.3389/fgene.2013.00298

Could magnetic resonance provide *in vivo* histology?

Marco Dominietto* and Markus Rudin

Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland

Edited by:
Enrico Capobianco, University of Miami, USA

Reviewed by:
Zhiqun Zhang, University of Florida, USA
Enrico Capobianco, University of Miami, USA

****Correspondence:***
Marco Dominietto, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, HIT-Wolfgang-Pauli-Street 27, Zurich 8093, Switzerland
e-mail: dominietto@biomed.ee.ethz.ch

The diagnosis of a suspected tumor lesion faces two basic problems: detection and identification of the specific type of tumor. Radiological techniques are commonly used for the detection and localization of solid tumors. Prerequisite is a high intrinsic or enhanced contrast between normal and neoplastic tissue. Identification of the tumor type is still based on histological analysis. The result depends critically on the sampling sites, which given the inherent heterogeneity of tumors, constitutes a major limitation. Non-invasive *in vivo* imaging might overcome this limitation providing comprehensive three-dimensional morphological, physiological, and metabolic information as well as the possibility for longitudinal studies. In this context, magnetic resonance based techniques are quite attractive since offer at the same time high spatial resolution, unique soft tissue contrast, good temporal resolution to study dynamic processes and high chemical specificity. The goal of this paper is to review the role of magnetic resonance techniques in characterizing tumor tissue *in vivo* both at morphological and physiological levels. The first part of this review covers methods, which provide information on specific aspects of tumor phenotypes, considered as indicators of malignancy. These comprise measurements of the inflammatory status, neo-vascular physiology, acidosis, tumor oxygenation, and metabolism together with tissue morphology. Even if the spatial resolution is not sufficient to characterize the tumor phenotype at a cellular level, this multiparametric information might potentially be used for classification of tumors. The second part discusses mathematical tools, which allow characterizing tissue based on the acquired three-dimensional data set. In particular, methods addressing tumor heterogeneity will be highlighted. Finally, we address the potential and limitation of using MRI as a tool to provide *in vivo* tissue characterization.

Keywords: *in vivo*, histology, MRI, tumor, classification, physiology, metabolism, tissue

INTRODUCTION

Imaging in diagnosis of suspected neoplastic lesion faces two basic problems: detection and identification of a tumor mass. Detection is based on achieving sufficient contrast (i.e., contrast-to-noise ratio) to enable discrimination of pathological from adjacent normal tissue. Critical factors are high SNR (signal-to-noise ratio) and high soft-tissue contrast, i.e., different tissues should be reflected by different intensity levels in the images and with high spatial resolution. Identification is more demanding and today still based in histological analysis, which faces however, some limitations. Histology is typically carried out on biopsy samples, which provide only focal information on a heterogeneous mass. Sample collection constitutes a burden for the patient and may not always be feasible. Furthermore, longitudinal analyses are difficult. On the other hand, histology yields unambiguous information critical for diagnosis that is based on cellular morphology or on the expression of a characteristic molecular signature expressed by the tissue. The possibility to simultaneously analyze multiple tissue parameters is essential for the identification of the tumor type.

Non-invasive imaging for tumor diagnosis offers unique advantages: minimal burden of the patient, full three-dimensional sampling of the heterogeneous lesion, dynamic measurement of physiological and metabolic processes complementing morphological information, and the possibility for longitudinal examinations. Yet, current imaging approaches are based on structural and physiological phenotypic readouts, which are sufficient for lesion detection and monitoring disease progression or therapy response, but most likely, will not allow identifying the lesion type. Analogous to histological tissue characterization it would be important to assess (a) molecular and cellular characteristics and (b) multiple complementary tissue features in order to achieve a high discriminative power.

As we will see later, the use of complementary imaging modalities that probe different aspects of the pathology would be most promising. Nevertheless, we will focus our current discussion on magnetic resonance based techniques, which are attractive as they provide high spatial resolution, unique soft tissue contrast, a temporal resolution sufficient for studying dynamic processes, and moreover are characterized by high chemical specificity, a feature that is extensively used for chemical and biochemical structure elucidation. In addition, the method can be easily translated into the clinics.

TISSUE CHARACTERIZATION BY MAGNETIC RESONANCE

Magnetic resonance images represent a weighted distribution of protons (^{1}H) in tissue, the predominant source of the signal being tissue water and lipids (adipose tissue). Obviously the signal is proportional to the density of protons in the respective tissue.

The weighting function is governed by the proton magnetic properties, which are affected by their local environments due to magnetic and chemical interactions which depend on the nature of tissue (Weishaupt et al., 2006). The effect of the environment on the MRI signal is lumped into parameters describing three distinct relaxation processes (Mark Haacke, 1999): (1) the longitudinal relaxation characterized by the relaxation time T1, which describes the interaction of the spin with its environment, hence the expression spin-lattice relaxation as a crystal lattice constituted the environment in early solid state physics nuclear magnetic resonance (NMR) experiments. T1 relaxation is based on energy exchange between the spin under investigation and its environment and occurs such that the system is driven back to its thermal equilibrium state. (2) Transverse relaxation, characterized by the relaxation time T2 that describes the interaction of the spin under interrogation with neighboring spins, hence the term spin-spin relaxation. T2 relaxation is based on dipole–dipole interactions between spin pairs that fluctuate with regard to their spatial alignment and hence is of stochastic nature. It leads to the irreversible loss of phase coherence and hence to a loss in signal intensity. (3) T2* relaxation is related to T2 and in addition to spin–spin interactions is governed by inhomogeneities in the local magnetic field, e.g., due to difference in magnetic susceptibility between tissues. This local field inhomogeneities are static and hence deterministic and can be accounted for when tailoring the MRI data acquisition (so-called spin-echo experiments). Nevertheless, T2* provides an additional source for contrast. Additional parameters that influence the modulate the interaction of the MRI signal with the environment and hence the MRI signal intensity are molecular diffusion, as well as mechanism leading to coherence/polarization transfer such as chemical exchange reactions or spin diffusion.

Relaxation processes can be influenced by administration of contrast agent, which are either paramagnetic (gadolinium based) or superparamagnetic agents (iron-oxide based). These agents contained unpaired electrons with a strong effect on the local magnetic field that is experience be nearby protons. The contrast mechanism of the two classes of agents is different, yet a detailed description is beyond the scope of this article (Rudin, 2005a). In the context of our discussion it suffices to state that paramagnetic agents enhance the longitudinal relaxation rate, i.e., they reduce T1, while superparamagnetic agents predominantly enhance the transverse relaxation rate, i.e., reduce T2. Apart from enhancing the contrast in static MR images to improve discrimination of distinct tissues, MRI allows monitoring dynamic changes following the contrast agent administration. The contrast change measured in a volume element (voxel) is proportional to the amount of contrast agent in this voxel, which by itself depends on the biodistribution (including compartments within a tissue) and pharmacokinetic properties of the agent. Such dynamic studies yield information on tissue perfusion, vascular leakage, or distribution volumes.

The magnetic resonance phenomena are not only restricted to the detection of protons of water and lipid molecules in tissue. Essentially all magnetic nuclei give rise to signal. The resonance frequency of a nucleus depends on its identity (characterized by the so-called gyromagnetic ratio) and its chemical environment. It is in particular the fact that the magnetic resonance sensitively probes the chemical structure to which the interrogated nucleus is attached that has made the method indispensable for chemical structure elucidation. The identification of a molecular entity is based on the detailed spectral analysis of its resonance frequencies. Translating these approaches to *in vivo* tissue characterization therefore bears considerable potential to enable a detailed (molecular) tissue characterization, which might be of high diagnostic value. Apart from protons, other nuclei such as phosphorus-31, carbon-13, constituents of many biologically relevant molecules are of interest for *in vivo* magnetic resonance spectroscopy (MRS). Yet this method suffers from the low intrinsic sensitivity of magnetic resonance, as these metabolites are typically present at millimolar to sub-millimolar concentration compared to water protons with tissue levels of approximately 80 M.

PHENOTYPIC TUMOR CHARACTERIZATION

If compared to healthy organs, tumor tissues present in general highly heterogeneous and chaotic architecture. Such heterogeneity is primarily due to the genetic instability of tumor cells that is responsible of the apparently chaotic tumor development, which is reflected in tissue architecture, tumor vasculature, host infiltrates, and metastasis formation (Heppner, 1984; Marusyk et al., 2012). This chaotic behavior occurs at a molecular, cellular, and microdomain level and determines also the interaction with the host environment. The result is the formation of different regions inside the tumor, which may exhibit completely different physiological behavior (Denysenko et al., 2010; Huse et al., 2013).

In order to rationalize the complexities of neoplastic disease, Hanahan and Weinberg (2000) have defined six phenotypic hallmarks of cancer, which correspond to six biological features acquired during tumor development. Those include sustained proliferative signaling, evasion of effects of growth suppressor, resistance to cell death program, acquisition of replicative immortality, development of a vascular network (angiogenesis), invasion of adjacent healthy tissue, and the formation of distant metastases. In a recent publication (Hanahan and Weinberg, 2011), these initial six hallmarks were complemented by four additional features related to the specific behavior of tumor tissue: genome instability, inflammation, reprogramming of energy metabolism, and evasion of immune surveillance.

An important aspect of tumor is that they are not only composed of cancer cells but contain a variety of host derived cells such as immune cells, endothelial cells, pericytes, fibroblasts, stem, and progenitor cells that characterize the hallmarks traits and constitute the tumor microenvironment (Swartz et al., 2012).

Considerable efforts have been invested to assess these tumor hallmarks non-invasively using imaging. Today, methods are available to study tumor proliferation (DNA, protein, and membrane synthesis) using PET and MRI methods, aspects of tumor metabolism using PET and MRS, aspects of tumor vessel architecture and physiology (MRI), apoptotic processes using PET, MRI, and fluorescence imaging, as well as of the invasive potential and propensity for metastasis formation using PET and fluorescence imaging. Yet, all these phenotypic readouts are not specific enough for an unambiguous identification of the tumor type, which is based on unique molecular markers. Secondly, many of these tools

are still in an early experimental stage and will not be available in a clinical setting soon.

TUMOR MORPHOLOGY

Damadian (1971) reported on the observation that T1 relaxation times in tumors are higher than in the adjacent normal tissue and suggested that this feature might be used for tumor detection. This constituted one of the prime motivations that later led to the development of MRI. Nowadays, modern MRI scanners offer several tools for detecting and characterize tumor.

Detection of tumors based on altered relaxivity values

Despite the fact that the basic biophysical mechanism leading to tissue specific relaxivity values are poorly understood, the evaluation of relaxivity parameters are of high diagnostic value.

According to the type of MR sequence and the relative parameters, it is possible to acquire a signal, which is mostly dominated by one of these contributions. Most established are T1-weighted, T2-weighted or proton density weighted images (Haacke et al., 1999). By optimizing the contrast between neoplastic and normal tissue it is in generally possible to detect the cancer lesion, to identify sub-regions displaying different tissue characteristics (dense versus non-dense tissue, poorly versus highly vascularized, necrotic areas, edematous tissue, etc.), and to monitor of tumor progression or regression. Yet, these phenotypic measurements are in general not sufficient for "histological" classification of the tumor. Instead some generic tissue features are reflected. For example, T1-weighted images are usually used to assess the gross morphology of the tumor as shown in **Figure 1** (left). As rule of thumb, regions with high water content appear dark, while regions with high fat content appear bright (Weishaupt et al., 2006). In combination with gadolinium-based contrast agent such as Gd-DTPA it is possible to assess regions displaying high uptake of the agent indicative of hemorrhage and leaky vessels. Areas, for which little uptake is observed are commonly associated with necrotic or edematous domains. Only when waiting sufficiently long these areas will accumulate extravasated contrast agent via passive diffusion.

In T2-weighted images areas with high water content appears bright. Since most diseases are characterized by increased water content in tissues associated with an inflammatory tissue response, T2-weighted are particularly useful for pathological investigation. Dark regions may indicate high blood content such as hemorrhage, vessels, or angiomas.

In proton weighted images (Westbrook, 2010), bright areas indicates high proton density tissue, such as cerebrospinal fluid or edema, while dark areas indicate low proton density such connective tissue (i.e., tendons) or cortical bone.

Nowadays, tumor detection based on altered T1 and T2 relaxivity values is commonly used to diagnose and follow-up different kinds of tumor comprising, among the others, brain tumor (Young, 2007), breast tumor (Heywang-Kobrunner et al., 1997), prostate cancer (Verma et al., 2012), and gastric cancer (Wang et al., 2000). By means of T1 and T2 weighted images and in combination with contrast agent, as Gd-DTPA or superparamagnetic nanoparticles, it is possible to assess tumor morphology and grossly identify edematous and necrotic regions. Moreover, kinetics and extent of contrast agent uptake are considered as an indicator of prognostic quality.

The possibility to obtain high-resolution and high-contrast images of soft tissue with similar density but different relaxivity values makes MRI the method of choice for the detection of solid tumors.

Alteration in cellularity: measuring the apparent diffusion coefficient

Diffusion Weighted Imaging (DWI) measures the random movement of the water molecules and allows deriving the so-called apparent diffusion coefficient (ADC) for each voxel (Haacke et al., 1999). "Apparent" since the measured coefficient corresponds to a weighted average across individual diffusion coefficients for all compartments contained in this voxel. Also, structural barriers

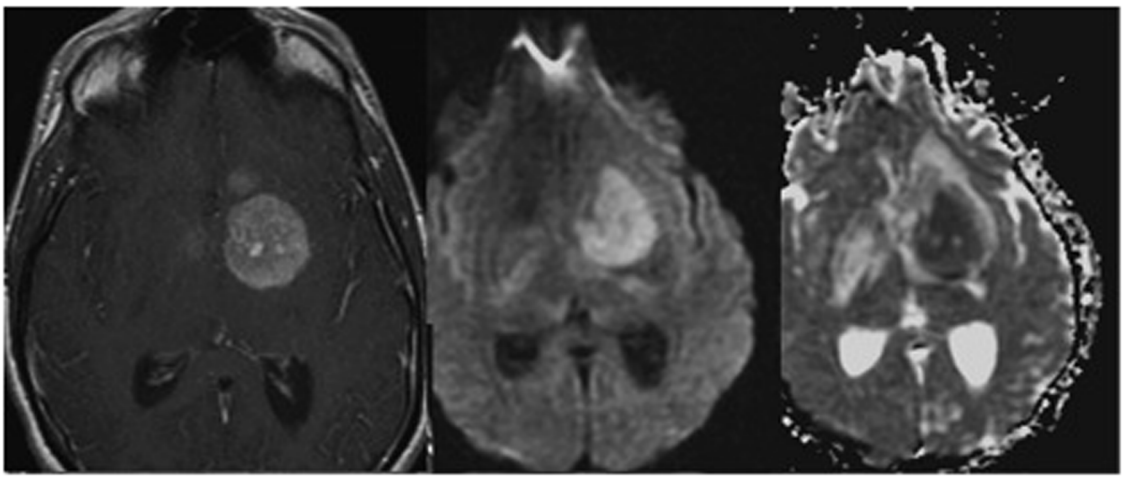


FIGURE 1 | T1-weighted image of a glioma following contrast enhancement using a gadolinium-based contrast agent (left). Diffusion weighted images DWI (**middle**), and apparent diffusion coefficient map ADC (**right**) of the same tumor patient. Adapted from Young (2007), reproduced with permission.

like cell membranes, or perfusion effects affect diffusion (Haacke et al., 1999). Given this definition and the fact that the diffusion coefficients within cells and in the extracellular space are different, with $D_{intracellular} < D_{extracellular}$, it becomes apparent that the ADC values are sensitive to the relative size of these two compartments. Hence, regions with densely packed cells will show low ADC values. This has been exploited in the characterization of brain neoplasms. High grade tumor neoplasms display significant reduction of ADC and correspondingly a higher signal in DWI as compared to lower grade (Okamoto et al., 2000; **Figure 1** middle and right). Fluid filled cysts or edematous regions appear hyperintense in ADC maps (and hypo-intense in DWI) when compared to the normal parenchyma because they largely correspond to bulk water enabling unrestricted diffusion (within the MRI timescale; Drevelegas and Papanikolaou, 2011).

Inflammatory status: edema formation and infiltration of immune cells

Recent data have expanded the concept that inflammation is a critical component of tumor progression (Coussens and Werb, 2002). The quantification of the inflammatory status is crucial in the determination of the tumor volume, since its value is an important prognostic factor with regard to the treatment of malignant tumors (Xie et al., 2005). Moreover, inflammation may also influence therapy outcome in two opposite ways, in particular for brain tumors such as gliomas (Kleijn et al., 2011). It can lead to tumor control, by killing cancer cells and establishing anti-cancer immunity, or it may further promote tumor growth, by participating in glioma reoccurrence and progression. It is therefore evident that the possibility to monitor the inflammation status *in vivo*, i.e., by monitoring immune cells, is a crucial step in tumor management. Traditionally, such evaluation is performed *ex vivo* using cytometry and immunohistochemistry methods, or *in vivo* using labeled-radionuclides for PET (Positron Emission Tomography) or SPET (Single Photon emission tomography) scanner (Ahrens and Bulte, 2013). However, recent developments, in particular the possibility to prepare non-toxic MRI probes for cell labeling, enables MRI based tracking of immune cells. Compared to PET or SPET, MRI has the advantages that it does not use ionizing radiation and provides higher spatial resolution.

Magnetic resonance imaging (MRI) cell tracking involves exogenous cell labels such as iron oxide nanoparticles, perfluorocarbon (PFC) nanoemulsion, or genetically encoded MRI reporters (Ahrens and Bulte, 2013; **Figure 2**). Immune cells can be labeled with superparamagnetic iron oxide based (SPIO) nanoparticles in two ways: (i) by *ex vivo* labeling of harvested cells that are incubated with SPIO nanoparticles in media typically using a transfection agent, or (ii) by non-selective *in situ* labeling of the phagocytic cells, such as macrophages, following intravenous injection of SPIO nanoparticles (Bhakoo et al., 2006). PFC emulsion can be used to track cells using the same labeling strategies. PFC-based cell tracking provides high specificity for cell detection (i.e., a high signal-to-background ratio can be achieved as there is no endogenous source of a fluorine signal) and enables the quantitative measurements of the amount of cells. Yet they require a specific MRI coil tuned to the resonance frequency of ^{19}F nuclei. Disadvantages of using passive labeling strategies are that only the presence of the label is detected, which is not necessarily identical with the presence of cells. Cells may release the label into the environment, e.g., after death, yielding to a false positive signal. Also, the presence of the label does not yield any information on the status of the cell, i.e., whether it is alive or dead. Finally, for dividing cells (which is not relevant for the immune cells) the label will be subsequently diluted. In addition, a passive label will be degraded over time. Genetic encoded reporters avoid some of these issues. They only yield a signal when the gene is expressed, i.e., when the cell is alive, and the presence of labels also indicates the presence of the cell. On the other hand, the sensitivity of genetic cell marking is in general inferior to that of potent exogenous labels.

Magnetic resonance imaging cell tracking can also be used to monitor inflammation related to other disease as neurological disorders, autoimmune diseases, or transplant rejection. Moreover, it is likely to become an important tool also in cell therapy (i.e., stem cells for different diseases) with the specific aim to guide cell injections and subsequently monitoring their migration (Bulte, 2009; Hong et al., 2010).

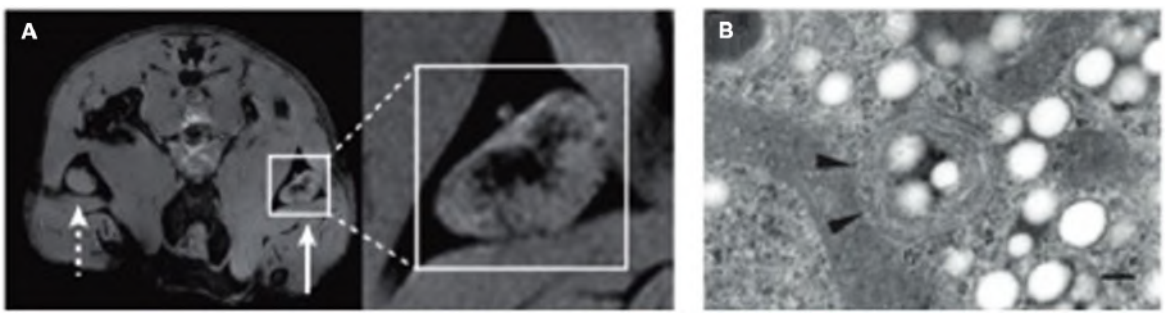


FIGURE 2 | Example of tracking immune cells with MRI using SPIO nanoparticles and PFC emulsions. (A) Imaging of *in vivo* antigen capture and trafficking of dendritic cells (DCs). Sentinel DCs were labeled *in situ* by intradermal injection of unlabeled (**dashed arrow**) or SPIO-labeled (**solid arrow**) irradiated cancer cells, which function as a vaccine. Following phagocytosis of both SPIO particles and tumour antigens in a process known as magnetovaccination, the hypointense DCs migrate into the medulla of the draining popliteal lymph node. **(B)** An electron micrograph of a perfluorocarbon (PFC)-labeled DC is shown. Numerous bright spots (PFC droplets) are observed inside the cell. Particles appear as smooth spheroids (Ogawa et al., 1990). Arrowheads indicate vesicles. The scale bar represents 200 nm. Adapted from Ahrens and Bulte (2013), reproduced with permission.

One of the consequences of the inflammatory status is the formation of a peritumoral edema which is the results of several cellular mechanism (Stummer, 2007). Although its prognostic value for diagnosis, as well its role in the course of disease is still a matter of discussion, peritumoral edema may cause severe neurological symptoms in case of brain tumor, and remains a challenge in the treatment of glioblastoma patients (Kleijn et al., 2011; Stummer, 2007).

The evaluation of edema by means of MRI is usually performed using T2-weighted sequences that are quite sensitive to water content, and by assessing changes in ADC. The regions affected by edema are characterized by prolonged T2 values and therefore appear hyperintense in T2-weighted images.

TUMOR PHYSIOLOGY

The physiology of tumor tissues is directly dependent on the structure and functionality of the vascular network developed during tumor growth. The newly formed vessels are responsible for the delivery of the nutrients from the hosting tissue to the tumor and for the removing of waste metabolites from the tumor. Characterization of the angiogenic process is therefore essential either for understanding the chaotic steps of tumor evolution or for the development of anti-angiogenic drugs (Marmé and Fusenig, 2007).

Tumor vasculature deviates profoundly from that of the normal organs both in vascular architecture and functionality. The vascular network of solid tumor does not show the hierarchical branching patterns characteristic for the majority of healthy organs. This is the results of the opportunistic nature of the angiogenic process, which in tumor seems not to follow physiological pre-determined steps (Tropres et al., 2001; Kiselev et al., 2005). Initially avascular tumor masses trigger the development of new angiogenic vessels as a consequence of hypoxia and the secretion of angiogenic factors (Lemasson et al., 2013). Alternatively, tumors may grow along one or more existing vessels and co-opt them in the tumor structure in a parasitic manner. In both cases vessels usually remain in a primitive status with immature vascular walls and proper support by the tissue matrix. Tumor vascular networks therefore consist of tortuous microvessels exerting chaotic branching, arterial-venous shunts, and are subject to acute or transient collapse (Heywang-Kobrunner et al., 1997).

The lack of maturation of the primitive vessel network gives origin to a few abnormalities in vascular function. Tumor capillaries show high permeability compared to the healthy ones (Tropres et al., 2004). This results in a profound extravasation of erythrocytes and plasma in the adjacent tissue leading to an elevated interstitial fluid pressure and to a rise in the viscous resistance to blood flow (Dominietto, 2012). Second, because of this resistance and chaotic structure, the blood circulation or perfusion within such vessels is rarely correlated to the metabolic demands of solid tumor (Heywang-Kobrunner et al., 1997). Moreover, the clearance of metabolites from the tissue and the drainage by the venous system do not work properly and are responsible of the accumulation of blood in the tumor tissue.

To complicate matters even more, the degree of abnormalities changes in different kinds of tumors and also during different stages of the same tumor. While from a biological point of view the origin of these physiological fluctuations is poorly understood, the assessment of vascular abnormalities constitute an attractive biomarker, as it clearly distinguishes neoplastic from normal tissue. Various structural and physiological aspects of tumor vasculature can be quantified by MRI and used for classification and staging of tumors.

NEOANGIOGENESIS: VASCULAR STATUS AND PHYSIOLOGY

The vascular network of bigger vessels (diameter $> 50\ \mu m$) can be directly visualized by means of magnetic resonance angiography (MRA) technique as shown in **Figure 3**. Three different methods are currently available: (a) time-of-flight (TOF), (b) contrast enhanced (CE), and (c) phase contrast (PC) MRA. All these approaches aim at generating a high contrast between the vascular lumen (blood compartment) and the surrounding tissue to enable the segmentation and extraction of vascular structures.

Time-of-flight angiography (Heverhagen et al., 2008) exploits the intrinsic differential behavior of protons in flowing blood

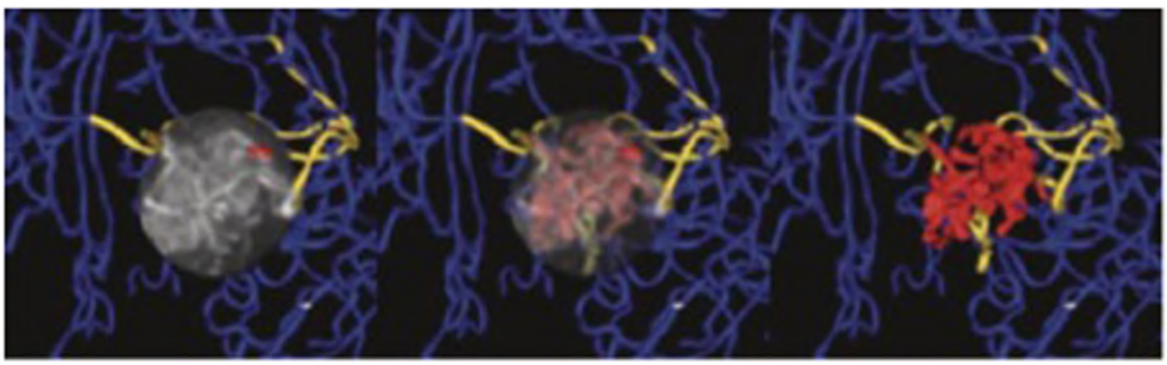


FIGURE 3 | Magnetic resonance angiography of a brain tumor to evaluate the tortuosity of the vascular network. Vessels within the tumor nidus are shown in red, vessels supplying or passing through the nidus in gold, while normal vessels outside the nidus are blue. The nidus, containing type II tortuosity vessels, is volume rendered at full opacity (**left**), at partial opacity (**center**), while vascular structures exclusively are shown at (**right**). Adapted from Bullitt and Gerig (2003), reproduced with permission.

as compared to stationary tissue and does not require the administration of contrast agents. Briefly, by a combination of radiofrequency excitation pulses all the spins of the excited volume will be saturated and, because of that, the signal will be largely suppressed. However, blood that has entered the imaged volume, will give rise to the full signal intensity, as it has not experienced previous saturation. Whether a vessel can be depicted using TOF-MRA depends on whether it can be reached by fresh blood during excitation.

Contrast enhanced (Chandra et al., 2012) takes the advantage of the administration as a bolus of a contrast agent in the blood stream during MRI acquisition. Gadolinium based contrast agent will produce an enhancement of the signal in T1-weighted sequences, while iron-based contrast agent will cause dephasing of the nuclear magnets decreasing the overall signal in T2-weighted acquisitions. Acquisition has to be fast enough that extravasation of the contrast agent remains minimal. Angiograms are then obtained by comparing pre- and post-contrast images.

Phase contrast (Thomas and Wells, 2011) utilizes the change in the phase shifts of the flowing protons in the region of interest to create an image. Spins moving along the direction of a magnetic field gradient receive a phase shift proportional to their velocity. This is usually accomplished by applying gradient pairs, which sequentially dephase and then rephase spins during the sequence. Use of phase-sensitive image reconstruction allows depticting the vascular systems exclusively and more over provides information on blood flow velocities.

Despite the high spatial resolution of MRI if compared to other diagnostic imaging modalities, it is not possible to depict the fine details vascular tree as (a) the typical vessel diameter of tumor vessels is in the range 5–50 μm, and (b) flow velocity in these vessels is typical small. Only with high-field magnets and sophisticate coils that are used in experimental studies in animals, enabling an isotropic spatial resolution of the order of 50 μm, it has been possible to depict larger branches of the tumor vasculature (>50 μm) using CE techniques in subcutaneous or orthotopic tumors in mice. Nevertheless, MRI offers the ability to indirectly investigate small vessels by means of a special CE technique called vessel size imaging (VSI).

Vessel size imaging (Tropres et al., 2001; Kiselev et al., 2005) allows the evaluation of the mean vascular density (MVD; Lemasson et al., 2013) and the average vessel diameter (AVD) in a voxel or in a volume (Tropres et al., 2004). The approach is based on the simultaneous measurement of the changes in T2 and T2* induced by the administration of an intravascular superparamagnetic contrast agent. While T2 depends on the dipolar interaction between the intravascular contrast agent and the tissue protons, which scales to the surface of the vessel T2* effects ar proportional to the bulk effect of the contrast agent to the local magnetic susceptibility, which scales to the vascular volume. From indirect measurements of vessel surface and volume we can infer on the average radius of the vessels in a given region-of-interest.

The dimension and density of the vessels is an important index when studying angiogenesis. When combined with an independent measurement of the tumor blood volume (TBV), it constitutes an index of the organization of the vascular network. Identification of vessels of various diameter (from big to small) indicates a hierarchical network, while the presence of only small vessels is an index of the poor organization of the vascular tree.

While information on the vascular architecture within the tumor is a downstream manifestation of the angiogenic process, it is important to derive physiological information in order to understand the implication on substrate delivery, which essentially determines the fate of the tumor. Capillary vessels like arterioles and venules are permeable to the substances present in the blood to enhancing compound exchange between the blood and tissue compartment. It has been shown that in tumors also relatively big vessels are highly permeable due to the immature structure of the vascular wall. This results on an almost completely leaky network with a highly non-uniform blood supply to tumor tissue (Dominietto, 2012).

The characteristically high permeability of tumor vessels has been suggested as biomarker for angiogenesis (Feng et al., 2008), and for evaluating antiangiogenic treatment efficacy (Alic et al., 2011; O'Connor et al., 2011; Najafi et al., 2012). Vascular permeability values are commonly assessed by means of T1-weighted dynamic contrast enhanced (DCE) acquisitions, involving serial images of the same region during the administration of a gadolinium-based contrast agent (Rudin et al., 2005). The measured MRI signal enhancement curve is fitted using a two-compartment model originally proposed by Tofts and Kermode (1991). In its simplest version the model comprises a vascular and an extracellular compartment. Fitting to the enhancement curve is carried out by optimizing two parameters, the vascular permeability defined by the transfer constant k^{trans}, a measure for the rate of contrast agent extravasation, and the volume of the extracellular compartment V_e.

Two other important parameters giving insight into the vessel functionality are tumor blood flow (TBF) and TBV (**Figure 4**). While TBV measure the volume of the vascular compartment in a region-of-interest, TBF assess the exchange of blood within this volume per unit time. Both parameters can be estimated by means of T2*-weighted dynamic susceptibility contrast (DSC) MRI experiments recording the change in signal intensity during the administration of a super-paramagnetic contrast agent (Barbier et al., 2001; Rudin et al., 2005). For data analysis, it is assumed that, due to its nanoparticulate size, the contrast agent remains confined to the blood compartment, at least for the duration of the measurement.

Tumor oxygenation

The oxygenation is another important factor in tissue characterization since abnormal oxygen levels have several implications in tumor progression and treatment (Nilesh and Quarles, 2011). In particular, a hypoxic environment is known to promote angiogenesis, inflammatory behavior, genetic instability, invasiveness, and metastasis formation. Hence, hypoxia is associated with increased malignancy and causes reduced efficacy of radio- and chemo-therapy.

Two MR based techniques have mainly developed to image tissue oxygenation status: BOLD-MRI and fluorine-19 NMR (^{19}F-NMR). BOLD (Blood Oxygen Level Dependent; **Figure 5**) contrast assesses alterations in the relative concentrations of deoxyhemoglobin (dHb) and oxyemoglobin (HbO_2) concentration in

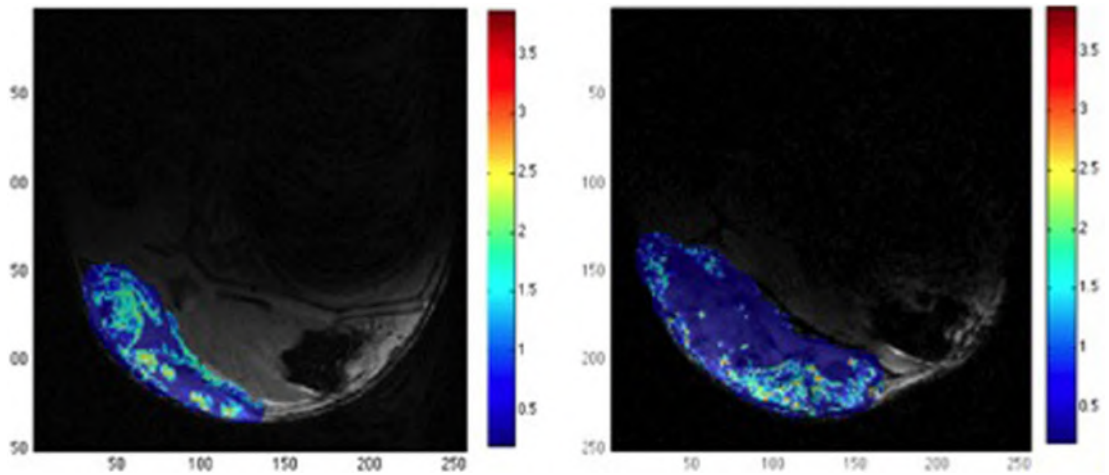


FIGURE 4 | Example of relative tumor blood volume rTBV (color) overlaid on a structural MR image (gray level). The images show the effect of DMOG treatment that affects angiogenesis process (**left**) versus placebo (**right**). DMOG treated tumor shows multiple small regions with relative high rTBV, while placebo treated tumor present only one big region with significant rTBV. The color bar indicates the rTBV values in arbitrary units. Adapted from Dominietto et al. (2012), reproduced with permission.

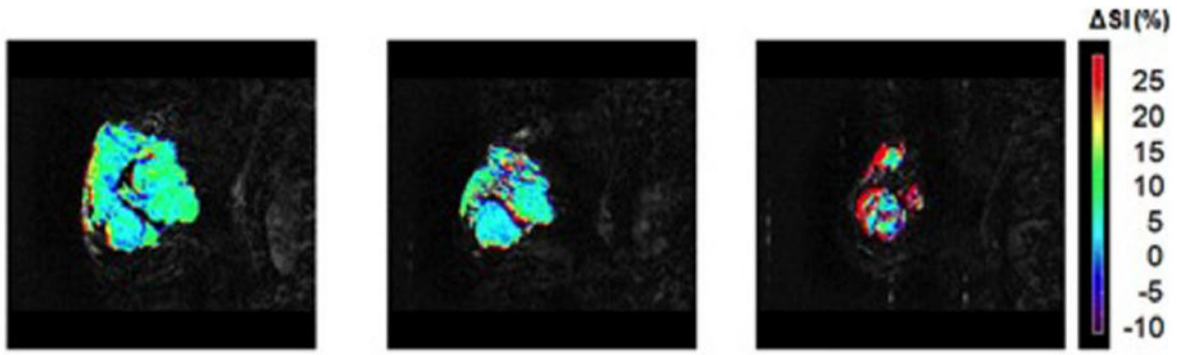


FIGURE 5 | BOLD MRI for a patient with breast tumor exhibiting a partial response to therapy. Images show a signal enhancement maps (**color**) overlaid on T2-weighted anatomical images. Images have been acquired 1 week before start of neoadjuvant chemotherapy (**left**), after one cycle of chemotherapy showing small signal response (**middle**) and after four cycles of chemotherapy demonstrating a striking change in tumor characteristics in response to therapy (**right**). Adapted from Jiang and Weatherall (2013), reproduced with permission.

blood (Ogawa et al., 1990). The blood oxygen saturation given by the ratio $[HbO_2]/([HbO_2] + [dHb])$ changes according to local cellular activity and hence oxygen consumption. Since dHb is paramagnetic, it induces local changes in magnetic susceptibility, and hence a decrease of T2*, in the region surrounding the vessel. Correspondingly, increased oxygen saturation will lead to an increased signal intensity when using T2*-weighted pulse sequences (Nilesh and Quarles, 2011). This method has been used to monitor treatment response during phototherapy (Gross et al., 2003), upon administration of vasomodulators (Robinson et al., 1995; Taylor et al., 2001), to predict the response radiotherapy response, which is known to critically depend on the oxygenation status of the tumor (Rodrigues et al., 2004), and to characterize vascular architecture in general (Robinson et al., 2003). While BOLD based methods provide accurate qualitative information of blood oxygenation it is difficult to extract reliable quantitative data.

^{19}F-NMR approaches involve the administration of PFCs, which are well known for their high oxygen carrying capacity. It has been demonstrated that the ^{19}F relaxation time T1 is linearly dependent on oxygen tension (Joseph et al., 1985; Fishman et al., 1989) and with proper calibration it is possible to quantitatively assess tissue oxygenation at equilibrium, or following a metabolic perturbation. However, given the difficulty of delivering sufficient quantities of PFCs to tumor tissue, as many of these agents require intra-tumoral injection, the method has remained a preclinical tool (Nilesh and Quarles, 2011).

Acidosis: link to metabolism

Metabolic reprogramming of tumor cells has been recognized already very early. It has been observed that neoplastic tissue exerts high glycolytic activity even under conditions of normoxia (Warburg effect; Gatenby and Gillies, 2004). In fact, measurement of enhanced glucose utilization with PET using

[^{18}F]-2-fluoro-2-doxyglucose (FDG) as tracer has emerged as important diagnostic tool for tumor diagnosis, in particular for detection of the metastatic burden. Only recently, molecular mechanism underlying this reprogramming, linking metabolic processes to altered gene expression are being elucidated (DeBerardinis et al., 2008; Ward and Thompson, 2012). Glycolysis leads to the production of lactic acid from pyruvic acid via pyruvate dehydrogenase, which is responsible for acidosis. Nevertheless, the intracellular pH of solid tumor, which is the result of a balance between metabolic proton production, proton buffering capacity and transport processes, is maintained within a range of pH = 7.0–7.2 (Zhang et al., 2010). Hence, despite increased acid production, tumor cells maintain a normal slightly alkaline intracellular pH. The major acid load is transported outside the cells but, since the acid cannot be easily removed by the abnormal vasculature, the microenvironment will become acidic (Zhang et al., 2010).

Tissue acidosis is an important feature of the tumor microenvironment which has been shown to drive local invasion and not surprisingly several approaches have been described to assess tumor pH non-invasively (**Figure 6**). *In vivo* MRI and MRS can be used to measure pH values *in vivo* either using endogenous or exogenous compounds (Raghunand, 2006). MRS methods are generally based on a difference in chemical shifts between pH-dependent and pH-independent resonances (Zhang et al., 2010). A resonance becomes pH dependent when the resonance frequency of the protonated form is distinct from that of the deprotonated form and when the exchange reaction is fast compared to the MRS time scale, which is defined by the frequency difference of the two resonances. Different nuclei can be used to determine tissue pH using this approach: ^{31}P (Gadian and Radda, 1981), ^{1}H and hyperpolarized ^{13}C (Gallagher et al., 2011).

An alternative approach using MRI relies on perturbing the relaxivity of water via pH-dependent relaxation agents. Small molecules Gd-based agents, whose relaxivity is pH dependent, have been recently synthesized (Zhang et al., 1999; Raghunand et al., 2002; Pierre et al., 2006). For the pH quantification, this method requires knowledge of the concentration of the agent in each voxel.

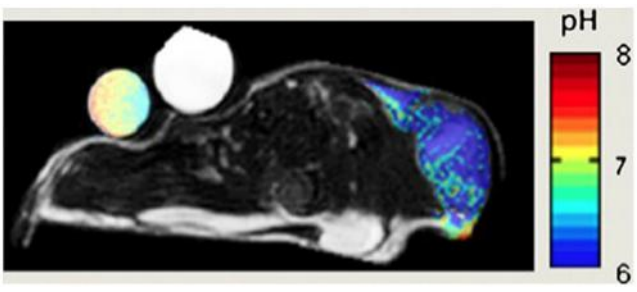


FIGURE 6 | pH map of mouse MCF-7 breast tumor model. pH was measured by administration of a paramagnetic CEST (Chemical Exchange Saturation Transfer) MRI using pH-sensitive contrast agent ytterbium-1,4,7,10-tetraazacyclododecane-1,4,7 tetraacetic acid, 10-oaminoanilide. Adapted from Zhang et al. (2010) reproduced with permission.

Finally, a new generation of agents that have been developed to generate contrast via chemical shift saturation transfer (CEST) enable pH measurement (Zhang et al., 2010). The dynamic process of CEST can be described by 2-pool chemical exchange model, wherein the magnetization is exchanged between a labile proton (e.g., an amide proton of proteins) and bulk water. The two resonances have to be distinguishable. In the experiment one of the two resonances (the smaller proton pool) is magnetically labeled (saturated) and the transfer of label to the exchange partner (the water proton) is monitored. For example, the resonance of amide protons is saturated and the transfer of saturation to the water resonance, i.e., the decrease of the water signal intensity, is analyzed. Mathematical modeling based on Bloch equations coupled by chemical exchange yields estimates for the exchange rate, which depend on pH. In general, exchange rates are slower at a low pH. There are three main categories of CEST imaging: diamagnetic (Pacheco-Torres et al., 2011), paramagnetic (Liu et al., 2012), and amide proton transfer (Sun et al., 2011).

TUMOR METABOLISM

The concentration various metabolites can be measured by means of MRS (**Figure 7**). Compounds accessible by MRS relate to the tumor hallmarks deregulated energy metabolism, sustained proliferation, and resisting cell death (Hanahan, 2000). Metabolites related to energy metabolism are the substrate glucose and the intermediates of glycolytic processing including pyruvate and lactate, which can be assessed using either ^{1}H or ^{13}C MRS. Recently, hyperpolarization techniques such as ^{13}C MRS combined with dynamic nuclear polarization (DNP) have been introduce. They enhance the sensitivity of MRI by three to four orders of magnitude, though the lifetime of the hyperpolarized state is typically less than 1 min in biological tissue, which limits the applicability of the method. Nevertheless, it could be shown using DNP ^{13}C MRS in addition to glycolytic processing of pyruvate that the label is also transferred to alanine, which indicates the increased anabolic (proliferative) activity of tumors. The prime energy substrate produced by anaerobic and aerobic glucose processing is adenosine-triphosphate (ATP), which can be assessed, together with other phosphorus containing metabolites such as phosphocreatine (PCr), nicotinamide adenine dinucleotide phosphate (NADP), or orthophosphate (HPO_4^{2-}/ $H_2PO_4^-$) using ^{31}P MRS. A characteristic of tumors is their acidic environment, which is related to their high glycolytic activity. Intracellular pH is commonly assessed by comparing the resonance frequency of the PCr and HPO_4^{2-}/ $H_2PO_4^-$ resonance. Due to the fast proton exchange (with regard to the MRS time scale) between HPO_4^{2-} and $H_2PO_4^-$ only one resonance signal is observed for the two compounds, the frequency of which depends on the relative concentration of the two and hence sensitive to the pH value. In contrast, the PCR signal does not depend on the pH value. Hence by measuring the frequency difference of the PCr versus the HPO_4^{2-}/ $H_2PO_4^-$ signal, the pH value can be accurately determined (Zhang et al., 2010). High proliferation capacity implies high rates of membrane synthesis. Not surprisingly tumor typically show high levels of phospholipid precursors such as choline/phosphocholine or ethanolamine/phospho-ethanolamine. While the non-phosphorylated compound are typically measured using ^{1}H MRS,

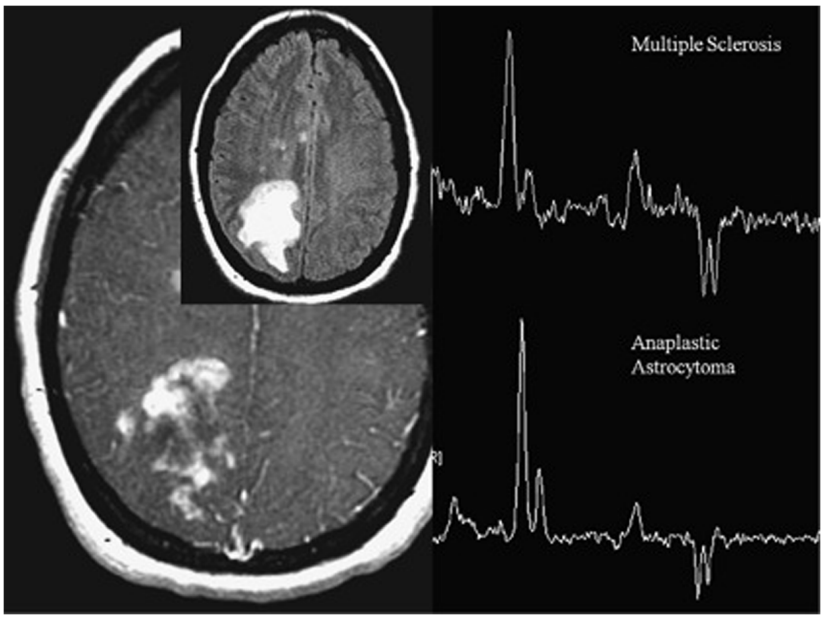


FIGURE 7 | Magnetic resonance spectroscopy from patient with heterogeneously enhancing white matter lesions. The indistinguishable spectra demonstrate elevated choline, low NAA, and moderate lactate. One spectrum represents tumefactive multiple sclerosis (MS), the other one anaplastic astrocytoma. In anaplastic astrocytoma, choline elevation reflects membrane synthesis as marker of active proliferation, whereas in MS, it represents membrane injury and degradation of membrane phospholipids. Adapted from Young (2007) reproduced with permission.

the phosphorylated analogs are detected as a phosphomonoester resonance using ^{31}P MRS. In fact the characteristic nature of this peak has been used to assess therapy response already very early (Ng et al., 1989). In clinical routine, these proliferation readouts are mainly used in the diagnosis and monitoring of brain tumors (Bhakoo et al., 2006; Ahrens and Bulte, 2013). Finally it has been shown that ^{1}H MRS of lipid signal may be used to study apoptotic signaling (Schmitz et al., 2005).

The evaluations of all the phenotypic readouts previously described are indirect measurement of processes that occur at a molecular level. Although these readouts provide relevant information on the tumor status, they are of generic nature and may lack the specificity required for the final diagnosis: different molecular processes, for example, can lead to almost identical phenotypes. The identification of tumor types is based on its molecular composition. Hence, similar to the histological analysis imaging, methods have to be developed that provide cellular and molecular information (see Assessing Cellular and Molecular: Molecular Imaging Approaches). Alternatively, we might consider compiling the various structural, physiological and metabolic information collected into a fingerprint that may provide the desired degree of specificity in selected cases (see Mathematical Tools for Handling Multi-Parametric Imaging Data: a Classification Problem).

ASSESSING CELLULAR AND MOLECULAR: MOLECULAR IMAGING APPROACHES

Final histological tumor diagnosis/classification is based on the expression of specific molecular markers, hence it becomes obvious that whenever non-invasive imaging should reach that stage, it mast yield temporal-spatially resolved information on the expression of such tumor-specific biomolecules, typically surface epitopes. This asks for molecular imaging solutions visualizing molecular targets or molecular processes occurring at the molecular and cellular levels (Martin, 2011). To achieve this goal, exogenous contrast agents coupled with a molecule that targets specific cell receptors or interacts with specific enzyme or proteins *in vivo* are needed. Quantification of results in molecular imaging refers to the ability to estimate the concentration of the exogenous agent that has reached a specific location at a specific time, and in special cases, to estimate the rate of a biochemical process, such as enzymatic cleavage.

Today, there are a considerable number of publications describing target specific compounds tested in *in vitro* assays that have the potential for *in vivo* imaging; yet only few studies are reported with living organism.

Antibody-based imaging agents constitute a large majority of tumor specific probes (Rudin, 2005b). The tyrosine kinase receptor Her-2/neu, for example, is a protein over-expressed on the surface of breast cancer cells, and other human tumors (Slamon et al., 1989). Approximately 30% of mammary carcinomas express this epithelial growth factor receptor. High expression levels correspond to poor prognosis; hence, Her-2/neu may constitute an attractive target for immunotherapeutic agents, such as humanized monoclonal antibody trastuzumab (Herceptin). By labeling trastuzumab with a superparamagnetic iron-oxide nanoparticles (SPIO) a specific agent able to target cancer cells that overexpressed Her-2/neu could be designed though *in vivo* validation of the approach is still lacking (Smith, 2010; Artemov et al., 2003).

Tissue homeostasis is normally achieved by a tight regulation of proliferation, differentiation, and apoptosis. Apoptosis, or programmed cells death, is downregulated in cancer cells. A general therapeutic strategy may be therefore to induce apoptosis. Development of such treatments would benefit from imaging assays that specifically target molecular players involved in apoptotic signaling or cell surface marker that are specifically expressed on the surface of cells undergoing programmed cell death (Rudin, 2005b). For example, cells undergoing apoptosis redistribute aminophospholipids, primarily phosphatidylserine, to the outer layer of the cell membrane. Phagocytic cells, thus constituting a signal for cell removal, recognize exposed phosphatidylserines. Phosphatidylserine is recognized by peptidic molecules such as annexin-V and synaptogamin I. The latter has been labeled with SPIO nanoparticles and used *in vivo* as apoptosis-specific contrast agent. The nanoparticulate probe can leave the vascular bed in tumors since tumor vessels are immature and leaky, hence uptake is likely to be non-specific. Nevertheless it could be shown that the target specific probe was better retained in subcutaneously implanted tumors in mice while non-targeted SPIO nanoparticles were rapidly cleared from the tumor site (Zhao et al., 2001).

Molecular imaging can also be used as a complementary tool to monitor angiogenesis. In particular, it offers the possibility to differentiate angiogenic vessels from normal blood vessels by detecting differences in the expression of molecular markers (McDonald and Choyke, 2003). In the angiogenic cascade, different cell surface receptors, including the $\alpha_v\beta_3$-integrin, are strongly expressed on activated endothelial cells. Mulder et al. (2005) have described the possibility to imaging angiogenesis using $\alpha_v\beta_3$-specific bimodal lipidic nanoparticle both with MRI and fluorescence imaging.

The motivation for using MRI-based contrast agents, instead of other imaging modalities, is the possibility to combine together both the target-specific information with the high anatomical definition. Moreover, MRI is able to provide three-dimensional imaging which enables the possibility for an accurate quantification of the probe concentration, which otherwise is not be possible in the case of two-dimensional techniques as SPECT or optical imaging. The drawback of MRI approach is the low sensitivity, i.e., high local concentration of the reporter construct is required to induce detectable changes in the relaxation rates (Rudin, 2005b). In addition, MRI reporter molecules are in general bulky and may not easily reach the target site. However, for tumors this might be less an issue due to the leaky vasculature. Today, none of the MRI based target-specific probes has been approved for clinical use.

MATHEMATICAL TOOLS FOR HANDLING MULTI-PARAMETRIC IMAGING DATA: A CLASSIFICATION PROBLEM

In each three-dimensional image dataset the object (tumor) is characterized by a set of voxels, with parameter values (features) that are characteristic for the respective measurement attributed to every voxel. Examples are values for the relaxation time, apparent water diffusion coefficient, or vascular permeability. Assuming that the dataset are properly coregistered all voxels $v_{x,y,z}$ are characterized by a vector, whose elements are the parameter values f_i allocated to the various measurements, i.e.,

$$v_{x,y,z;t} = v_{x,y,z;t}\ (f_1, f_2, ..., f_N).$$

The dimension of this data set is $D \times T \times N$, where D is the number of voxels, T the number of time points measured ($T = 1$ for static measurements) and N the number of features evaluated.

In mathematical terms these set of voxels (three-dimensional maps) form a dataset that contains all the information collected for the tumor. Although all the data are stored in a simple structure as a basic database, it is not easy to extract and quantify information from it. Usually, radiologists consider just few features and mentally divide the tumor in macro-regions, for which individual parameters are analyzed. Obviously this type of analysis discards many the majority of features contained in the dataset and the validity of conclusion critically depends on the experience of the reader. There is no way for human brain to systematically process all the available information voxel by voxel.

The three-dimensional maps contain all measured information on the object reflecting both morphological aspects and physiological behavior. Information regarding the heterogeneity of the object is intrinsically contained. Taking into account this huge amount of information requires mathematical tools that allow a data reduction in a robust manner. One output of such tools is to classify each voxel of the tumor according to the measured features, and finally generate a map of the different tissues types present in the tumor. Several mathematical methods, which come from the field of information theory, have been developed for this purpose. A schematic workflow of the quantification process is shown in **Figure 8**.

EXTRACTING OBJECT FEATURES FOR CLASSIFICATION

As mentioned before, all the information are stored in a *dataset*, where the *features* are any kind of map (measured by MRI, **Figure 9**) and the *subject* are the individual voxels voxel of the three-dimensional matrix.

The first step of the classification process consists of the selection of useful features from the dataset. This process called *feature selection* aims at taking into account only features that contain significant and non-redundant information in order to minimize the confusion intrinsic noise of the data (Umbaugh, 2011). For this purpose different approaches, that describe the variability of the dataset, can be pursued. The traditional way, which comprises a set of techniques that perform a simultaneous statistical analysis of all features, is called *multivariate analysis*. Such techniques include multivariate analysis of variance (MANOVA), principal component analysis (PCA), factor analysis, multidimensional scaling, and correspondence analysis. All of them have as goal to determine a new set of synthetic variables that best represent the samples in a statistical interval.

Another approach consist of considering all the features and assign them a ranking score according to their discriminant power and accuracy, and then simply select the top ranked ones as final features used for the classification (Press et al., 2007; Zacharaki et al., 2009). These methods can be divided in three main categories: filter algorithm, wrapper, and embedded methods. For a comprehensive mathematical description of these methods the reader is referred to (Guyon and Elisseff, 2003).

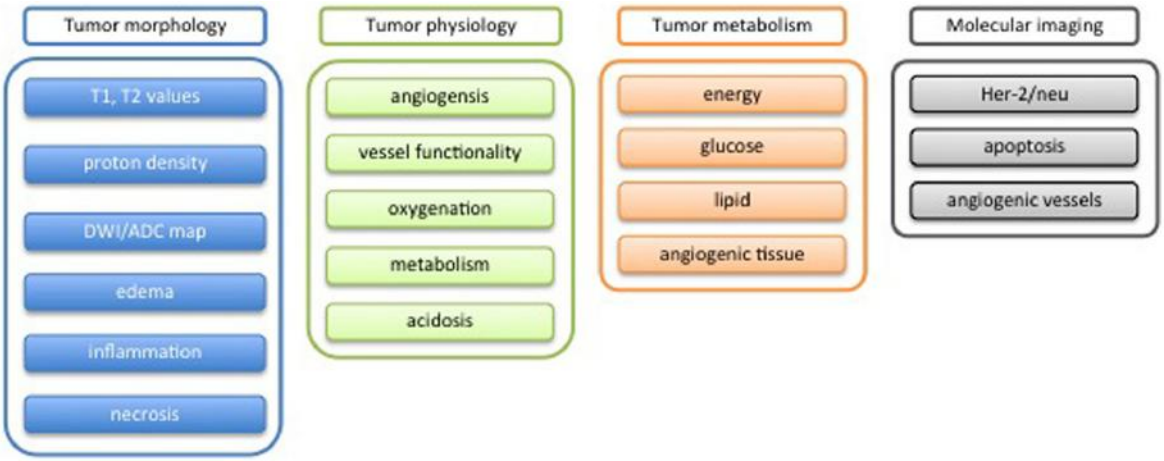


FIGURE 8 | Scheme for potential tumor phenotypic characterization by mean of MRI.

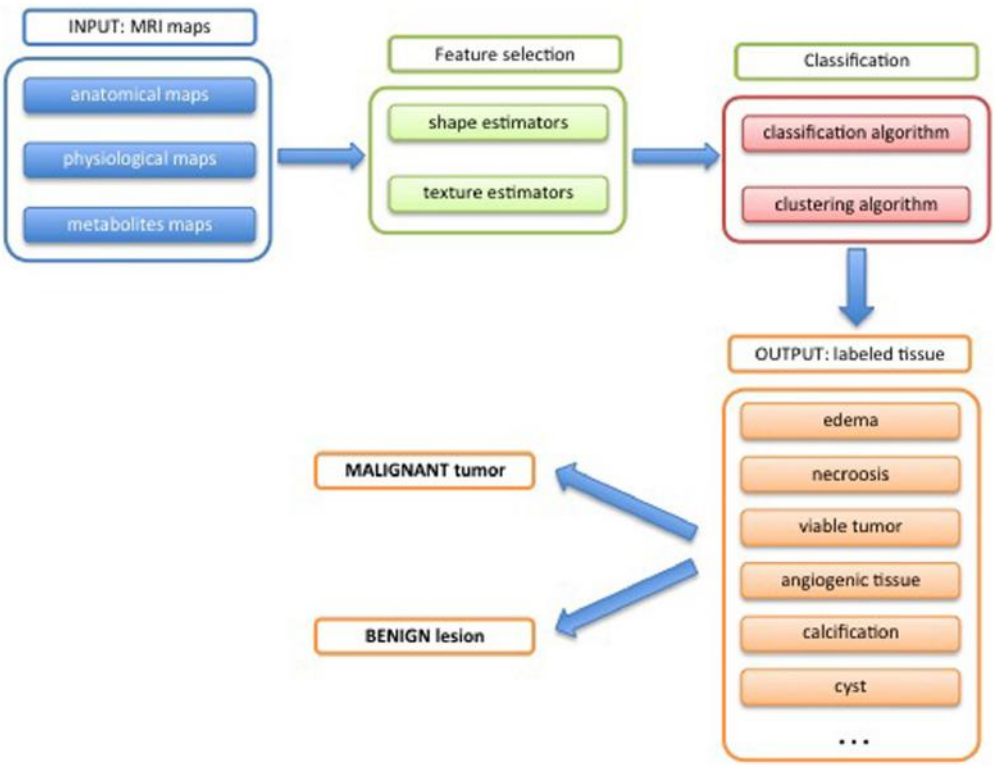


FIGURE 9 | Schematic workflow of the quantification process.

Another important issue is the quantifications of the characteristics inherent in 3D feature maps. In other words, specific estimators that take into account the heterogeneity and the complexity of the object (tumor) are determined (Dominietto et al., 2012). Two types of estimators are commonly used: shape and texture estimators. The first group describes the geometry of the object (whole tumor or specific region), and extracts shape descriptors such as volume, surface area, compactness and signature (Rangayyan and Nguyen, 2007; Rangayyan et al., 2010). Texture estimators are related with the contents of the object and in particular to its texture by means of a set of estimators as fractal dimension (Lopes and Betrouni, 2009), lacunarity (Plotnick et al., 1996), Laws' measures (Rangayyan, 2005), and Haralick's measures (Haralick, 1979). Both shape and texture estimators also used in the geometrical segmentation of anatomical structure (Rangayyan, 2005).

CLASSIFICATION

For classification two common techniques are currently used: pattern recognition and clustering technique (Umbaugh, 2010).

In general terms, for a given group of objects (i.e., different kinds of tissue or different type of tumors), pattern recognition algorithms aim at identifying the individual objects and assign them the correct label. In order to perform this operation, the algorithm has been "trained" previously with a dataset consisting of known objects (training dataset) by means of which it learns to recognize the objects from their features. This process is called supervised machine learning (Bishop, 2006).

The clustering approach is different as it does not require previous knowledge on the objects. Briefly, for analyzing multiparametric static data each voxel represent a subject in a $N.D$-dimensional space, where N is the number of features and D the number of voxels. Voxels that share similar properties will have similar features values and therefore will form a group (or cluster) of points in the $N.D$ space. The objective of using the cluster algorithm is to identify the different groups of points (Theodoridis and Koutroumbas, 2006). The combination of features expressed by each group characterizes its morphological, physiological, metabolic, or molecular properties: it is therefore necessary, but not always straightforward, to translate the feature fingerprint into biomedically relevant information.

For both approaches, the most critical point is feature selection as subsequent tissue classification, e.g., differentiating tumor from healthy tissue or classifying subregions within a tumor, critically depends on the discriminative nature of the features.

Most of the studies to classify tumor tissues relate to brain tumors. Brain, in fact, offers many advantages related with the image acquisitions: easy positioning and fixation, absence of or minimal physiological movements, availability of several anatomical landmarks that renders co-registration rather straightforward in case of multi-modalities acquisitions. Different approaches have been described in the literature to classify and segment brain tumors using texture analysis (Qurat-Ul-Ain et al., 2010), neural networks (Arizmendi et al., 2012), linear discriminant analysis decision tree support vector machine (Zacharaki et al., 2009), and clustering (Jagadeesan and Sivanandam, 2013). Similar studies have been reported for breast tumor in order first to discriminate between malignant tumors and benign microcalcifications (Rangayyan and Nguyen, 2007; Mu et al., 2008), and second to classify tumor lesions (Zheng et al., 2007; Tang et al., 2009; Glasser et al., 2013).

IN VIVO HISTOLOGY USING MRI/MULTIMODAL ANALYSIS: POTENTIAL AND ISSUES

Multiple features have to be evaluated in order to comprehensively characterize biological tissue. Histological analysis, the gold standard for such investigation, used morphological features as well as specific molecular markers to unambiguously identify a specific tissue type. Yet, histology is based on tissue specimen, which for diagnosis are typically obtained via biopsy. Standard biopsy involves focal sampling of only small portions of tissue, and hence carries the risk, that critical regions may be missed in particular when sampling highly heterogeneous tissue such as tumors. The possibility to acquire *in vivo* 3D multi-parametric information on tissues, in our context tumors, in a non-invasive manner might offer important benefits in management of cancer patients. Compared to biopsy, imaging (MRI) based tissue characterization allows analyzing the whole tumor yielding information over its entire volume thereby avoiding the problem of sampling errors. As the measurement is non-invasive, changes in tissue features can be monitored longitudinally, which is highly relevant for prognosis and for evaluating therapy response. The comprehensive nature of tissue analysis provided by imaging supports histological analysis by guiding biopsy sampling thereby minimizing the possibility of sampling errors.

An important advantage of the *in vivo* measurement is the possibility to study physiological processes, which evidently cannot be assessed *ex vivo*. Measurements of processes such as tumor angiogenesis, perfusion, metabolism, or oxygen consumption provide essential information for determining the stage of the tumor. Also it has been shown that such readouts may be early indicators of therapy response, proceeding morphological changes. Similar to morphological features, tumor physiological and metabolic parameters are highly heterogeneous, for example different tumor stages may coexist in the same proliferative mass in glioma patients (Zacharaki et al., 2009). Apart from spatial heterogeneity tumor physiology and metabolism also fluctuate over time (Bonadonna et al., 2007).

In vivo tissue characterization based on imaging has emerged as important tool for the detection and characterization of solid tumors including metastases (Mia, 2011). Today, MRI together with PET (Positron Emission Tomography), SPECT (Single Photon Emission Computer Tomography), CT (Computer Tomography), and US (Ultra Sound) provide a platform that provides multiparametric information characterizing tumor morphology, physiology, metabolism as well as cellular and molecular properties. These techniques are currently used in the clinic to gain as comprehensive information as possible before deciding the best treatment for the patient. Nevertheless, the evaluation of this huge amount of data is usually qualitative and relies on skills of the radiologist. A standardize quantitative evaluation, which gives robust and reproducible results is at the moment missing.

At present there is a huge diversity of imaging/MRI methods that are used in experimental animal studies that provide the multiparametric information required for using the classification tools. However, only a few are being used in the clinics, standard features derived from DCE, FLAIR, T1w, and T2w images and used as qualitative indicators of tumor stage (Young, 2007). More sophisticated techniques as DTI (Diffusion Tensor Imaging), MRS together with machine learning infrastructure, can provide complementary features that better characterize tumor physiology and micro-environment behavior, which would enhance the value of multiparametric analysis. It is important to introduce such method in a standardized manner into radiological practice.

Obviously MRI does not reach microscopic resolution; (Heyn et al., 2005; Martin, 2011), for *in vivo* experiments, the detection limit is in the range between 100 and 500 cells (Heyn et al., 2005; Muja and Bulte, 2009). This is relevant insofar, as final diagnosis is based on the cellular (type and shape) and molecular information (surface epitopes expressed by the cells) derived

from histology. In order to reach this detail of information at the macroscopic level sampled by MRI, target specific contrast agents have to be used. We have seen, that such agents can be developed; yet there are substantial hurdles to overcome, before such agents will make it to the clinics. Scientific hurdles mainly relate to probe specificity and even more so probe delivery. MRI contrast agents are bulky and in general do not cross tissue barriers (membranes). Despite substantial efforts, this still constitute a major problem. The second hurdle relates to economics: development of such an agent is expensive. MRI probes are not administered in tracer amounts, which requires full safety and toxicology analysis. Multicenter clinical trials to demonstrate diagnostic relevance have to be carried. The complexity of developing MRI contrast agents to the market is reflected by the fact that only a very small number of generic agents is currently available for clinical use and it is unlikely that this is going to change in the near future. Hence, MRI methods to be used in clinical setting have to exploit endogenous contrast and rely on the contrast agents currently available. Nevertheless, together with spectroscopic readouts this already constitutes a fair basis for tissue characterization.

Multiparametric imaging based tumor characterization using morphological, physiological, metabolic – and eventually also cellular and molecular – features that can be monitored longitudinally in individual patients might open a way to personalization of the treatment. Today, for many tumor standard treatment protocols that are nevertheless tuned to the specific situation of each patient, are being pursued. This approach does not permit to exploit all possibilities offered today for tumor treatment. Highly specific drugs, new detailed reclassifications of tumor diseases, genetic characterization of several tumors as well as improvements in diagnostic technologies are dramatically changing the landscape of oncology toward patient-specific personalized treatments (Tursz et al., 2011). On the other hand, given the high genetic instability of tumors, it has been questioned whether such approaches are in fact viable (Gillies et al., 2012). Nevertheless, it is beyond doubt that the combined analysis of multi-parametric readouts will improve the diagnostic accuracy, which ultimately should translate into an improved management of cancer patients

REFERENCES

Ahrens, E. T., and Bulte, J. W. (2013). Tracking immune cells in vivo using magnetic resonance imaging. *Nat. Rev. Immunol.* 13, 755–763. doi: 10.1038/nri3531

Alic, L., van Vliet, M., van Dijke, C. F., Eggermont, A. M., Veenland, J. F., and Niessen, W. J. (2011). Heterogeneity in DCE-MRI parametric maps: a biomarker for treatment response? *Phys. Med. Biol.* 56, 1601–1616. doi: 10.1088/0031-9155/56/6/006

Arizmendi, C., Vellido, A., and Romero, E. (2012). Classification of human brain tumours from MRS data using discrete wavelet transform and Bayesian neural networks. *Expert Syst. Appl.* 39, 5223–5232. doi: 10.1016/j.eswa.2011.11.017

Artemov, D., Mori, N., Okollie, B., and Bhujwalla, Z. M. (2003). MR molecular imaging of the Her-2/neu receptor in breast cancer cells using targeted iron oxide nanoparticles. *Magn. Reson. Med.* 49, 403–408. doi: 10.1002/mrm.10406

Barbier, E. L., Lamalle, L., and Decorps, M. (2001). Methodology of brain perfusion imaging. *J. Magn. Reson. Imaging* 13, 496–520. doi: 10.1002/jmri.1073

Bhakoo, K., Chapon, C., Jackson, J., and Jones, W. (2006). "Application of MRI to cell tracking," in *Modern Magnetic Resonance*, ed. G. Webb (Dordrecht: Springer Netherlands), 879–890.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York: Springer-Verlag.

Bonadonna, G., Robustelli Della Cuna, G., and Valagussa, P. (2007). *Medicina Oncologica*. Amsterdam: Elsevier.

Bullitt, E., and Gerig, G. (2003). Measuring tortuosity of the intracerebral vasculature from MRA images. *IEEE Trans. Med. Imaging* 22, 1163–1171. doi: 10.1109/TMI.2003.816964

Bulte, J. W. (2009). In vivo MRI cell tracking: clinical studies. *AJR Am. J. Roentgenol.* 193, 314–325. doi: 10.2214/AJR.09.3107

Chandra, T., Pukenas, B., Mohan, S., and Melhem, E. (2012). Contrast-enhanced magnetic resonance angiography. *Magn. Reson. Imaging Clin. N. Am.* 20, 687–698. doi: 10.1016/j.mric.2012.08.007

Coussens, L. M., and Werb, Z. (2002). Inflammation and cancer. *Nature* 420, 860–867. doi: 10.1038/nature01322

Damadian, R. (1971). Tumor detection by nuclear magnetic resonance. *Science* 171, 1151–1153. doi: 10.1126/science.171.3976.1151

DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G., and Thompson, C. B. (2008). The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* 7, 11–20. doi: 10.1016/j.cmet.2007.10.002

Denysenko, T., Gennero, L., Roos, M. A., Melcarne, A., Juenemann, C., Faccani, G., et al. (2010). Glioblastoma cancer stem cells: heterogeneity, microenvironment and related therapeutic strategies. *Cell Biochem. Funct.* 28, 343–351. doi: 10.1002/cbf.1666

Dominietto, M. (2012). *Fractal Physiology of Tumor Angiogenesis*: Saarbrücken: LAP Lambert Academic Publishing.

Dominietto, M., Lehmann, S., Keist, R., Rudin, M. (2012). Pattern analysis accounts for heterogeneity observed in MRI studies of tumor angiogenesis. *Magn. Reson. Med.* 70, 1481–1490. doi: 10.1002/mrm.24590

Drevelegas, A., and Papanikolaou, N. (2011). "Imaging modalities in brain tumors," in *Imaging of Brain Tumors with Histological Correlations*, ed. A. Drevelegas (Berlin: Springer), 13–33. doi: 10.1007/978-3-540-87650-2_2

Feng, Y., Jeong, E. K., Mohs, A. M., Emerson, L., and Lu, Z. R. (2008). Characterization of tumor angiogenesis with dynamic contrast-enhanced MRI and biodegradable macromolecular contrast agents in mice. *Magn. Reson. Med.* 60, 1347–1352. doi: 10.1002/mrm.21791

Fishman, J. E., Joseph, P. M., Carvlin, M. J., Saadi-Elmandjra, M., Mukherji, B., and Sloviter, H. A. (1989). In vivo measurements of vascular oxygen tension in tumors using MRI of a fluorinated blood substitute. *Invest. Radiol.* 24, 65–71. doi: 10.1097/00004424-198901000-00014

Gadian, D. G., and Radda, G. K. (1981). NMR studies of tissue metabolism. *Annu. Rev. Biochem.* 50, 69–83. doi: 10.1146/annurev.bi.50.070181.000441

Gallagher, F. A., Kettunen, M. I., and Brindle, K. M. (2011). Imaging pH with hyperpolarized 13C. *NMR Biomed.* 24, 1006–1015. doi: 10.1002/nbm.1742

Gatenby, R. A., and Gillies, R. J. (2004). Why do cancers have high aerobic glycolysis? *Nat. Rev. Cancer* 4, 891–899. doi: 10.1038/nrc1478

Gillies, R. J., Verduzco, D., and Gatenby, R. A. (2012). Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat. Rev. Cancer* 12, 487–493. doi: 10.1038/nrc3298

Glasser, S., Niemann, U., Preim, U., Preim, B., and Spiliopoulou, M. (2013). "Classification of benign and malignant DCE-MRI breast tumors by analyzing the most suspect region," in *Bildverarbeitung for die Medizin 2013*, eds H.-P. Meinzer, T. M. Deserno, H. Handels, and T. Tolxdorff (Berlin: Springer), 45–50.

Gross, S., Gilead, A., Scherz, A., Neeman, M., and Salomon, Y. (2003). Monitoring photodynamic therapy of solid tumors online by BOLD-contrast MRI. *Nat. Med.* 9, 1327–1331. doi: 10.1038/nm940

Guyon, I., and Elisseff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Hanahan, D. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proc. IEEE* 67, 786–804. doi: 10.1109/PROC.1979.11328

Heppner, G. H. (1984). Tumor heterogeneity. *Cancer Res.* 44, 2259–2265.

Heverhagen, J. T., Bourekas, E., Sammet, S., Knopp, M. V., and Schmalbrock, P. (2008). Time-of-flight magnetic resonance angiography at 7 Tesla. *Invest. Radiol.* 43, 568–573. doi: 10.1097/RLI.0b013e31817e9b2c

Heyn, C., Bowen, C. V., Rutt, B. K., and Foster, P. J. (2005). Detection threshold of single SPIO-labeled cells with FIESTA. *Magn. Reson. Med.* 53, 312–320. doi: 10.1002/mrm.20356

Heywang-Kobrunner, S. H., Viehweg, P., Heinig, A., and Kuchler, C. (1997). Contrast-enhanced MRI of the breast: accuracy, value, controversies, solutions. *Eur. J. Radiol.* 24, 94–108. doi: 10.1016/S0720-048X(96)01142-4

Hong, H., Yang, Y., Zhang, Y., and Cai, W. (2010). Non-invasive cell tracking in cancer and cancer therapy. *Curr. Top. Med. Chem.* 10, 1237–1248. doi: 10.2174/156802610791384234

Huse, J. T., Holland, E., and DeAngelis, L. M. (2013). Glioblastoma: molecular analysis and clinical implications. *Annu. Rev. Med.* 64, 59–70. doi: 10.1146/annurev-med-100711-143028

Jiang, L., and Weatherall, P. T. (2013). Blood oxygenation level-dependent (BOLD) contrast magnetic resonance imaging (MRI) for prediction of breast cancer chemotherapy response: a pilot study. *J. Magn. Reson. Imaging* 37, 1083–1092. doi: 10.1002/jmri.23891

Joseph, P. M., Fishman, J. E., Mukherji, B., and Sloviter, H. A. (1985). In vivo 19F NMR imaging of the cardiovascular system. *J. Comput. Assist. Tomogr.* 9, 1012–1019. doi: 10.1097/00004728-198511000-00003

Kiselev, V. G., Strecker, R., Ziyeh, S., Speck, O., and Hennig, J. (2005). Vessel size imaging in humans. *Magn. Reson. Med.* 53, 553–563. doi: 10.1002/mrm.20383

Kleijn, A., Chen, J. W., Buhrman, J. S., Wojtkiewicz, G. R., Iwamoto, Y., Lamfers, M. L., et al. (2011). Distinguishing inflammation from tumor and peritumoral edema by myeloperoxidase magnetic resonance imaging. *Clin. Cancer Res.* 17, 4484–4493. doi: 10.1158/1078-0432.CCR-11-0575

Lemasson, B., Valable, S., Farion, R., Krainik, A., Remy, C., and Barbier, E. L. (2013). In vivo imaging of vessel diameter, size, and density: a comparative study between MRI and histology. *Magn. Reson. Med.* 69, 18–26. doi: 10.1002/mrm.24218

Liu, G., Li, Y., Sheth, V. R., and Pagel, M. D. (2012). Imaging in vivo extracellular pH with a single paramagnetic chemical exchange saturation transfer magnetic resonance imaging contrast agent. *Mol. Imaging* 11, 47–57.

Lopes, R., and Betrouni, N. (2009). Fractal and multifractal analysis: a review. *Med. Image Anal.* 13, 634–649. doi: 10.1016/j.media.2009.05.003

Mark Haacke, E. (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Hoboken, NJ: Wiley-Liss.

Haacke, M. E., Brown, R. W., Thompson, M. R., and Venkatesan, R. (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design,* 1st Edn. Hoboken, NJ: Wiley-Liss.

Marmé, D., and Fusenig, N. E. (2007). *Tumor Angiogenesis: Basic Mechanisms and Cancer Therapy.* Berlin: Springer.

Martin, L. (2011). *Molecular and Cellular Imaging. Quantitative MRI in Cancer, Imaging in Medical Diagnosis and Therapy*. Oxford: Taylor & Francis.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* 12, 323–334. doi: 10.1038/nrc3261

McDonald, D. M., and Choyke, P. L. (2003). Imaging of angiogenesis: from microscope to clinic. *Nat. Med.* 9, 713–725. doi: 10.1038/nm0603-713

Mia, L. (2011). *Clinical Assessment of the Response of Tumors to Treatment with MRI. Quantitative MRI in Cancer, Imaging in Medical Diagnosis and Therapy*. Oxford: Taylor & Francis.

Mu, T., Nandi, A. K., and Rangayyan, R. M. (2008). Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers. *J. Dig. Imaging* 21, 153–169. doi: 10.1007/s10278-007-9102-z

Muja, N., and Bulte, J. W. (2009). Magnetic resonance imaging of cells in experimental disease models. *Prog. Nucl. Magn. Reson. Spectrosc.* 55, 61–77. doi: 10.1016/j.pnmrs.2008.11.002

Mulder, W. J., Strijkers, G. J., Habets, J. W., Bleeker, E. J., van der Schaft, D. W., Storm, G., et al. (2005). MR molecular imaging and fluorescence microscopy for identification of activated tumor endothelium using a bimodal lipidic nanoparticle. *FASEB J.* 19, 2008–2010.

Najafi, M., Soltanian-Zadeh, H., Jafari-Khouzani, K., Scarpace, L., and Mikkelsen, T. (2012). Prediction of glioblastoma multiform response to bevacizumab treatment using multi-parametric MRI. *PLoS ONE* 7:e29945. doi: 10.1371/journal.pone.0029945

Ng, T. C., Grundfest, S., Vijayakumar, S., Baldwin, N. J., Majors, A. W., Karalis, I., et al. (1989). Therapeutic response of breast carcinoma monitored by 31P MRS in situ. *Magn. Reson. Med.* 10, 125–134. doi: 10.1002/mrm.1910100112

Nilesh, M., and Quarles, C. C. (2011). *Imaging Tissue Oxygenation Status with MRI. Quantitative MRI in Cancer, Imaging in Medical Diagnosis and Therapy*. Oxford: Taylor & Francis.

O'Connor, J. P., Rose, C. J., Jackson, A., Watson, Y., Cheung, S., Maders, F., et al. (2011). DCE-MRI biomarkers of tumour heterogeneity predict CRC liver metastasis shrinkage following bevacizumab and FOLFOX-6. *Br. J. Cancer* 105, 139–145. doi: 10.1038/bjc.2011.191

Ogawa, S., Lee, T. M., Nayak, A. S., and Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* 14, 68–78. doi: 10.1002/mrm.1910140108

Okamoto, K., Ito, J., Ishikawa, K., Sakai, K., and Tokiguchi, S. (2000). Diffusion-weighted echo-planar MR imaging in differential diagnosis of brain tumors and tumor-like conditions. *Eur. Radiol.* 10, 1342–1350. doi: 10.1007/s003309900310

Pacheco-Torres, J., Calle, D., Lizarbe, B., Negri, V., Ubide, C., Fayos, R., et al. (2011). Environmentally sensitive paramagnetic and diamagnetic contrast agents for nuclear magnetic resonance imaging and spectroscopy. *Curr. Top. Med. Chem.* 11, 115–130. doi: 10.2174/156802611793611904

Pierre, V. C., Botta, M., Aime, S., and Raymond, K. N. (2006). Fe(III)-templated Gd(III) self-assemblies-a new route toward macromolecular MRI contrast agents. *J. Am. Chem. Soc.* 128, 9272–9273. doi: 10.1021/ja061323j

Plotnick, R. E., Gardner, R. H., Hargrove, W. W., Prestegaard, K., and Perlmutter, M. (1996). Lacunarity analysis: a general technique for the analysis of spatial patterns. *Phys. Rev. E Stat. Phys. Plasmas. Fluids Relat. Interdiscip. Topics* 53, 5461–5468. doi: 10.1103/PhysRevE.53.5461

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Qurat-Ul-Ain, Q.-U.-A., Latif, G., Kazmi, S. B., Jaffar, M. A., and Mirza, A. M. (2010). "Classification and segmentation of brain tumor using texture analysis," in *Proceedings of the 9th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases.* Cambridge: World Scientific and Engineering Academy and Society, 147–155.

Raghunand, N., Zhang, S., Sherry, A. D., and Gillies, R. J. (2002). In vivo magnetic resonance imaging of tissue pH using a novel pH-sensitive contrast agent, GdDOTA-4AmP. *Acad. Radiol.* 9(Suppl. 2), S481–S483. doi: 10.1016/S1076-6332(03)80270-2

Raghunand, N. (2006). "Tissue pH measurement by magnetic resonance spectroscopy and imaging," in *Magnetic Resonance Imaging,* Vol. 124, *Methods in Molecular Medicine*, ed. P. Prasad (Totowa, NJ: Humana Press), 347–364.

Rangayyan, R. M. (2005). *Biomedical Image Analysis.* Boca Raton, FL: London CRC Press.

Rangayyan, R. M., and Nguyen, T. M. (2007). Fractal analysis of contours of breast masses in mammograms. *J. Dig. Imaging* 20, 223–237. doi: 10.1007/s10278-006-0860-9

Rangayyan, R. M., Oloumi, F., and Nguyen, T. M. (2010). Fractal analysis of contours of breast masses in mammograms via the power spectra of their signatures. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2010, 6737–6740.

Robinson, S. P., Howe, F. A., and Griffiths, J. R. (1995). Noninvasive monitoring of carbogen-induced changes in tumor blood flow and oxygenation by functional magnetic resonance imaging. *Int. J. Radiat. Oncol. Biol. Phys.* 33, 855–859. doi: 10.1016/0360-3016(95)00072-1

Robinson, S. P., Rijken, P. F., Howe, F. A., McSheehy, P. M., van der Sanden, B. P., Heerschap, A., et al. (2003). Tumor vascular architecture and function evaluated by non-invasive susceptibility MRI methods and immunohistochemistry. *J. Magn. Reson. Imaging* 17, 445–454. doi: 10.1002/jmri.10274

Rodrigues, L. M., Howe, F. A., Griffiths, J. R., and Robinson, S. P. (2004). Tumor R2* is a prognostic indicator of acute radiotherapeutic response in rodent tumors. *J. Magn. Reson. Imaging* 19, 482–488. doi: 10.1002/jmri.20024

Rudin, M. (2005a). *Molecular Imaging: Basic Principles and Applications in Biomedical Research.* London: Imperial College Press.

Rudin, M. (2005b). *Molecular Imaging: Basic Principles and Applications in Biomedical Research.* London: Imperial College Press.

Rudin, M., McSheehy, P. M., Allegrini, P. R., Rausch, M., Baumann, D., Becquet, M., et al. (2005). PTK787/ZK222584, a tyrosine kinase inhibitor of vascular endothelial growth factor receptor, reduces uptake of the contrast agent GdDOTA by

murine orthotopic B16/BL6 melanoma tumours and inhibits their growth in vivo. *NMR Biomed.* 18, 308–321. doi: 10.1002/nbm.961

Schmitz, J. E., Kettunen, M. I., Hu, D. E., and Brindle, K. M. (2005). 1H MRS-visible lipids accumulate during apoptosis of lymphoma cells in vitro and in vivo. *Magn. Reson. Med.* 54, 43–50. doi: 10.1002/mrm.20529

Jagadeesan, R. and Sivanandam, S. N., (2013). A novel clustering and classification based approaches for identifying tumor in MRI brain images. *Int. J. Comput. Appl.* 67, 16–21.

Slamon, D. J., Godolphin, W., Jones, L. A., Holt, J. A., Wong, S. G., Keith, D. E., et al. (1989). Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* 244, 707–712. doi: 10.1126/science.2470152

Smith, T. A. (2010). Towards detecting the HER-2 receptor and metabolic changes induced by HER-2-targeted therapies using medical imaging. *Br. J. Radiol.* 83, 638–644. doi: 10.1259/bjr/31053812

Stummer, W. (2007). Mechanisms of tumor-related brain edema. *Neurosurg. Focus* 22, E8. doi: 10.3171/foc.2007.22.5.9

Sun, P. Z., Cheung, J. S., Wang, E., and Lo, E. H. (2011). Association between pH-weighted endogenous amide proton chemical exchange saturation transfer MRI and tissue lactic acidosis during acute ischemic stroke. *J. Cereb. Blood Flow Metab.* 31, 1743–1750. doi: 10.1038/jcbfm.2011.23

Swartz, M. A., Iida, N., Roberts, E. W., Sangaletti, S., Wong, M. H., Yull, F. E., et al. (2012). Tumor microenvironment complexity: emerging roles in cancer therapy. *Cancer Res.* 72, 2473–2480. doi: 10.1158/0008-5472.CAN-12-0122

Tang, J., Rangayyan, R. M., Xu, J., Naqa, I. E., and Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *Trans. Info Tech. Biomed.* 13, 236–251. doi: 10.1109/TITB.2008.2009441

Taylor, N. J., Baddeley, H., Goodchild, K. A., Powell, M. E., Thoumine, M., Culver, L. A., et al. (2001). BOLD MRI of human tumor oxygenation during carbogen breathing. *J. Magn. Reson. Imaging* 14, 156–163. doi: 10.1002/jmri.1166

Theodoridis, S., and Koutroumbas, K. (2006). *Pattern Recognition.* 3rd Edn, San Diego, CA: Academic Press, Inc.

Thomas, D., and Wells, J. (2011). MR angiography and arterial spin labelling. *Methods Mol. Biol.* 711, 327–345. doi: 10.1007/978-1-61737-992-5_16

Tofts, P. S., and Kermode, A. G. (1991). Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. 1. Fundamental concepts. *Magn. Reson. Med.* 17, 357–367. doi: 10.1002/mrm.1910170208

Tropres, I., Grimault, S., Vaeth, A., Grillon, E., Julien, C., Payen, J. F., et al. (2001). Vessel size imaging. *Magn. Reson. Med.* 45, 397–408. doi: 10.1002/1522-2594(200103)45:3<397::AID-MRM1052>3.0.CO;2-3

Tropres, I., Lamalle, L., Peoc'h, M., Farion, R., Usson, Y., Decorps, M., et al. (2004). In vivo assessment of tumoral angiogenesis. *Magn. Reson. Med.* 51, 533–541. doi: 10.1002/mrm.20017

Tursz, T., Andre, F., Lazar, V., Lacroix, L., and Soria, J. C. (2011). Implications of personalized medicine – perspective from a cancer center. *Nat. Rev. Clin. Oncol.* 8, 177–183. doi: 10.1038/nrclinonc.2010.222

Umbaugh, S. E. (2010). *Digital Image Processing and Analysis: Human and Computer Vision Applications with CVIPtools.* 2nd Edn, Boca Raton, FL: CRC Press, Inc.

Umbaugh, S. E. (2011). *Digital Image Processing and Analysis: Human and Computer Vision Applications with CVIPtools.* Boca Raton, FL: CRC Press.

Verma, S., Turkbey, B., Muradyan, N., Rajesh, A., Cornud, F., Haider, M. A., et al. (2012). Overview of dynamic contrast-enhanced MRI in prostate cancer diagnosis and management. *AJR Am. J. Roentgenol.* 198, 1277–1288. doi: 10.2214/AJR.12.8510

Wang, C. K., Kuo, Y. T., Liu, G. C., Tsai, K. B., and Huang, Y. S. (2000). Dynamic contrast-enhanced subtraction and delayed MRI of gastric tumors: radiologic-pathologic correlation. *J. Comput. Assist. Tomogr.* 24, 872–877. doi: 10.1097/00004728-200011000-00009

Ward, P. S., and Thompson, C. B. (2012). Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell* 21, 297–308. doi: 10.1016/j.ccr.2012.02.014

Weishaupt, D., Köchli, V. D., and Marincek, B. (2006). *How Does MRI Work? An Introduction to the Physics and Function of Magnetic Resonance Imaging.* Berlin: Springer.

Westbrook, C. (2010). *MRI at a Glance.* Malden, MA: Wiley-Blackwell.

Xie, K., Yang, J., Zhang, Z. G., and Zhu, Y. M. (2005). Semi-automated brain tumor and edema segmentation using MRI. *Eur. J. Radiol.* 56, 12–19. doi: 10.1016/j.ejrad.2005.03.028

Young, G. S. (2007). Advanced MRI of adult brain tumors. *Neurol. Clin.* 25, 947–973. doi: 10.1016/j.ncl.2007.07.010

Zacharaki, E. I., Wang, S. M., Chawla, S., Yoo, D. S., Wolf, R., Melhem, E. R., et al. (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn. Reson. Med.* 62, 1609–1618. doi: 10.1002/mrm.22147

Zhang, X., Lin, Y., and Gillies, R. J. (2010). Tumor pH and its measurement. *J. Nucl. Med.* 51, 1167–1170. doi: 10.2967/jnumed.109.068981

Zhang, S., Wu, K., and Sherry, A. D. (1999). A novel pH-sensitive MRI contrast agent. *Angew. Chem. Int. Ed. Engl.* 38, 3192–3194. doi: 10.1002/(SICI)1521-3773(19991102)38:21<3192::AID-ANIE3192>3.0.CO;2-#

Zhao, M., Beauregard, D. A., Loizou, L., Davletov, B., and Brindle, K. M. (2001). Non-invasive detection of apoptosis using magnetic resonance imaging and a targeted contrast agent. *Nat. Med.* 7, 1241–1244. doi: 10.1038/nm1101-1241

Zheng, Y., Baloch, S., Englander, S., Schnall, M. D., and Shen, D. (2007). Segmentation and classification of breast tumor using dynamic contrast-enhanced MR images. *Med. Image Comput. Comput. Assist. Interv.* 10(Pt 2), 393–401.

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

Received: 29 September 2013; paper pending published: 19 November 2013; accepted: 06 December 2013; published online: 13 January 2014.

*Citation: Dominietto M and Rudin M (2014) Could magnetic resonance provide in vivo histology? Front. Genet. **4**:298. doi: 10.3389/fgene.2013.00298*

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

Copyright © 2014 Dominietto and Rudin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 07 April 2014
doi: 10.3389/fgene.2014.00070

Inflammation blood and tissue factors of plaque growth in an experimental model evidenced by a systems approach

Gualtiero Pelosi[1], Silvia Rocchiccioli[1], Antonella Cecchettini[1,2], Federica Viglione[1], Mariarita Puntoni[1], Oberdan Parodi[1], Enrico Capobianco[3,4] and Maria G. Trivella[1]*

[1] Institute of Clinical Physiology, Consiglio Nazionale delle Ricerche, Pisa, Italy
[2] Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy
[3] Laboratory of Integrative Systems Medicine, Institute of Clinical Physiology, Consiglio Nazionale delle Ricerche, Pisa, Italy
[4] Center for Computational Science, University of Miami, Miami, FL, USA

Edited by:
Pietro Lio, University of Cambridge, UK

Reviewed by:
Tianxiao Huan, Framingham Heart Study, USA
Lifan Zeng, Indiana University School of Medicine, USA

***Correspondence:**
Maria G. Trivella, Institute of Clinical Physiology, Consiglio Nazionale delle Ricerche, Via Moruzzi 1, 56124 Pisa, Italy
e-mail: trivella@ifc.cnr.it

Purpose: The multifactorial pathogenesis of coronary atherosclerotic lesion formation has been investigated in a swine model of high cholesterol diet induced atherogenesis and data processed by a systems approach.

Methods: Farm pigs were fed on standard or high cholesterol diet of 8 and 16 weeks duration. Plasma assessment of total cholesterol, HDL, LDL, and ELISA of some cytokines and ICAM-1 were performed on baseline and end-diet samples. Segments of the right coronary artery were incubated for 24 h in serum-free medium to collect secreted proteins and their expression analyzed by mass spectrometry. Data of plasma and tissue factors were processed by a statistical systems inference approach: both histologic parameters of coronary intimal thickness (IT) and of lesion area (LA) were chosen as dependent variables (coronary atherosclerotic burden).

Results: Relations among plasma adhesion molecules, cytokines, lipoproteins, tissue proteins and histology indexes were integrated in a model regression scheme. Bayesian model averaging (BMA) variable selection was chosen as a method to identify relevant factors associated to atherosclerotic burden: TNFα was identified as an associated plasma marker, oxLDL and HDL as relevant lipoproteins; macrophage function related antioxidant Catalase enzyme, lysosome associated Cathepsin D, S100-A10, and Transforming growth factor-beta-induced protein ig-h3 were identified and selected as associated to atherogenesis outcome.

Conclusions: The results of this systems approach are consistent with the hypothesis that, in high cholesterol diet-induced experimental atherogenesis, the interaction between plasma cytokines, lipoproteins and artery-specific proteins, influences lesion initiation and growth. In particular, some macrophage function related proteins are found significantly and positively associated to atherosclerotic burden, suggesting a novel molecular framework into the atherogenesis-inflammatory disorder.

Keywords: systems biomedicine, coronary atherogenesis, swine model, vascular inflammation, Bayesian model averaging

INTRODUCTION

Atherogenesis is the initiating step of atherosclerosis, and can be considered the key-point for a better understanding of the entire process, as several factors and mechanisms are also related to plaque progression in the clinical scenario (Weber and Noels, 2011).

Plaque initiation steps take place in the following environments:

1. Systemic blood environment (proatherogenic or atheroprotective) constituted of inflammatory and lipid factors
2. Endothelial blood-vessel interface, which expresses adhesion molecules for monocyte intra-lesional transfer
3. Sub-endothelial intimal space, where proteoglycans retain LDL
4. Intimal and intima-media interface, scenario for vascular smooth cell (VSMC) phenotype switch and activation toward migratory and proliferative conditions (Libby et al., 2010)

Traditional views of atherosclerosis, basically seen as a lipid-based disorder, have been modified by the recognition of the multifactorial etiology of this disease (Lamon and Hajjar, 2008), involving the interplay of genetic, phenotypic and environmental factors that have to be integrated into a unified scheme. According to this theory, the most likely sequence of events occurring in the initial phase of atherosclerosis comprises vascular dysfunction and/or injury, monocyte recruitment and foam cell formation, lipid deposition, vascular smooth muscle cell proliferation and synthesis of extracellular matrix (Libby, 2002). The interaction of all these factors

confers to the resulting atherosclerotic plaque its typical features.

In this study, circulatory systemic and locally expressed artery factors in a high cholesterol diet animal model of coronary atherogenesis have been collected and inter-related using a Bayesian Model Averaging (BMA) (Leamer, 1978; Raftery, 1995) computational approach, which is also suggested as a useful strategy to unravel novel actors and pathways outlining this complex framework.

A statistical regression framework based on BMA to account for model uncertainty determined by many variables of heterogeneous nature has been used. In such circumstances, the choice of an encompassing model is not easy, and needs to be a statistically reasonable decision. BMA is a suitable model strategy, which presents several advantages and reasonable computational requirements.

MATERIALS AND METHODS

EXPERIMENTAL DESIGN, CIRCULATORY-TISSUE DATA COLLECTION AND HISTOLOGY

Animal experiment protocol was approved by the Animal Care Committee of the Minister of Public Health according with guidelines (protocol number: 06/2009-B-2009/01/26). Atherogenesis has been studied in 13 farm pigs fed on a high cholesterol (4%) high fat (27%) diet for 8 (HF, 4 cases) and 16 weeks (HHF, 6 cases) and controls fed on standard diet (CNTL, 3 cases). Data on plasma lipids, cytokines and cell adhesion markers have been collected before and at the end of the diet period in all animals. Total cholesterol, High Density Lipoprotein (HDL) and triglycerides (TG) were measured by standard enzymatic techniques (Synchron CX9 Pro, Beckman Coulter Inc., Fullerton, CA, USA). Low density lipoprotein (LDL) was calculated according to Friedewald et al. (1972) IL-6, TNFα, and ICAM-1 were purchased by Abcam (Cambridge, UK), while oxLDL was a product of Antibodies-Online (Atlanta, GA, USA).

At the end of diet period, animals were anesthetized by intramuscular administration of 10 mg/kg of Zoletil® and 0.05 mg/Kg of atropine, plus 5 mg/kg/h of propofol intravenous infusion and sacrificed by KCl i.v. bolus injection. Upon heart explantation, a 3 mm long segment of the proximal tract of right coronary artery (RCA), 1 cm below the ostium, was harvested and placed in serum free solution to collect secreted/released proteins (Rocchiccioli et al., 2013)

Following heart fixation in 10% buffered formalin (7–10 days), 5–10 mm thick transverse arterial samples were collected from left main, left anterior descending, left circumflex and right coronary arteries for routine histologic processing for paraffin embedding. Consecutive cross-sections were obtained from each coronary segment (rotary microtome Microm HM 300, Bio-optica) for Haematoxylin and Eosin, Mallory trichrome and Weigert van Gieson staining and examined under light microscopy (Olympus BX43, Italy) from 2× to 40× original magnification. Images were digitized by a video system (Olympus DP20 camera, Italy) interfaced to a computer with dedicated software (CellSens Dimension, Olympus, Italy) for morphometric analysis. Intimal thickness (IT, mm), i.e. maximal radial expansion of the lesion, and lesional area (LA, mm^2) i.e., entire lesion area in each cross-section, were used as representative morphometric indexes of overall atherosclerotic burden in each individual case. Both mean and median of all the IT and LA values of all cross-sectioned coronary lesions of each case were calculated (Viglione et al., 2013).

LIQUID CHROMATOGRAPHY (LC) SEPARATION, MASS SPECTROMETRY (MS) ANALYSES AND DATA POST-PROCESSING

Chromatographic separation of digested peptides obtained from secreted proteins was performed using an Ultimate 3000 nano-HPLC system (LC Packings, DIONEX, USA) and peptides eluted from chromatography were directly processed using TripleTOF™ 5600 mass spectrometer (AB SCIEX, Toronto, Canada) (Rocchiccioli et al., 2013). MS/MS data were processed with ProteinPilot™ Software (AB SCIEX, Toronto, Canada), using the Paragon™ and Pro Group™ Algorithms and SwissProt 2012 as protein database for *Sus scrofa*. The false discovery rate (FDR) analysis was done using the integrated tools in ProteinPilot software and a confidence level of 95% was set. Expression data for proteins were obtained using MarkerView^{TM} software 1.2.1 (AB SCIEX). Normalization of the total artery tissue size was accomplished with a global normalization of profiles (total protein content) using Marker View 1.2 software.

MATHEMATICAL MODEL APPROACH: IMPLEMENTATION OF R ENVIRONMENT, BMA PACKAGE

Circulatory and omics data have been processed and related to histology parameters of mean and median coronary IT and LA of each case of HF and HHF groups. All dependent and independent variables of diet treated cases (HF and HHF) were normalized to average values of standard diet CNTL cases which are taken as reference. The effect of normalization, together with a logarithmic transform taken to minimize variability, is a better control of the wide range of magnitude for the absolute values of histology, circulatory and omics variables. The independent variables are considered plasma lipoproteins (total cholesterol, LDL, HDL) oxidized LDL, circulatory cytokines (IL6 and TNFα, ICAM-1 and several coronary proteins identified by LC-MS reported in the **Supplementary Table 1**. At first, the model has been applied to all diet treated cases as a whole group, while it was subsequently applied to HF and HHF groups separately.

In general, BMA is employed when multiple models may be statistically reasonable, and selecting a single particular model can lead to the underestimation of the uncertainty related to the model form underlying the variables of interest. In such cases, BMA can quickly determine suitable models through specified sets of explanatory variables with high likelihoods. Equivalently, averaging across a large set of such models allows to determine the variables which are relevant to the data generating process for a given set of priors used in the analysis.

The implementation of BMA was done within the R environment (Raftery et al., 2010), by averaging the best models of a certain class, and according to the approximate posterior model probability which was computed in each case. For instance, the class "bicreg" in the BMA R package identifies the linear regression models, and is the one chosen among other possible tested classes. In this way the analysis has been kept at its simplest and

most interpretable level. In particular, the option "iBMA" represents the iterated BMA method for variable selection, and works by repeatedly calling BMA, i.e., iterating through the variables in a fixed order based on some measure of goodness of fit. After each call, only the variables with posterior probability greater than a specified threshold are retained, the rest being replaced by other variables.

The summary function was used to provide concise and summarized information about the variables that have been examined up to the last iteration. Each model, and set of variables, is weighted and the final estimates are constructed as a weighted average of the parameter estimates from each of the models. All the variables are considered, but some are subject to shrinking by setting to zero the model weights, and depending upon features such as the choice of prior (see also Supplementary Material).

POST-PROCESSING OF MODEL RESULTS

The adopted strategies to assess relevance of model selections were:

1. Congruence of model selection by histology indexes.

Among all selected variables, those with only IT or LA association have been discarded. Congruence of selected variables with both maximal radial expansion and circumferential extension of lesions was thus ensured.

2. Congruence of model selection by regression coefficients (value and sign).

It has also been checked whether relevant variables according to step 1 had regression coefficients of comparable size, and similar direction of association (negative or positive sign); this strategy allowed for a more robust combination of factors which strongly relate to atherogenesis outcome. Variables were discarded, when the corresponding coefficients had comparable size and opposite sign, indicating inappropriate selection, as well as when absolute values were very low (<0.001) irrespective of sign congruence.

RESULTS

Circulatory and omics data (**Supplementary Table 1**) have been processed by BMA and related to histology parameters of mean and median coronary IT (mm) and LA (mm^2) of HF and HHF groups. Circulatory data were measured by antibody-based kits and expressed as a concentration in serum. Protein data were measured by mass spectrometry and protein expression was measured by peptide peak area using arbitrary units (normalized counts).

SELECTED VARIABLES BY FIRST IMPLEMENTATION RUN OF THE MODEL: ALL HIGH CHOLESTEROL DIET-TREATED ANIMALS

IT and LA are the dependent variables that have been chosen for BMA approach to provide different and complementary information on atherosclerotic lesions, depending on lesion shape and its mainly eccentric or concentric growth. IT is more representative of maximal radial expansion of the lesion, whilst LA is more related to the circumferential extension.

Also Mean and Median values of the two indexes provide distinct information, mean values being more representative of mild localized rather than of severe and diffuse atherosclerotic changes: different distribution patterns of lesions are present along each coronary artery, related to single lesion severity and extent of coronary involvement. But generally, mean and median values tend to coincide when lesions are present in all examined segments, whilst they diverge when atherosclerotic changes are localized only in few segments, such as in the proximal portion of main coronary arteries (**Figure 1**).

Independent variables selected by the model as associated to atherosclerotic burden and derived from implementations run on all diet treated cases are reported in **Table 1** (lipoproteins and circulatory factors) and **Table 2** (artery secreted proteins). As described in the Methods section, only variables with combined association of IT and LA histology indexes of atherosclerotic changes are reported and considered relevant.

Among lipoproteins, oxLDL, and HDL are found significantly associated to arterial pathology. Plasma cytokine TNFα, as well as adhesion molecule ICAM-1 are also relevantly selected variables.

Among artery secreted proteins, the most selected and associated to all histology indexes of atherosclerotic burden are Catalase (CATA) and Cathepsin D (CATD). Transforming growth factor-beta-induced protein ig-h3(BGH3), S100A10 (S10AA) and Glyceraldehyde-3-phosphate dehydrogenase (G3P) are also selected and are congruent with 3 out of the 4 histology indexes.

SELECTED VARIABLES BY SECOND IMPLEMENTATION RUN OF THE MODEL: TWO DISTINCT GROUPS (HF AND HHF)

The model has been also applied to 8 weeks (HF) and 16 weeks (HHF) high cholesterol diet treated animals separately.

When considering systemic variables, separate analysis of early atherogenesis HF group does not provide further relevant information in addition to what previously derived from model run on pooled data: this is likely due to the limited number of HF cases and/or to the very low grade of atherogenesis after 8 weeks high cholesterol diet.

On the other hand, for local factors, the model selects Moesin and Osteonectine (MOES, SPRC) that had not been picked in the first run, as well as the already selected Apolipoprotein A4 (APOA4), Byglican (HPLN1), G3P, BGH3 and Calpastatin (ICAL), all related to lesion development in model run on HF and also on HHF group data. Annexin 1 (ANXA1) is the only protein selected from HF group omics data and unselected in HHF.

EVALUATION OF REGRESSION COEFFICIENTS IN FIRST AND SECOND IMPLEMENTATION

Regression coefficients (absolute beta values, considering beta as the regression coefficients) of all the selected independent variables have been analyzed as an index of robustness of results and a qualitative measure of their relation with dependent variables IT and LA (mean and/or median values).

In pooled case run (first implementation), analysis of congruence by regression coefficients confirms systemic lipid and proinflammatory variables associated to atherosclerotic burden (association with at least one IT plus one LA index). A positive association is present for oxLDL, HDL and TNFα. On the other

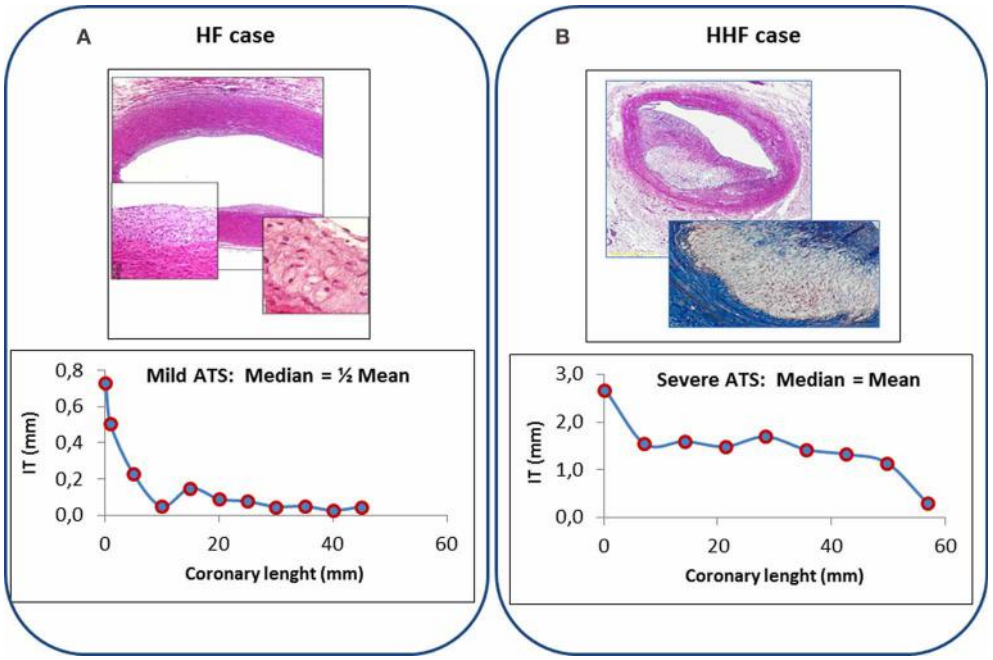


FIGURE 1 | Top: Histologic features of coronary lesions in a typical HF (panel A, fatty streak, H&E 2×, insets 10× and 20×) and HHF case (Panel B, atheroma, H&E 4×, inset Mallory trichrome 10×). Bottom: Coronary profiling (left anterior descending artery) of IT values of observed lesions in 11 consecutive segments of a HF case **(panel A)** and in 9 consecutive segments of a HHF case **(panel B)**. Median of IT values is about one half of the mean of IT values in mild localized atherosclerotic changes of HF case (left), whilst it is equal to the mean of IT values when diffuse severe changes are present (HHF case, right).

Table 1 | Lipoproteins and inflammatory factors.

	Model selected circulatory variables		
	IT mean	**IT median**	**LA mean**
IT median	ICAM-1 BAS, IL6 BAS		
LA mean	TNFα BAS	TNFα END-DIET	
LA median	OX-LDL END-DIET	HDL END-DIET	ICAM-1 END-DIET

Systemic inflammation selected variables: ICAM-1 (congruence with all indexes), TNFα (congruence with IT mean, IT median, LA mean). Lipoproteins selected: ox-LDL (congruence with IT mean, LA median), HDL end-diet (congruence with IT median, LA median).

Table 2 | Artery secreted proteins.

	Model selected proteins		
	IT mean	**IT median**	**LA mean**
IT median	CATA, G3P, S10AA, CPNS1		
LA mean	CATA, CATD,BGH3, S10AA, CPNS1, ANXA4, FABPH	CATA, S10AA, CPNS1, PPCE	
LA median	CATA, BGH3, G3P, CATD, ANXA4	CATA, G3P, PPCE	BGH3, CATA, CATD, PPCE

Selected proteins: CATA (Catalase) is congruent with all indexes and combinations of histology indexes, CATD (Cathepsin D) with all indexes but not all combinations, S10AA (S100 A10) and CPNS1 (Calpain small subunit 1) with IT mean, IT median, LA mean, G3P (Glyceraldehyde-3-phosphate dehydrogenase) with IT mean, IT median, LA median, BGH3 (Transforming growth factor-beta-induced protein ig-h3) and ANXA4 (Annexin A4) with IT mean, LA mean, LA median, PPCE (Prolyl endopeptidase) with IT median, LA mean, LA median.

side, when considering artery specific factor congruence is limited to the combination of one IT and one LA index for CATA (positive association), CATD (positive), BGH3 (positive), S10AA (positive) and for Fatty acid-binding protein 3 (FABPH, with a negative association) (**Figure 2**).

Inclusion of regression coefficients, in variables selected by the model from separate run on HF and HHF groups, strongly restricts the relevance of results. No congruence is present, neither in absolute values nor in the sign of coefficients, whatever combination of dependent variables (histologic indexes) is considered.

DISCUSSION

The aim of this study is to propose a systems biology oriented approach as a tool to associate circulatory and tissue markers with coronary lesion development in a high cholesterol diet swine model. Using animal models, systemic and tissue data can be collected and analyzed at the early stages of diet-accelerated process and can be useful to define the timing of events that are mostly uninvestigable in the clinical setting.

Among animal models of atherogenesis, pig is currently considered the most suitable among those closer to human pathology (Vilahur et al., 2011).

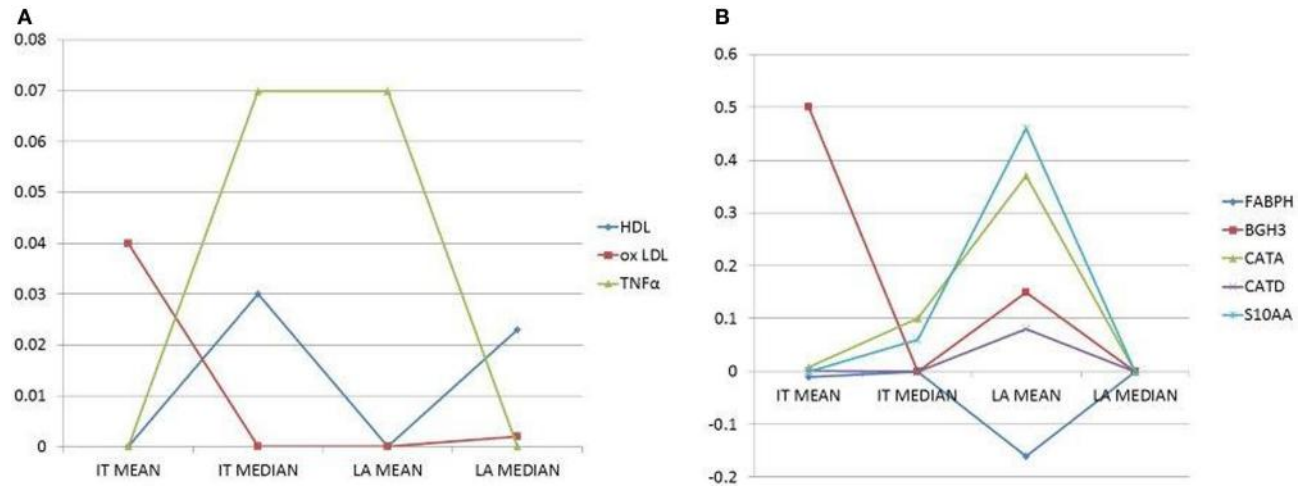


FIGURE 2 | (A) Plots of selected circulatory variables (as shown by different symbols) with their coefficients (Y-axis). **(B)** Plots of selected, secreted proteins (as shown by different symbols) with their coefficients (Y-axis). Histomorphometry data on the X-axis are the dependent variables. Only independent variables (circulatory factors or secreted proteins) with the combined highest coefficients and congruent signs (i.e., inverse, negative sign, or direct, positive sign, correlations), for at least two dependent variables, are chosen and reported.

Coronary histology indexes, plasma lipoproteins, circulatory cytokines, adhesion molecules and coronary specific secreted proteins were provided to the mathematical model and were chosen considering the current knowledge on factors involved in atherogenesis (Libby et al., 2002; Mohler et al., 2008).

The rationale of exploiting systems approaches through statistical models to elucidate the association between all these factors originates from the need of pointing out relationships strongly associated to coronary early atherosclerotic changes. It is known that the interplay between circulatory and tissue markers and the association between molecular factors and plaque growth represent the crossroad of blood-artery wall events during atherogenesis (Döring et al., 2012). Computational tools like those described, which run on a multitude of variables simultaneously, perform variable selection and model optimization, may help toward the ultimate aim of predictive inference, without bringing the burden of noisy and spurious correlative associations.

BMA APPROACH TO EXPERIMENTAL DATA

The model application to the provided data sets, followed by a post-processing exclusion based on the criteria of absolute values and sign of correlation coefficients, has evidenced that no congruence of any of the independent variables considered for all the chosen histology indexes is present, neither in the pooled nor in the separate HF and HHF data implementation runs. This finding is not surprising and underlies the limitations of this approach for pathophysiologic investigations when a reduced number of data is provided to the statistical tool. Despite such limitation, a restricted number of variables (two lipoproteins, one cytokine and five proteins) is finally suggested as robustly associated to both IT and LA morphologic indexes of atherogenesis outcome when HF and HHF cases are pooled. Separate analysis of the two groups does not lead to a robust selection of any variable under the criteria adopted, possibly because of the further reduction of data available for the model.

BIOLOGICAL RELEVANCE OF MODEL RESULTS

The most robust association between dependent (histology indexes of atherosclerotic burden) and independent (circulatory and local factors) variables has been found when considering HF and HHF cases as a single group. This finding may be the consequence of model limitations (low number of early atherogenesis HF cases) although common mechanisms of initiation and of early plaque growth can also be hypothesized.

Circulatory associated variables are LDL, oxLDL, and TNFα and artery-specific variables are CATA, CATD, S100-A10, BGH3, and FABPH. It must be emphasized that, at variance with conventional statistical tools, the BMA mathematical model accounts for all the possible associations and blood-tissue factor interrelations in selecting those relevant for histologically determined atherogenesis outcome.

BMA selection of circulatory variables supports the current view that atherogenesis in a high cholesterol diet experimental model is related to systemic proinflammatory cytokines and adhesion molecules under a LDL-rich blood environment. The impact of inflammation-immunity state on pathology outcome has been demonstrated by several previous experimental and clinical studies, both as a strong proatherogenic determinant of plaque initiation as well as of its progression and evolution (Lamon and Hajjar, 2008; Merched et al., 2008).

By the mathematical model, identified local artery-specific factors, relevantly associated to lesion initiation and growth, are those mainly involved in macrophage/phagocytosis function and immunity-inflammatory pathways (CATA, CATD, S100-A10, FABPH, BGH3) (Haidar et al., 2006; Nacu et al., 2008; O'Connell et al., 2010; Lee et al., 2013). These proteins may be viewed as mediators and possible markers of a local inflammation scenario with pro- and anti-inflammatory elements playing a role in both initiation and early growth of high cholesterol diet-induced coronary atherosclerotic lesions. Among those, negative association is

evidenced only for FABPH, in contrast with current knowledge on the role of this protein in atherogenesis (Lee et al., 2013).

CONCLUDING REMARKS

An integrative systems approach is proposed to study the association between circulatory markers and omics data to coronary atherosclerosis severity. BMA variable selection was chosen as a method to identify relevant factors associated to atherosclerosis. Specifically, TNFα was identified as an associated plasma marker, oxLDL and HDL were confirmed as relevant lipoproteins, macrophage related antioxidant Catalse enzyme, lysosome associated Cathepsin D, S100-A10 and Transforming growth factor-beta-induced protein ig-h3 were selected as associated to atherogenesis outcome.

The proposed approach has been shown to be feasible from a computational standpoint and capable of helping in understanding the association of multilevel factors in atherosclerotic plaque initiation with early growth.

The results of this study suggest a relevant conclusion: in a high-cholesterol diet-induced model of coronary artery disease, systemic inflammation impacts on atherogenesis outcome and it is specifically reflected by macrophage/phagocytosis-related artery-specific protein expression. Further studies integrating genomics, epigenomics and transcriptomics are needed for a better assessment of causative mechanisms and sequence of events in the early phase of atherogenesis in coronary artery disease.

ACKNOWLEDGMENTS

Experimental data and financial support have been obtained from the EU Project ARTreat (FP7-224297 for Large-scale Integrating Project) "Multi-level patient-specific artery and atherogenesis model for outcome prediction, decision support treatment, and virtual hand-on training" (http://www.artreat.org/).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00070/abstract

Supplementary Table 1 | Expression data for secretome proteins and circulatory factors in CNTL, HF and HHF groups.

REFERENCES

Döring, Y., Noels, H., and Weber, C. (2012). The use of high-throughput technologies to investigate vascular inflammation and atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 32, 182–195. doi: 10.1161/ATVBAHA.111.232686

Friedewald, W. T., Levi, R. I., and Fredrickson, D. S. (1972). Estimation of the concentration of low density lipoproteins cholesterol in plasma without use of the ultracentrifuge. *Clin. Chem.* 18, 499–502.

Haidar, B., Kiss, R. S., Sarov-Blat, L., Brunet, R., Harder, C., McPherson, R., et al. (2006). Cathepsin D, a lysosomal protease, regulates ABCA1-mediated lipid efflux. *J. Biol. Chem.* 281, 39971–39981. doi: 10.1074/jbc.M605095200

Lamon, B. D., and Hajjar, D. P. (2008). Inflammation at the molecular interface of atherogenesis: an anthropological journey. *Am. J. Pathol.* 173, 1253–1264. doi: 10.2353/ajpath.2008.080442

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Non-Experimental Data*. New York, NY: Wiley.

Lee, K., Santibanez-Koref, M., Polvikoski, T., Birchall, D., Mendelow, A. D., and Keavney, B. (2013). Increased expression of fatty acid binding protein 4 and leptin in resident macrophages characterises atherosclerotic plaque rupture. *Atherosclerosis* 226, 74–81. doi: 10.1016/j.atherosclerosis.2012.09.037

Libby, P. (2002). Inflammation in atherosclerosis. *Nature* 420, 868–874. doi: 10.1038/nature01323

Libby, P., Di Carli, M., and Weissleder, R. (2010). The vascular biology of atherosclerosis and imaging targets. *J. Nucl. Med.* 51(Suppl. 1), 33S–37S. doi: 10.2967/jnumed.109.069633

Libby, P., Ridker, P. M., and Maseri, A. (2002). Inflammation and atherosclerosis. *Circulation* 105, 1135–1143. doi: 10.1161/hc0902.104353

Merched, A. J., Ko, K., Gotlinger, K. H., Serhan, C. N., and Chan, L. (2008). Atherosclerosis: evidence for impairment of resolution of vascular inflammation governed by specific lipid mediators. *FASEB J.* 22, 3595–3606. doi:10.1096/fj.08-112201

Mohler, E. R., Sarov-Blat, L., Shi, Y., Hamamdzic, D., Zalewski, A., Macphee, C., et al. (2008). Site-specific atherogenic gene expression correlates with subsequent variable lesion development in coronary and peripheral vasculature. *Arterioscler. Thromb. Vasc. Biol.* 28, 850–855. doi: 10.1161/ATVBAHA.107.154534

Nacu, N., Luzina, I. G., Highsmith, K., Lockatell, V., Pochetuhen, K., Cooper, Z. A., et al. (2008). Macrophages produce TGF-β-induced (β-ig-h3) following ingestion of apoptotic cells and regulate MMP14 levels and collagen turnover in fibroblasts. *J. Immunol.* 180, 5036–5044.

O'Connell, P. A., Surette, A. P., and Liwski, R. S. (2010). S100A10 regulates plasminogen dependent macrophage invasion. *Blood* 116, 1136–1146. doi: 10.1182/blood-2010-01-264754

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163. doi: 10.2307/271063

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2010). *BMA: Bayesian Model Averaging*. R package version 3.13. Available online at: http://CRAN.R-project.org/package=BMA

Rocchiccioli, S., Pelosi, G., Rosini, S., Marconi, M., Viglione, F., Citti, L., et al. (2013). Secreted proteins from carotid endarterectomy: an untargeted approach to disclose molecular clues of plaque progression. *J. Trasl. Med.* 11, 260. doi: 10.1186/1479-5876-11-260

Viglione, F., Sbrana, S., Puntoni, M., Rocchiccioli, S., Cecchettini, A., Trivella, M. G., et al. (2013). Circulatory inflammation molecules and extracellular matrix proteoglycans: local and systemic modulated markers in an atherogenesis model. *Eur. Heart J.* 34(Suppl.), 443. doi: 10.1093/eurheartj/eht308.P2399

Vilahur, G., Padro, T., and Badimon, L. (2011). Atherosclerosis and thrombosis: insights from large animal models. *J. Biomed. Biotechnol.* 2011:907575. doi: 10.1155/2011/907575.

Weber, C., and Noels, H. (2011). Atherosclerosis: current pathogenesis and therapeutic options. *Nature* 17, 1410–1422. doi: 10.1038/nm.2538

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: 05 December 2013; accepted: 17 March 2014; published online: 07 April 2014.

*Citation: Pelosi G, Rocchiccioli S, Cecchettini A, Viglione F, Puntoni M, Parodi O, Capobianco E and Trivella MG (2014) Inflammation blood and tissue factors of plaque growth in an experimental model evidenced by a systems approach. Front. Genet. **5**:70. doi: 10.3389/fgene.2014.00070*

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

Copyright © 2014 Pelosi, Rocchiccioli, Cecchettini, Viglione, Puntoni, Parodi, Capobianco and Trivella. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in GENETICS

PERSPECTIVE ARTICLE
published: 22 April 2014
doi: 10.3389/fgene.2014.00080

How integration of global omics-data could help preparing for pandemics – a scent of influenza

Lieuwe D. J. Bos[1,2,3]*, Menno D. de Jong[4], Peter J. Sterk[2] and Marcus J. Schultz[1,3]

[1] Department of Intensive Care Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands
[2] Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands
[3] Laboratory of Experimental Intensive Care and Anesthesiology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands
[4] Department of Medical Microbiology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands

Edited by:
Pietro Lio, University of Cambridge, UK

Reviewed by:
Julio Vera González, University Hospital Erlangen, Germany
Anatoly Sorokin, Russian Academy of Sciences, Russia

***Correspondence:**
Lieuwe D. J. Bos, Department of Intensive Care Medicine, Academic Medical Center, University of Amsterdam, G3–228, Meibergdreef 9, 1105 AZ Amsterdam, Netherlands
e-mail: l.d.bos@amc.uva.nl

Pandemics caused by novel emerging or re-emerging infectious diseases could lead to high mortality and morbidity world-wide when left uncontrolled. In this perspective, we evaluate the possibility of integration of global omics-data in order to timely prepare for pandemics. Such an approach requires two major innovations. First, data that is obtained should be shared with the global community instantly. The strength of rapid integration of simple signals is exemplified by Google's™ Flu Trend, which could predict the incidence of influenza-like illness based on online search engine queries. Second, omics technologies need to be fast and high-throughput. We postulate that analysis of the exhaled breath would be a simple, rapid and non-invasive alternative. Breath contains hundreds of volatile organic compounds that are altered by infection and inflammation. The molecular fingerprint of breath (breathprint) can be obtained using an electronic nose, which relies on sensor technology. These breathprints can be stored in an online database (a "breathcloud") and coupled to clinical data. Comparison of the breathprint of a suspected subject to the breathcloud allows for a rapid decision on the presence or absence of a pathogen.

Keywords: pandemic, exhaled breath, systems biology, diagnosis, metabolomics, metabolite profiling

RATIONALE

Respiratory tract infections are the primary cause of death by communicable diseases (Lopez et al., 2006). The global burden is estimated to be around 3.7 million deaths yearly. Novel emerging or re-emerging infectious diseases could increase the number of victims substantially, as exemplified by the approximately 50 million deaths in the Spanish influenza pandemic in 1918–1919 (Taubenberger and Morens, 2006). Subsequent influenza pandemics and the emergence of novel animal-origin influenza (H5N1, H7N9) and coronaviruses (SARS-CoV, MERS-CoV) that cause severe infections in humans illustrate a continuous and ongoing threat of new pandemics. Intensive farming and changing climate enhance the likelihood of (zoonotic) transmission of animal-origin pathogens to humans and subsequent evolution of such pathogens to efficient infection of – and transmission between humans. Globalization, migrations and intensive tourism further increase the chance of rapid spread of such agents (Jones et al., 2008).

Since infectious disease outbreaks typically emerge unexpectedly and can advance swiftly, rapid detection of (re)-emerging pathogens is of utmost importance. Rapid detection allows for optimal preparation on the level of individuals (e.g., early recognition, quarantine and swift start of adequate treatment of individual patients), on the level of populations (e.g., fast vaccination and other preventive measures), and on the level of organizations (e.g., timely preparation and education of hospital personnel, adequate distribution of therapeutics and medical equipment, and preparation of research infrastructures), thereby hopefully limiting burden caused by each novel rapidly spreading disease.

The two most important challenges are timely recognition of infected individuals and sufficient and timely monitoring of global spread of an outbreak. The first step to optimal preparation may therefore be earlier recognition of infected individuals and global availability of data on spread of outbreaks. In this perspective, we describe a novel vision of how pandemics could be monitored in the future, using global omics-data. We will use influenza as an example as this has been an important causative infection in the past and is likely to cause successive pandemics in the near future.

THE STATE OF THE ART FOR INFLUENZA DIAGNOSIS AND TREATMENT

As conventional diagnostic methods such as viral culture and detection of antigens or antibodies have limitations due to low sensitivity and delay in time, the officious gold standard for laboratory diagnosis is detection of viral nucleic acids by reverse-transcriptase polymerase chain reaction (RT-PCR; George, 2012). Using this highly sensitive technique, minute amounts of virus can be detected and influenza virus subtypes can be differentiated in less than a few hours. The obvious disadvantages are that skilled personal is needed to perform the tests and that they may not be available during evenings, nights and weekend. This potentially causes delay in the diagnosis of an influenza infection in individual patients (Writing Committee of the WHO Consultation on Clinical Aspects of Pandemic (H1N1) 2009 Influenza et al., 2010).

Especially in case of severe illness, any delay is unwanted as it could hamper timely and life-saving measures and treatment (Writing Committee of the WHO Consultation on Clinical Aspects

of Pandemic (H1N1) 2009 Influenza et al., 2010; Ryoo et al., 2013). Of course, one could decide to quarantine and treat every critically ill patient who presents with influenza-like symptoms empirically (Writing Committee of the WHO Consultation on Clinical Aspects of Pandemic (H1N1) 2009 Influenza et al., 2010). Potential side-effects of treatment and high costs associated with quarantine, however, are arguments against such "unselected" treatment, especially on a large scale.

In the scenario of a pandemic with new emerging or re-emerging infections, every delay in adequate diagnosis halts precautionary measures to protect the uninfected individuals. These individuals could be vaccinated, if possible, which may (partially) protect them against the virus when being carried sufficiently ahead of time. Adequate vaccination may also limit further spread of the virus under specific circumstances, as vaccinated individuals not only stay healthy but also will not become contagious themselves. On the level of organization, timely detection allows for preparation and education of hospital personal and distribution of therapeutics and equipment to the desired location.

GLOBAL RECOGNITION OF FLU THROUGH INTEGRATION OF SIGNALS

Equally important to the time to diagnostic test results is the time to global availability of these data, especially if preparation at the population level and global organization for pandemia prevention is a goal. A system that obliges clinicians to report new cases of severe respiratory viral infections, called the public health response, is available for the clinical suspicion of specific pathogens but this requires a pro-active effort of doctors. Automated integration of test results through an online platform would allow for real-time surveillance. This approach is nicely illustrated by "Flu Trend" in Google™ (http://www.google.org/flutrends/). Using online search engine query data such as "influenza complication" and "flu remedy," Google™ is able to detect epidemics of respiratory viruses (Ginsberg et al., 2009). This method could predict the incidence of influenza-like illness 1 week before the Center of Disease Control (Ginsberg et al., 2009). However, as the input data for this model are not specific for influenza infection, this tool is helpful for influenza-like illness but probably not sufficient for monitoring the spread of a specific strain of the influenza virus. To capture this complexity, more specific viral signals should be investigated.

SYSTEMS BIOLOGY AND "OMICS" TECHNOLOGIES

Search engine queries rely on the phenotypic presentation of people with symptoms of influenza-like illnesses. Symptoms are non-specific results of physiological, cellular and molecular changes in the body that occur during viral infection. Systems biology aims to integrate the signals from all these levels into an understanding of the complete system (Josset et al., 2013). Following this philosophy, several "omics" technologies have been developed to measure the molecular landscape in an integrative fashion within one domain. "Genomics" can be used to study genetic risk factors for disease susceptibility of the host (Keynan et al., 2013) and for understanding of the pathogenicity of the pathogen in this context (Kash, 2009). Analysis of mRNA using "transcriptomics" potentially allows for simultaneous measurement of the expression of 10s of 1000s genes. Transcription research allows for rapid testing (e.g., in the order of hours) and provides more information on functionality than genomics alone. However, proteins are ultimately responsible for the function of cells. "Proteomics" has therefore the potential to uncover important interactions between the virus and the host. Studies show that influenza induces rapid changes in the host transcriptome and proteome (Liu et al., 2012; Pommerenke et al., 2012), as soon as 1 h after infection (Cheung et al., 2012). These very early temporal changes after infection are also observed at the metabolite level (Lin et al., 2012). "Metabolomics" is "the global assessment of endogenous metabolites within a biologic system and represents a snapshot-reading of gene function, enzyme activity and the physiological landscape" (Serkova et al., 2011). Treatment of influenza induces many metabolic changes that can be traced back to specific pathways (Lu et al., 2012).

"OMICS" TECHNOLOGIES FOR GLOBAL INFLUENZA SURVEILLANCE

The systems biology approach has the potential to increase understanding of the spread of a pandemic and the adaptations that viruses undergo meanwhile, as exemplified recently in outstanding research on the influenza virus reservoir in birds (Huang et al., 2013). The major problem with these technologies for monitoring is that they are very time-consuming and expensive, thus conclusions can only be drawn after the pandemic has ended, which is obviously too late. As such, they may only be sufficient for research on the pathogenesis of a pathogen responsible for the pandemic of interest. However, the unbiased approaches of systems biology can be used for unsupervised previsions about disease spreading if this information could be obtained rapidly and at the bedside.

FOCUS ON EXHALED BREATH

Exhaled breath of infected individuals contains aerosols filled with influenza viruses, mostly present in coarse particles (<5 μm; Milton et al., 2013). This in fact is an important route for the virus to spread. Breath also contains thousands of volatile organic compounds (VOCs), metabolites in gas-phase produced by both physiological and pathophysiological processes (Pauling et al., 1971; Moser et al., 2005). Pulmonary infection, inflammation and oxidative stress may alter the concentration of certain VOCs in exhaled breath (Bos et al., 2013a,b). VOC-patterns identified by smell have been used to diagnose disease and intoxication for ages (e.g., scent of acetone in uncontrolled diabetes; Manolis, 1983). Thus, influenza diagnosis based on exhaled breath analysis could take two forms: detection of aerosols with viral RNA, or an influenza-specific VOC-patterns.

So far, both these methods have relied on relatively time-consuming methods, RT-PCR and gas-chromatography coupled to mass-spectrometry, respectively. Rapid technological innovation in sensors, however, allows for detection of these signals using re-usable, rapid and easy nanosensor arrays. For example, a silicon nano-wire sensor device would allow influenza detection in half the time of RT-PCR in the clinical setting (Shen et al., 2012). Devices using sensor-based detection of VOCs are

called "electronic noses," following their apparent similarities to olfaction (Röck et al., 2008). Electronic noses integratively capture complex VOC mixtures using an array of different sensors (Röck et al., 2008). Sensors have individual sensitivities and specificities for multiple VOCs. The composite signal of all sensors can be analyzed using pattern-recognition algorithms. Electronic nose analysis of breath results in a unique fingerprints of exhaled metabolites, called "breath-prints." This allows rapid identification, recognition and comparison of VOC mixtures. Thereby, these breath-prints can be used for diagnostic and monitoring purposes, which do not require identification of individual molecular constituents. Breath-prints have found to be different in a wide range of respiratory diseases (Hockstein et al., 2005; Machado et al., 2005; Fens et al., 2009, 2010).

MONITORING OF MALODOR USING AN ELECTRONIC NOSE

Electronic nose technology is not sufficiently mature for widespread application in clinical practice. However, we can look at other applications to glance at the possibilities for global monitoring using this technology. In the port of Rotterdam, the Netherlands, with 10 km × 10 km one of the largest ports in the world, odor nuisance is a major problem. 30 metal oxide sensor based electronic noses were installed throughout the portal area, to monitor odor emissions in order to timely prevent nuisance (Bootsma and Milan, 2010; Milan et al., 2012). After a training period in which sensors were learned to recognize "malodor," the electronic noses were able to recognize more than 90% of the reported odor complaints in advance. Based on comparison between fingerprints of the recognized odor and an online database of previous events (an "odor-cloud," in line with the popular expression for an online virtual server application) the most probable chemical characteristics of the scent can be estimated (Apostulou, 2012). Combined with the temporal findings in different sensors and the direction and speed of the wind, the most probable source of pollution can be identified (**Figure 1**). This approach has allowed for prevention of the development of odor nuisance and environmental pollution by refineries, but also passing cargo ships.

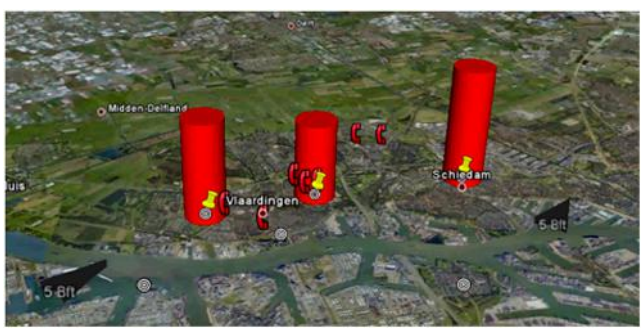


FIGURE 1 | Odor signatures in the harbor of Rotterdam. The odor signature disseminates in the direction of the wind leading to increased complaints of inhabitants (telephone symbols on the figure), with permission of Simon Bootsma.

MONITORING OF EXHALED BREATH USING ELECTRONIC NOSES

The parallels between monitoring of the spread of malodor in industrialized areas and of viral infections are striking. In both situations, the timing of the event is unknown, which requires continuous surveillance, and source identification is necessary for early intervention. In the case of exhaled breath analysis, the source cannot be identified with the direction of the wind, but by movement of hosts or patient populations. Therefore, we postulate that exhaled breath tests can best be positioned in places where large groups of people assemble for traveling. One could think of airports, train stations or border control. Here, a very sensitive test may identify infected individuals who are to contribute to the global spread of the pathogen. Importantly, the technique should be high-throughput and the result should be available instantaneously. Thus, in line with environmental surveillance, exhaled breath analysis of patients would allow for the construction of an online database with previously observed breath-prints (a "breath-cloud"; **Figure 2**). Linked with the clinical characteristics of these patients, an exhaled breath pattern for influenza infection can be identified and subsequently used for characterization of new patients. When the clinical information of these patients is known, the breath-cloud can be updated and identification can be improved, allowing for repeated, cyclic improvement of the diagnostic algorithm. There are important differences in the type and concentration of VOCs in environmental and breath analysis. In the environment, the molecules of interest are mostly present in parts-per-million concentration, in contrast to 10s to 100s parts-per-billion in breath. This means that sensors for breath monitoring need to be more sensitive. The VOCs of interest are mostly sulfur-containing and cyclic compounds in studies on maladour but breath research is not limited to those. Therefore it is anticipated that the sensor array ought to be larger and more versatile in breath analysis.

REQUIRED STEPS FOR INTEGRATING THE METHODOLOGY

Several steps are needed to accomplish the above-suggested approach:

- The VOCs that can be used for early diagnosis of a viral infection need to be identified. A very sensitive diagnosis is a first requirement for global screening as the goal is to isolate a small potion of the population while maintaining a very high negative predictive value.
- An appropriate array of sensors needs to be assembled. These arrays need to contain several sensors that are designed to react selectively with the previously identified VOCs and a wide variety of general, semi-selective (cross-reactive) sensors (Konvalina and Haick, 2013). The former are used for specific identification of a viral infection but are prone to changes in virus induced exhaled breath profile that may occur over time, due to viral mutations or phenotypic changes. The latter allows for plasticity of the diagnostic algorithm by reacting with unselected VOCs.
- An online centralized database, were to breath-prints are uploaded, instantly should be created. Thus, the electronic

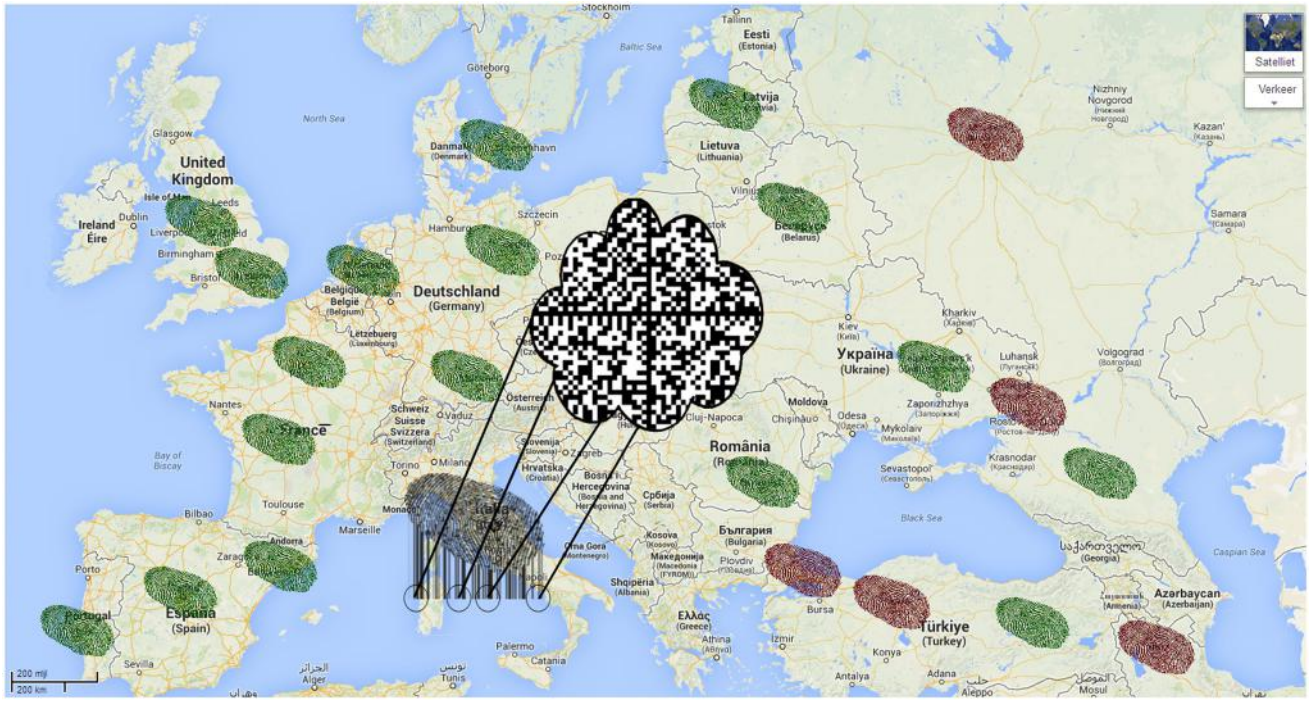


FIGURE 2 | Future perspective of integration of global omics signals. There is a centralized database containing molecular fingerprints of past cases, the "breath-cloud" (middle of the picture). Green fingerprints represent suspected cases that were found to be most similar to the non-infected profile. Red fingerprints represent cases that were most similar to the infected profile. There is a new suspected case in Italy, the molecular signature is now compared to the profiles in the breath-cloud.

noses should be connected to the Internet, which allows for synchronization with the breath-cloud.

- The electronic noses should be readily available in the areas where the first signals of an outbreak are visible.

OTHER TECHNOLOGIES

Exhaled breath analysis by electronic nose may not be the only technology that is continuously available and provides a direct test result. Any technology that gives a rapid result can be used for the same purpose. All signals can be used complementary by uploading them together to the same online database. Here a pattern recognition algorithm can treat these signals similar and will select only the most discriminative, updating prior beliefs with every additional information that becomes available. The development of bedside PCR machines is exciting in this respect as these could allow for very accurate and rapid detection of viral RNA (Centers for Disease Control and Prevention, 2013). Further miniaturization and optimization toward a "lab on a chip" will further improve the time to result and bedside use (Sun et al., 2011; Song et al., 2012).

CONCLUSION

To conclude, there is a need for rapid diagnosis of specific infections (including but not restricted to outbreaks of influenza), especially during a pandemic. High-throughput chemical profiling of patient material could provide a fast and objective means to diagnose patients. At this moment, portable electronic nose technology is a good example of how these infections can be captured rapidly at the bedside and can be shared with the world in the form of a "breath-cloud." In principle, any technology that provides test results rapidly could attribute to this online database. Pattern recognition software can subsequently be used to diagnose new suspected cases based on previous profiles from patients all over the world.

REFERENCES

Apostulou, M. (2012). *A Toolkit for Training Odour Monitoring Systems.* Msc thesis, Institute for Informatics, University of Amsterdam, Amsterdam.

Bootsma, S., and Milan, B. (2010). Odour monitoring with E-noses in the Port of Rotterdam. *Chem. Eng. Trans.* 23, 147–152. doi: 10.3303/CET1023025

Bos, L. D. J., Sterk, P. J., and Schultz, M. J. (2013a). Volatile metabolites of pathogens: a systematic review. *PLoS Pathog.* 9:e1003311. doi: 10.1371/journal.ppat.1003311

Bos, L. D. J., van Walree, I. C., Kolk, A. H. J., Janssen, H.-G., Sterk, P. J., and Schultz, M. J. (2013b). Alterations of exhaled breath metabolite-mixtures in two rat models of lipopolysaccharide-induced lung injury. *J. Appl. Physiol.* 115, 1487–1495. doi: 10.1152/japplphysiol.00685.2013

Centers for Disease Control and Prevention. (2013). *Rapid Diagnostic Testing for Influenza.* Available at: http://www.cdc.gov/flu/professionals/diagnosis/rapidlab.htm#table2

Cheung, C. Y., Chan, E. Y., Krasnoselsky, A., Purdy, D., Navare, A. T., Bryan, J. T., et al. (2012). H5N1 virus causes significant perturbations in host proteome very early in influenza virus-infected primary human monocyte-derived macrophages. *J. Infect. Dis.* 206, 640–645. doi: 10.1093/infdis/jis423

Fens, N., Douma, R. A., Sterk, P. J., and Kamphuisen, P. W. (2010). Breathomics as a diagnostic tool for pulmonary embolism. *J. Thromb. Haemost.* 8, 2831–2833. doi: 10.1111/j.1538-7836.2010.04064.x

Fens, N., Zwinderman, A. H., van der Schee, M. P., de Nijs, S. B., Dijkers, E., Roldaan, A. C., et al. (2009). Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. *Am. J. Respir. Crit. Care Med.* 180, 1076–1082. doi: 10.1164/rccm.200906-0939OC

George, K. S. (2012). Diagnosis of influenza virus. *Methods Mol. Biol.* 865, 53–69. doi: 10.1007/978-1-61779-621-0_4

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014. doi: 10.1038/nature07634

Hockstein, N. G., Thaler, E. R., Lin, Y., Lee, D. D., and Hanson, C. W. (2005). Correlation of pneumonia score with electronic nose signature: a prospective study. *Ann. Otol. Rhinol. Laryngol.* 114, 504–508.

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian, W., et al. (2013). The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.* 45, 776–783. doi: 10.1038/ng.2657

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., et al. (2008). Global trends in emerging infectious diseases. *Nature* 451, 990–993. doi: 10.1038/nature06536

Josset, L., Tisoncik-Go, J., and Katze, M. G. (2013). Moving H5N1 studies into the era of systems biology. *Virus Res.* 178, 151–167. doi: 10.1016/j.virusres.2013.02.011

Kash, J. C. (2009). Applications of high-throughput genomics to antiviral research: evasion of antiviral responses and activation of inflammation during fulminant RNA virus infection. *Antiviral Res.* 83, 10–20. doi: 10.1016/j.antiviral.2009.04.004

Keynan, Y., Malik, Y., and Fowke, K. R. (2013). The role of polymorphisms in host immune genes in determining the severity of respiratory illness caused by pandemic H1N1 influenza. *Public Health Genomics* 16, 9–16. doi: 10.1159/000345937

Konvalina, G., and Haick, H. (2013). Sensors for breath testing: from nanomaterials to comprehensive disease detection. *Acc. Chem. Res.* 47, 66–76. doi: 10.1021/ar400070m

Lin, S., Liu, N., Yang, Z., Song, W., Wang, P., Chen, H., et al. (2012). GC/MS-based metabolomics reveals fatty acid biosynthesis and cholesterol metabolism in cell lines infected with influenza A virus. *Talanta* 83, 262–268. doi: 10.1016/j.talanta.2010.09.019

Liu, L., Zhou, J., Wang, Y., Mason, R. J., Funk, C. J., and Du, Y. (2012). Proteome alterations in primary human alveolar macrophages in response to influenza A virus infection. *J. Proteome Res.* 11, 4091–4101. doi: 10.1021/pr3001332

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., and Murray, C. J. L. (2006). *Global Burden of Disease and Risk Factors.* Washington: The World Bank and Oxford University Press. doi: 10.1596/978-0-8213-6262-4

Lu, C., Jiang, Z., Fan, X., Liao, G., Li, S., He, C., et al. (2012). A metabonomic approach to the effect evaluation of treatment in patients infected with influenza A (H1N1). *Talanta* 100, 51–56. doi: 10.1016/j.talanta.2012.07.076

Machado, R. F., Laskowski, D., Deffenderfer, O., Burch, T., Zheng, S., Mazzone, P. J., et al. (2005). Detection of lung cancer by sensor array analyses of exhaled breath. *Am. J. Respir. Crit. Care Med.* 171, 1286–1291. doi: 10.1164/rccm.200409-1184OC

Manolis, A. (1983). The diagnostic potential of breath analysis. *Clin. Chem.* 29, 5–15.

Milan, B., Bootsma, S., and Bilsen, I. (2012). Advances in odour monitoring with E-Noses in the Port of Rotterdam. *Chem. Eng. Trans.* 30, 145–150. doi: 10.3303/CET1230025

Milton, D. K., Fabian, M. P., Cowling, B. J., Grantham, M. L., and McDevitt, J. J. (2013). Influenza virus aerosols in human exhaled breath: particle size, culturability, and effect of surgical masks. *PLoS Pathog.* 9:e1003205. doi: 10.1371/journal.ppat.1003205

Moser, B., Bodrogi, F., Eibl, G., Lechner, M., Rieder, J., and Lirk, P. (2005). Mass spectrometric profile of exhaled breath – field study by PTR-MS. *Respir. Physiol. Neurobiol.* 145, 295–300. doi: 10.1016/j.resp.2004.02.002

Pauling, L., Robinson, A. B., Teranishi, R., and Cary, P. (1971). Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc. Natl. Acad. Sci. U.S.A.* 68, 2374–2376. doi: 10.1073/pnas.68.10.2374

Pommerenke, C., Wilk, E., Srivastava, B., Schulze, A., Novoselova, N., Geffers, R., et al. (2012). Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses. *PLoS ONE* 7:e41169. doi: 10.1371/journal.pone.0041169

Röck, F., Barsan, N., and Weimar, U. (2008). Electronic nose: current status and future trends. *Chem. Rev.* 108, 705–725. doi: 10.1021/cr068121q

Ryoo, S. M., Kim, W. Y., Sohn, C. H., Seo, D. W., Oh, B. J., Lee, J. H., et al. (2013). Factors promoting the prolonged shedding of the pandemic (H1N1) 2009 influenza virus in patients treated with oseltamivir for 5 days. *Influenza Other Respir. Viruses* 7, 833–837. doi: 10.1111/irv.12065

Serkova, N. J., Standiford, T. J., and Stringer, K. A. (2011). The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. *Am. J. Respir. Crit. Care Med.* 184, 647–655. doi: 10.1164/rccm.201103-0474CI

Shen, F., Wang, J., Xu, Z., Wu, Y., Chen, Q., Li, X., et al. (2012). Rapid flu diagnosis using silicon nanowire sensor. *Nano Lett.* 12, 3722–3730. doi: 10.1021/nl301516z

Song, H.-O., Kim, J.-H., Ryu, H.-S., Lee, D.-H., Kim, S.-J., Kim, D.-J., et al. (2012). Polymeric LabChip real-time PCR as a point-of-care-potential diagnostic tool for rapid detection of influenza A/H1N1 virus in human clinical specimens. *PLoS ONE* 7:e53325. doi: 10.1371/journal.pone.0053325

Sun, Y., Dhumpa, R., Bang, D. D., Hogberg, J., Handberg, K., and Wolff, A. (2011). A lab-on-a-chip device for rapid identification of avian influenza viral RNA by solid-phase PCR. *Lab Chip* 11, 1457–1463. doi: 10.1039/c0lc00528b

Taubenberger, J. K., and Morens, D. M. (2006). 1918 influenza: the mother of all pandemics. *Emerg. Infect. Dis.* 12, 15–22. doi: 10.3201/eid1201.050979

Writing Committee of the WHO Consultation on Clinical Aspects of Pandemic (H1N1) 2009 Influenza, Bautista, E., Chotpitayasunondh, T., Gao, Z., Harper, S. A., Shaw, M., et al. (2010). Clinical aspects of pandemic 2009 influenza a (H1N1) virus infection. *N. Engl. J. Med.* 362, 1708–1719. doi: 10.1056/NEJMra1000449

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: 27 November 2013; accepted: 25 March 2014; published online: 22 April 2014.

*Citation: Bos LDJ, de Jong MD, Sterk PJ and Schultz MJ (2014) How integration of global omics-data could help preparing for pandemics – a scent of influenza. Front. Genet. **5**:80. doi: 10.3389/fgene.2014.00080*

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

Copyright © 2014 Bos, de Jong, Sterk and Schultz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
GENETICS

MINI REVIEW ARTICLE
published: 14 May 2014
doi: 10.3389/fgene.2014.00132

Non-coding RNAs in pluripotency and neural differentiation of human pluripotent stem cells

Dunja Lukovic[1], Victoria Moreno-Manzano[2], Martin Klabusay[3], Miodrag Stojkovic[4,5], Shomi S. Bhattacharya[1] and Slaven Erceg[1]*

[1] *Retina Group, Cell therapy and Regenerative Medicine, Centro Andaluz de Biología Molecular y Medicina Regenerativa, Sevilla, Spain*
[2] *Neuronal and Tissue Regeneration Lab, Centro de Investigación Príncipe Felipe, Valencia, Spain*
[3] *Integrated Center of Cellular Therapy and Regenerative Medicine – International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic*
[4] *Spebo Medical, Leskovac, Serbia*
[5] *Human Genetics Department, Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia*

Edited by:
Enrico Capobianco, University of Miami, USA

Reviewed by:
Sudipto Saha, Bose Institute, India
Noriko Hiroi, Keio University, Japan

****Correspondence:***
Slaven Erceg, Retina Group, Cell therapy and Regenerative Medicine, Centro Andaluz de Biología Molecular y Medicina Regenerativa, Avenida Americo Vespucio s/n, Parque Científico y Tecnológico Cartuja, Isla de la Cartuja, Sevilla, Spain
e-mail: slaven.erceg@cabimer.es

Several studies have demonstrated the important role of non-coding RNAs as regulators of posttranscriptional processes, including stem cells self-renewal and neural differentiation. Human embryonic stem cells (hESCs) and induced pluripotent stem cells (ihPSCs) show enormous potential in regenerative medicine due to their capacity to differentiate to virtually any type of cells of human body. Deciphering the role of non-coding RNAs in pluripotency, self-renewal and neural differentiation will reveal new molecular mechanisms involved in induction and maintenances of pluripotent state as well as triggering these cells toward clinically relevant cells for transplantation. In this brief review we will summarize recently published studies which reveal the role of non-coding RNAs in pluripotency and neural differentiation of hESCs and ihPSC.

Keywords: pluripotent stem cells, pluripotency, non-coding RNA, differentiation, human embryonic stem cells

INTRODUCTION

Personalized medicine is expected to benefit from the combination of genomic information with the high throughput studies including transcriptomic, proteomic and metabolomic profiling. Measuring gene expression in individual cells is crucial for understanding the gene regulatory network. In order to decipher the genetic regulatory network in cells significant efforts have been made over the years to develop technology platforms for transcriptome characterization such as DNA microarray hybridization, serial analysis of gene expression (SAGE; Velculescu et al., 1995) or next-generation RNA sequencing often called RNA-seq (Mortazavi et al., 2008).

The latest techniques which involve bioinformatic expertise made a revolution in transcriptome analysis enabling not only the identification of cDNA and gene isoforms but discovery of long non-coding RNA (large intergenic non-coding RNA, lincRNA; >200 nucleotides in length) and short non-coding RNA (sncRNA, <200 nucleotides in length). Non-coding RNAs include transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear and small nucleolar RNA, microRNA (miRNA), and small interfering RNA (siRNA), which do not encode any proteins. Several of these non-coding RNA species, like miRNA or SiRNAs, are of particular interest to transcriptomic and particularly in stem cell research due to their role in post-transcriptional regulation of numerous biological processes (Morozova and Marra, 2008; Roukos, 2010). During the last several years many studies were published in order to determine the function of these non-coding transcripts including novel miRNA (Hafner et al., 2008) that exhibit different cell-type and tissue specificity (Guttman and Rinn, 2012). Although the functions of the majority of newly discovered non-coding RNAs are still unknown, some were found to play important roles in the regulation of stem cells. Recent studies concentrate on miRNAs (Wilson et al., 2009; Kim et al., 2011; Lipchina et al., 2011). In the context of stem cell biology, of particular interest is the role of these RNAs in expression of renewal genes in human embryonic stem cells (hESCs) or in regulation of induced pluripotency (Li et al., 2011). In this review, we focus on recent discoveries of non-coding RNA roles in human pluripotent stem cell biology and differentiation.

HUMAN EMBRYONIC STEM CELLS AND INDUCED PLURIPOTENT STEM CELLS

Human pluripotent stem cells encompassing hESCs and induced pluripotent stem cells (ihPSCs) show great potential for regenerative biology providing the unique human *in vitro* platforms for studying diseases, basic cell biology and develop ment.

Human embryonic stem cells can be derived from inner mass from human blastocyst maintaining unique capacity for unlimited self-renewal through long-term maintenance using laboratory culture conditions (Thomson et al., 1998). Since the generation of the first hESCs line in 1998 (Thomson et al., 1998), research in this area has progressed at a rapid pace, developing efficient protocols globally for differentiation of these cells to clinically relevant cell types (Erceg et al., 2008, 2009, 2010, 2012). hESCs represent a useful model for studying early human embryology and cell differentiation and have limited capacity for disease modeling in

human cells (Biancotti et al., 2010). hESCs bear the advantage over any other stem cells in that they are pluripotent, providing an unlimited starting cell source for differentiation to any type of tissue of the human body. The perspective of clinical use of these cells and their derivates is huge. The hESCs-based therapy is increasingly recognized as a promising strategy for degenerative disorders entering already in clinic to treat spinal cord injury or recently published encouraging results in human clinical trial investigating their use in age-related macular degeneration (Schwartz et al., 2012). The main disadvantage of use of hESCs in regenerative medicine is the fact that derivation of hESCs requires the destruction of human embryos which generates the ethical concerns.

Besides the abundance and efficient differentiation without traces of pluripotency, the main requisite for personalized regenerative medicine is to derive disease cells that genetically match the patient. Although the technique of somatic cell nuclear transfer (SCNT) and successive derivation of hESCs (Tachibana et al., 2013) could be a promising approach in the future to create patient specific cells, major technical and ethical obstacles related with this technique are present.

The discovery of human ihPSCs originally generated by ectopic expression of four transcription factors Oct4, Sox2, Klf4, and cMyc (Takahashi et al., 2007) in human fibroblast cells presents a novel tool to obtain disease cells. This Nobel Prize winner technology was substantially improved by introducing non-integrative transgene expression (Jin et al., 2012) and targeting different somatic tissues. Patient-specific ihPSCs derived from somatic cells are devoid of immnunological and ethical concerns, allow the generation of disease-specific stem cells providing a platform to study molecular mechanisms of genetic diseases. The ihPSCs show morphological, transcriptional, epigenetic, and phenotypic similarity to hESCs and can differentiate toward any cell of human body. Until now a number of studies has shown that ihPSCs can be successively generated from patients carrying different diseases and be a faithful platform for disease modeling *in vitro* (Gunaseeli et al., 2010; Hargus et al., 2010; Jin et al., 2011, 2012; Pedrosa et al., 2011; Kumano et al., 2012; Oh et al., 2012; Sun et al., 2012; Cocks et al., 2013; Gross et al., 2013; Tubsuwan et al., 2013).

Pluripotent stem cells possess two major characteristics: self-renewal and differentiation into other cell types. The investigators put the major effort in development of new protocols and moving these cells to clinics but it is crucial to understand these two main characteristics in order to enter deeply in basic biology of these cells. For example it is still to be elucidated reprogramming mechanisms in target cells and why only small population of cells becomes fully reprogrammed. In order to decipher molecular mechanisms of reprogramming the role of RNA and related global gene expression changes is of particular interest in order to increase reproducibility and efficiency of reprogramming processes. Reproducible generation of specific cellular type without traces of ihPSCs is one of the crucial issues in order to prevent teratoma generation in host. Improvements of the differentiation protocols are required as a basis for further cost-efficient industrial processes of large-scale for future application in clinics. To reach this also extensive characterization of differentiated cell has to be performed and subsequently compared with undifferentiated counterparts. Comparative transcriptome analyses using microarray also indicate that hESCs and hiPSCs have similar, highly alike gene expression patterns. Gene expression pattern of ihPSCs is separate from the originating somatic cells with possibility of retaining some transcriptional differences or an epigenetic memory of the starting cells (Plath and Lowry, 2011). Transcriptome characterization would undoubtedly provide insights into the genetic regulatory networks involved in maintaining pluripotency and directing differentiation. In order to define molecularly the various phases of the reprogramming process, as well as full pluripotent stem cells state global gene expression and proteomic patterns of clonal cell populations or enriched populations need to be performed in different stages after inicial reprogramming induction.

PLURIPOTENCY

Generally, a definition of pluripotency is related to ability of cell to give rise three germ layers: endoderm, ectoderm, and mesoderm and their derivates. This ability has only a small number of cells such as hESCs and ihPSCs and their maintenance involves core transcription factors: Oct4, Sox2, and Nanog (Boyer et al., 2005; Kim et al., 2009). A spectrum of different miRNA was detected in embryonic stem cell as pluripotency-specific markers which expression was downregulated during the induction of differentiation (**Table 1**; Wilson et al., 2009; Lee et al., 2010). A family of miRNA that includes AAGUGC seed sequence is of particular interest in pluripotent stem cells for its high expression in hESCs and ihPSC. The most abundant miRNA transcript in hESCs is *mir-302* which encodes for miR-302a/b/c/d and mir-367 (Suh et al., 2004) and is under the control of Oct4, Sox2, and Nanog. This miRNA is involved in maintenance of pluripotency, self-renewal, regulation of cell cycle, and fate specification during differentiation of hESCs (Suh et al., 2004; Landgraf et al., 2007; Bar et al., 2008; Lipchina et al., 2011) probably inhibiting neural differentiation by modulation of BMP signaling targeting its inhibitors: TOB2, DAZAP2, and SLAIN1 (Lipchina et al., 2011). Rosa and Brivanlou (2011) have shown that Oct4 and miR-302 inhibit NR2F2, which in turn inhibits Oct4. The expression of gene *NR2F2* is increased during differentiation when the expression of *OCT4* gene and miR-302 declines (Rosa and Brivanlou, 2011). This study showed important biological function of mir-302 and NR2F2 in human early development and cell fate determination. It seems that other miRNAs such as miR-145 has the opposite role in maintenance of pluripotency (Xu et al., 2009). The expression of this miRNA is low in undifferentiated hESCs but its increased expression is related to inhibition of hESCs self-renewal and induction of lineage-restricted differentiation (Xu et al., 2009).

Elucidation of the precise molecular and cellular mechanisms which convert human fibroblasts or other somatic cells to ihPSCs was the main challenge among the investigators during the last years. Reprogramming somatic cells into pluripotent cellular identity requires tightly regulated and coordinated changes in expression of many genes. Understanding the genetic network involved in cellular reprogramming is crucial to elucidate pluripotency in order to increase the reprogramming efficiency and cell renewal. These mechanisms will reveal why only small

Table 1 | Different roles of non-coding RNA in pluripotency and neural differentiation.

Type of cells	Processes involved	Non-coding RNA	Reference
hESC	Pluripotency, self-renewal, cell cycle and fate specification	miR-302	Suh et al. (2004), Bar et al. (2008), Lipchina et al. (2011)
hESC	Inhibition of pluripotency	miR-145	Xu et al. (2009)
iPSC	Pluripotency	miR-17, miR-106b, and miR-106a	Li et al. (2011)
Fibroblasts to iPSC	Reprogramming	miR-302, miR-372	Anokye-Danso et al. (2011, 2012), Subramanyam et al. (2011)
Fibroblasts to iPSC	Reprogramming	Combination of miR-302, miR-200c, and miR-369	Miyoshi et al. (2011)
iPSC	Reprogramming	LincRNAs	Loewer et al. (2010)
hESC	Neural differentiation	LincRNAs	Ng et al. (2012)
iPS-derived neural progenitors	Neural differentiation	LincRNAs	Lin et al. (2011)
hESC	Differentiation to neuroectoderm	miR-200, miR-96	Du et al. (2013)
hESC-derived neural stem cells	Suppression of selfrenewal, neural differentiation	miR-124, miR-125b and miR-9/9	Roese-Koerner et al. (2013)
hESC	Neural differentiation	miR7	Liu et al. (2012)
hESC	Neural differentiation	miR125	Boissart et al. (2012)

hESC, human embryonic stem cells; iPSC, induced pluripotent stem cells.

population of somatic cells undergo full reprogramming. Different gene expression patterns and post-transcriptional events, including mRNA decay, between pluripotent and differentiated cells could reveal the reprogramming mechanisms of the fibroblasts into ihPSCs. The study of Buganim et al. (2012) showed that reprogramming involves stochastic gene expression in early phase followed by a late hierarchical phase with activation of *SOX2* gene, which then triggers a stepwise gene activation that allows the cells to enter the pluripotent state. SOX2 represents a group of pluripotency initiating factors (PIFs) indispensable for endogenous activation of *OCT4*, *SOX2*, and *NANOG* (Boyer et al., 2005) which further maintain the ihPSCs state. Some of these genes maintain pluripotency by blocking the gene machinery involved in differentiation.

In the study of Li et al. (2011) was observed that three miRNA clusters: miR-17, miR-106b, and miR-106a were significantly upregulated that interfere with iRNA machinery directly connected with important reprogramming pathways: TGF-β signaling and cell cycle. These results suggest that transcription factors that modulate miRNA decay could have crucial role in reprogramming differentiated cells or in maintaining pluripotency, but future studies have to be performed to confirm whether these factors can be efficient target to induce or maintain the pluripotency or trigger the differentiation.

Several miRNA, especially miR-302 and miR-372 have been directly involved in enhancing of HFF reprogramming (Subramanyam et al., 2011) revealing the possibility to directly target these miRNAs to reprogram the HFF without Yamanaka factors. The recent study of Morrisey and colleague (Anokye-Danso et al., 2011, 2012), confirmed that reprogramming can be achieved by using miRNAs without protein-coding factors. Another study confirmed that fully pluripotent stem cells can be obtained by introducing other miRNA such as combination of miR-302, miR-200c, and miR-369 (Miyoshi et al., 2011). Different studies speculated about the mechanisms and signaling pathways by which these miRNAs exert their reprogramming function such as regulation of different genes involved in cell cycle, epithelial-mesenchymal transition, epigenetic regulation and vesicular transport (Subramanyam et al., 2011).

On the other hand, the abundance of lincRNAs in mammalian transcriptome reveals their role as key regulators of biological processes. These RNA transcripts have little or no protein coding potential but some studies point out their possible participation in pluripotency, differentiation and self-renewal (Guttman et al., 2009, 2011; Sheik Mohamed et al., 2010; Guttman and Rinn, 2012). Several studies have recently discovered a novel class of lincRNAs possible involved in reprogramming processes, pluripotency and lineage commitment (Boyer et al., 2006; Lee et al., 2006; Loewer et al., 2010).

Some of these lincRNAs act directly as regulators of reprogramming (RoR) called lincRNA-RoR (Loewer et al., 2010). Overexpression of these RNAs significantly enhances the reprogramming efficiency and their downregulation decreases the generation of ihPSC colonies possibly by mechanism of negative regulation of p53 (Zhang et al., 2013).

These studies indicate that non-coding RNAs, especially miRNAs have the potential to be used as small-molecule therapeutics to promote more efficient reprogramming or to induce the pluripotent stem cells toward other cell lineages.

DIFFERENTIATION

In the context of regenerative medicine it is crucial to develop protocols for efficient and reproducible differentiation of pluripotent stem cells toward homogeneous population of desired cells without traces of pluripotency. Since the generation of the first hESCs line (Thomson et al., 1998) and derivation of ihPSC (Takahashi et al., 2007), research in this area has progressed at a rapid pace, developing efficient protocols globally for differentiation of these cells to clinically relevant cell types. As already mentioned, hESCs and ihPSCs bear the advantage over any other stem cells in that they are pluripotent, providing an unlimited starting cell source for differentiation to any type of tissue of the human body. Understanding the regulatory mechanisms which orchestrate the hESCs and ihPSCs during differentiation is of enormous importance because coordinated changes in gene expression during the differentiation of hESC and ihPSC are crucial for lineage specification. Beside the gene expression changes in coding RNA it is a clear to investigate whether non-coding RNA play important role in early differentiation of pluripotent stem cells. Although recent studies have shown that ihPSCs lines exert better differentiation capacity when compared with hESCs (Hu et al., 2010) direct comparison of differentiated cells versus undifferentiated counterparts is crucial in order to find signaling mechanisms involved in differentiation. In the recent study Gifford et al. (2013) performed comprehensive transcriptional profiling of cell populations generated by directed differentiation of hESCs.

To reveal whether lincRNAs play important role in hESCs and neural differentiation Stanton and colleague (Ng et al., 2012), employed a highly efficient protocol for neural differentiation of hESCs based on stromal-derived induction activity (SDIA) using co-culture of hESCs with PA6 mouse stromal cells. This procedure, used by many groups, was designed to generate homogeneous population of neural progenitor cells and further dopaminergic neurons (Kawasaki et al., 2000, 2002; Zeng et al., 2004). About 36 lincRNAs were identified which were associated with pluripotency making the complex with SOX2, and SUZ12, well known genes involved in pluripotency. Association of newly discovered lincRNAs with MIR-125B and LET7A reveal important role of these lincRNA in neurogenesis and neural differentiation. These results demonstrate that lincRNAs represents indispensable components in regulation of biological processes such as neural differentiation and pluripotency.

In order to clarify the contribution of lincRNA in developmental and neurological disorders, Lin et al. (2011) were performed Genome-wide analysis using next-generation sequencing (RNA-Seq) of neural progenitors derived from ihPSCs. They found that early differentiated cells underwent dramatic quantitative changes in gene expression especially lincRNAs. The authors associated many lincRNAs with *HOX* gene (*HOXA* and *HOXB*), genes involved in early patterning of anterior posterior axis during the neural development. These results coincided with results obtained with neural progenitors derived from hESCs as an additional prove that these two sources of pluripotent stem cells has similar neuronal differentiation potential (Wu et al., 2010). The author's general aim in this article is to associate the obtained results with some neuropsychiatric disorders in order to establish faithful lincRNA markers. The RNA-Seq findings highlighted possible non-coding RNA variants as feasible candidates which mutations are involved in many neuropsychiatric disorders mostly schizophrenia, bipolar disorders and autism spectrum disorders. These transcription factors and chromatin modifiers candidate are: *POU3F2*, *MYTIL*, *RFX4*, *ZNF804A*, *SMARCA2,* and *NPAS3*. These changes in the transcriptome profiles and the role of lincRNA during early human neural differentiation using pluripotent stem cells reveals important use of ihPSC technology in studying human disease as a unique human assay of human neurogenesis. Integration the novel transcripts in more global systems of analysis is must in order to elucidate their abnormally regulation in a subgroup of patients.

Comparing the miRNA profiles of neuroectodermal cells to epidermal cells both derived from hESC, Zhang and colleague (Du et al., 2013) identified the downreglation of two miRNA families in neuroectodermal differentiated cells, miR-200 and miR-96. Investigating the function of these miRNA it was discovered that miR-200 regulates the level of zinc-finger E-box-binding homeobox (ZEB), transcription factor family involved in inhibition of expression of BMP and its downstream genes, thus promoting neural differentiation (Postigo et al., 2003), while miR-96 regulates PAX6 (paired box 6), well known transcription factor characteristic for neuroectoderm. The authors also find that upregulation of these miRNA suppresses differentiation of hESCs toward neural lineage (Du et al., 2013). Recent article examined the role of the neural-associated miR-124, miR-125b, and miR-9/9 in human neural stem cells derived from human pluripotent stem cells (Roese-Koerner et al., 2013) and showed that overexpression of these miRNA suppress self-renewal and induce further differentiation into neurons. Providing additional evidence of involvement of other miRNA such as miR7 (Liu et al., 2012) and miR125 (Boissart et al., 2012) in neural differentiation of hESCs, these studies showed that neural stem cells derived from pluripotent stem cells could be a faithful model for investigation of role of miRNA in modulating of stemness and neuronal differentiation capacity of these cells.

CONCLUSION

Studying of non-coding RNA in modeling exhaustive networks of gene interactions as an ultimate application of systems biology in systems biomedicine, could substantially contribute to understanding and modulation of developmental and differentiation processes in humans. Although the expression of newly correlated non-coding RNA is strongly associated to pluripotency and neural differentiation their possible role in different neurodegenerative disorders is still to be elucidated. These studies undoubtedly contribute to better understanding of the biological processes during pluripotency and neural differentiation and reveal the important interplay between multiple pluripotency transcription factors and non-coding RNAs especially miRNAs. However, the understanding of the impact of

miRNA-based regulation in human neural development is still at its dawn. The future studies will confirm the potential of controlling differentiation and pluripotency of human pluripotent stem cells by modulating the expression of selected non-coding RNAs and integrate them into models that reveal the global behavior of the biological process in biomedicine and neural diseases in order to ultimately improve patients' quality of life.

ACKNOWLEDGMENTS

This work was supported by funds for research from "Miguel Servet" contract of Instituto de Salud Carlos III of Spanish Ministry of Science and Innovation (Slaven Erceg), Fund for Health of Spain PI10-01683 (Victoria Moreno-Manzano), Junta de Andalucía PI-0113-2010 (Slaven Erceg) and Supported by European Regional Development Fund – Project FNUSA-ICRC (No. CZ.1.05/1.1.00/02.0123).

REFERENCES

Anokye-Danso, F., Snitow, M., and Morrisey, E. E. (2012). How microRNAs facilitate reprogramming to pluripotency. *J. Cell Sci.* 125, 4179–4187. doi: 10.1242/jcs.095968

Anokye-Danso, F., Trivedi, C. M., Juhr, D., Gupta, M., Cui, Z., Tian, Y., et al. (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell* 8, 376–388. doi: 10.1016/j.stem.2011.03.001

Bar, M., Wyman, S. K., Fritz, B. R., Qi, J., Garg, K. S., Parkin, R. K., et al. (2008). MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells* 26, 2496–2505. doi: 10.1634/stemcells.2008-0356

Biancotti, J. C., Narwani, K., Buehler, N., Mandefro, B., Golan-Lev, T., Yanuka, O., et al. (2010). Human embryonic stem cells as models for aneuploid chromosomal syndromes. *Stem Cells* 28, 1530–1540. doi: 10.1002/stem.483

Boissart, C., Nissan, X., Giraud-Triboult, K., Peschanski, M., and Benchoua, A. (2012). miR-125 potentiates early neural specification of human embryonic stem cells. *Development* 139, 1247–1257. doi: 10.1242/dev.073627

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956. doi: 10.1016/j.cell.2005.08.020

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441, 349–353. doi: 10.1038/nature04733

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., et al. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* 150, 1209–1222. doi: 10.1016/j.cell.2012.08.023

Cocks, G., Curran, S., Gami, P., Uwanogho, D., Jeffries, A. R., Kathuria, A., et al. (2013). The utility of patient specific induced pluripotent stem cells for the modelling of autistic spectrum disorders. *Psychopharmacology (Berl.)* 231, 1079–1088. doi: 10.1007/s00213-013-3196-4

Du, Z. W., Ma, L. X., Phillips, C., and Zhang, S. C. (2013). miR-200 and miR-96 families repress neural induction from human embryonic stem cells. *Development* 140, 2611–2618. doi: 10.1242/dev.092809

Erceg, S., Lainez, S., Ronaghi, M., Stojkovic, P., Perez-Arago, M. A., Moreno-Manzano, V., et al. (2008). Differentiation of human embryonic stem cells to regional specific neural precursors in chemically defined medium conditions. *PLoS ONE* 3:e2122. doi: 10.1371/journal.pone.0002122

Erceg, S., Lukovic, D., Moreno-Manzano, V., Stojkovic, M., and Bhattacharya, S. S. (2012). Derivation of cerebellar neurons from human pluripotent stem cells. *Curr. Protoc. Stem Cell Biol.* Chap. 1, Unit 1H.5. doi: 10.1002/9780470151808.sc01h05s20

Erceg, S., Ronaghi, M., Oria, M., Rosello, M. G., Arago, M. A., Lopez, M. G., et al. (2010). Transplanted oligodendrocytes and motoneuron progenitors generated from human embryonic stem cells promote locomotor recovery after spinal cord transection. *Stem Cells* 28, 1541–1549. doi: 10.1002/stem.489

Erceg, S., Ronaghi, M., and Stojkovic, M. (2009). Human embryonic stem cell differentiation toward regional specific neural precursors. *Stem Cells* 27, 78–87. doi: 10.1634/stemcells.2008-0543

Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., et al. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* 153, 1149–1163. doi: 10.1016/j.cell.2013.04.037

Gross, B., Sgodda, M., Rasche, M., Schambach, A., Gohring, G., Schlegelberger, B., et al. (2013). Improved generation of patient-specific induced pluripotent stem cells using a chemically-defined and matrigel-based approach. *Curr. Mol. Med.* 13, 765–776. doi: 10.2174/1566524011313050008

Gunaseeli, I., Doss, M. X., Antzelevitch, C., Hescheler, J., and Sachinidis, A. (2010). Induced pluripotent stem cells as a model for accelerated patient- and disease-specific drug discovery. *Curr. Med. Chem.* 17, 759–766. doi: 10.2174/092986710790514480

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature07672

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477, 295–300. doi: 10.1038/nature10398

Guttman, M., and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* 482, 339–346. doi: 10.1038/nature10887

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., et al. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* 44, 3–12. doi: 10.1016/j.ymeth.2007.09.009

Hargus, G., Cooper, O., Deleidi, M., Levy, A., Lee, K., Marlow, E., et al. (2010). Differentiated Parkinson patient-derived induced pluripotent stem cells grow in the adult rodent brain and reduce motor asymmetry in Parkinsonian rats. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15921–15926. doi: 10.1073/pnas.1010209107

Hu, B. Y., Weick, J. P., Yu, J., Ma, L. X., Zhang, X. Q., Thomson, J. A., et al. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4335–4340. doi: 10.1073/pnas.0910012107

Jin, Z. B., Okamoto, S., Osakada, F., Homma, K., Assawachananont, J., Hirami, Y., et al. (2011). Modeling retinal degeneration using patient-specific induced pluripotent stem cells. *PLoS ONE* 6:e17084. doi: 10.1371/journal.pone.0017084

Jin, Z. B., Okamoto, S., Xiang, P., and Takahashi, M. (2012). Integration-free induced pluripotent stem cells derived from retinitis pigmentosa patient for disease modeling. *Stem Cells Transl. Med.* 1, 503–509. doi: 10.5966/sctm.2012-0005

Kawasaki, H., Mizuseki, K., Nishikawa, S., Kaneko, S., Kuwana, Y., Nakanishi, S., et al. (2000). Induction of midbrain dopaminergic neurons from ES cells by stromal cell-derived inducing activity. *Neuron* 28, 31–40. doi: 10.1016/S0896-6273(00)00083-0

Kawasaki, H., Suemori, H., Mizuseki, K., Watanabe, K., Urano, F., Ichinose, H., et al. (2002). Generation of dopaminergic neurons and pigmented epithelia from primate ES cells by stromal cell-derived inducing activity. *Proc. Natl. Acad. Sci. U.S.A.* 99, 1580–1585. doi: 10.1073/pnas.032662199

Kim, H., Lee, G., Ganat, Y., Papapetrou, E. P., Lipchina, I., Socci, N. D., et al. (2011). miR-371-3 expression predicts neural differentiation propensity in human pluripotent stem cells. *Cell Stem Cell* 8, 695–706. doi: 10.1016/j.stem.2011.04.002

Kim, J. B., Sebastiano, V., Wu, G., Arauzo-Bravo, M. J., Sasse, P., Gentile, L., et al. (2009). Oct4-induced pluripotency in adult neural stem cells. *Cell* 136, 411–419. doi: 10.1016/j.cell.2009.01.023

Kumano, K., Arai, S., Hosoi, M., Taoka, K., Takayama, N., Otsu, M., et al. (2012). Generation of induced pluripotent stem cells from primary chronic myelogenous leukemia patient samples. *Blood* 119, 6234–6242. doi: 10.1182/blood-2011-07-367441

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

Lee, T. H., Song, S. H., Kim, K. L., Yi, J. Y., Shin, G. H., Kim, J. Y. et al. (2010). Functional recapitulation of smooth muscle cells via induced pluripotent stem cells from human aortic smooth muscle cells. *Circ. Res.* 106, 120–128. doi: 10.1161/CIRCRESAHA.109.207902

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., et al. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125, 301–313. doi: 10.1016/j.cell.2006.02.043

Li, Z., Yang, C. S., Nakashima, K., and Rana, T. M. (2011). Small RNA-mediated regulation of iPS cell generation. *EMBO J.* 30, 823–834. doi: 10.1038/emboj.2011.2

Lin, M., Pedrosa, E., Shah, A., Hrabovsky, A., Maqbool, S., Zheng, D., et al. (2011). RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS ONE* 6:e23356. doi: 10.1371/journal.pone.0023356

Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., et al. (2011). Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. *Genes Dev.* 25, 2173–2186. doi: 10.1101/gad.17221311

Liu, J., Githinji, J., Mclaughlin, B., Wilczek, K., and Nolta, J. (2012). Role of miRNAs in neuronal differentiation from human embryonic stem cell-derived neural stem cells. *Stem Cell Rev.* 8, 1129–1137. doi: 10.1007/s12015-012-9411-6

Loewer, S., Cabili, M. N., Guttman, M., Loh, Y. H., Thomas, K., Park, I. H., et al. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* 42, 1113–1117. doi: 10.1038/ng.710

Miyoshi, N., Ishii, H., Nagano, H., Haraguchi, N., Dewi, D. L., Kano, Y., et al. (2011). Reprogramming of mouse and human cells to pluripotency using mature microRNAs. *Cell Stem Cell* 8, 633–638. doi: 10.1016/j.stem.2011.05.001

Morozova, O., and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92, 255–264. doi: 10.1016/j.ygeno.2008.07.001

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Ng, S. Y., Johnson, R., and Stanton, L. W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* 31, 522–533. doi: 10.1038/emboj.2011.459

Oh, Y., Wei, H., Ma, D., Sun, X., and Liew, R. (2012). Clinical applications of patient-specific induced pluripotent stem cells in cardiovascular medicine. *Heart* 98, 443–449. doi: 10.1136/heartjnl-2011-301317

Pedrosa, E., Sandler, V., Shah, A., Carroll, R., Chang, C., Rockowitz, S., et al. (2011). Development of patient-specific neurons in schizophrenia using induced pluripotent stem cells. *J. Neurogenet.* 25, 88–103. doi: 10.3109/01677063.2011.597908

Plath, K., and Lowry, W. E. (2011). Progress in understanding reprogramming to the induced pluripotent state. *Nat. Rev. Genet.* 12, 253–265. doi: 10.1038/nrg2955

Postigo, A. A., Depp, J. L., Taylor, J. J., and Kroll, K. L. (2003). Regulation of Smad signaling through a differential recruitment of coactivators and corepressors by ZEB proteins. *EMBO J.* 22, 2453–2462. doi: 10.1093/emboj/cdg226

Roese-Koerner, B., Stappert, L., Koch, P., Brustle, O., and Borghese, L. (2013). Pluripotent stem cell-derived somatic stem cells as tool to study the role of microRNAs in early human neural development. *Curr. Mol. Med.* 13, 707–722. doi: 10.2174/1566524011313050003

Rosa, A., and Brivanlou, A. H. (2011). A regulatory circuitry comprised of miR-302 and the transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation. *EMBO J.* 30, 237–248. doi: 10.1038/emboj.2010.319

Roukos, D. H. (2010). Next-generation sequencing and epigenome technologies: potential medical applications. *Expert Rev. Med. Devices* 7, 723–726. doi: 10.1586/erd.10.68

Schwartz, S. D., Hubschman, J. P., Heilwell, G., Franco-Cardenas, V., Pan, C. K., Ostrick, R. M., et al. (2012). Embryonic stem cell trials for macular degeneration: a preliminary report. *Lancet* 379, 713–720. doi: 10.1016/S0140-6736(12)60028-2

Sheik Mohamed, J., Gaughwin, P. M., Lim, B., Robson, P., and Lipovich, L. (2010). Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* 16, 324–337. doi: 10.1261/rna.1441510

Subramanyam, D., Lamouille, S., Judson, R. L., Liu, J. Y., Bucay, N., Derynck, R., et al. (2011). Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. *Nat. Biotechnol.* 29, 443–448. doi: 10.1038/nbt.1862

Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., et al. (2004). Human embryonic stem cells express a unique set of microRNAs. *Dev. Biol.* 270, 488–498. doi: 10.1016/j.ydbio.2004.02.019

Sun, N., Yazawa, M., Liu, J., Han, L., Sanchez-Freire, V., Abilez, O. J., et al. (2012). Patient-specific induced pluripotent stem cells as a model for familial dilated cardiomyopathy. *Sci. Transl. Med.* 4, 130ra147. doi: 10.1126/scitranslmed.3003552

Tachibana, M., Amato, P., Sparman, M., Gutierrez, N. M., Tippner-Hedges, R., Ma, H., et al. (2013). Human embryonic stem cells derived by somatic cell nuclear transfer. *Cell* 153, 1228–1238. doi: 10.1016/j.cell.2013.05.006

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., et al. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861–872. doi: 10.1016/j.cell.2007.11.019

Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., et al. (1998). Embryonic stem cell lines derived from human blastocysts. *Science* 282, 1145–1147. doi: 10.1126/science.282.5391.1145

Tubsuwan, A., Abed, S., Deichmann, A., Kardel, M. D., Bartholoma, C., Cheung, A., et al. (2013). Parallel assessment of globin lentiviral transfer in induced pluripotent stem cells and adult hematopoietic stem cells derived from the same transplanted beta-thalassemia patient. *Stem Cells* 31, 1785–1794. doi: 10.1002/stem.1436

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270, 484–487. doi: 10.1126/science.270.5235.484

Wilson, K. D., Venkatasubrahmanyam, S., Jia, F., Sun, N., Butte, A. J., and Wu, J. C. (2009). MicroRNA profiling of human-induced pluripotent stem cells. *Stem Cells Dev.* 18, 749–758. doi: 10.1089/scd.2008.0247

Wu, J. Q., Habegger, L., Noisa, P., Szekely, A., Qiu, C., Hutchison, S., et al. (2010). Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5254–5259. doi: 10.1073/pnas.0914114107

Xu, N., Papagiannakopoulos, T., Pan, G., Thomson, J. A., and Kosik, K. S. (2009). MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell* 137, 647–658. doi: 10.1016/j.cell.2009.02.038

Zeng, X., Cai, J., Chen, J., Luo, Y., You, Z. B., Fotter, E., et al. (2004). Dopaminergic differentiation of human embryonic stem cells. *Stem Cells* 22, 925–940. doi: 10.1634/stemcells.22-6-925

Zhang, A., Zhou, N., Huang, J., Liu, Q., Fukuda, K., Ma, D., et al. (2013). The human long non-coding RNA-RoR is a p53 repressor in response to DNA damage. *Cell Res.* 23, 340–350. doi: 10.1038/cr.2012.164

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

Received: 19 December 2013; accepted: 24 April 2014; published online: 14 May 2014.
*Citation: Lukovic D, Moreno-Manzano V, Klabusay M, Stojkovic M, Bhattacharya SS and Erceg S (2014) Non-coding RNAs in pluripotency and neural differentiation of human pluripotent stem cells. Front. Genet. **5**:132. doi: 10.3389/fgene.2014.00132*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.

Copyright © 2014 Lukovic, Moreno-Manzano, Klabusay, Stojkovic, Bhattacharya and Erceg. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 23 May 2014
doi: 10.3389/fgene.2014.00131

Identification of potential therapeutic targets in a model of neuropathic pain

Hemalatha B. Raju[1,2], Zoe Englander[3], Enrico Capobianco[1,4], Nicholas F. Tsinoremas[1] and Jessica K. Lerch[5]*

[1] Center for Computational Science, Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, USA
[2] Human Genetics and Genomics Graduate Program, University of Miami Miller School of Medicine, Miami, FL, USA
[3] Department of Biomedical Engineering, Duke University, Durham, NC, USA
[4] Laboratory of Integrative Systems Medicine, National Research Council (CNR), Pisa, Italy
[5] Department of Neuroscience, Center for Brain and Spinal Cord Repair, The Ohio State University, Columbus, OH, USA

Edited by:
Pietro Lio, University of Cambridge, UK

Reviewed by:
Ying Xu, West Virginia University, USA
Guanglong Jiang, Capital Normal University, China

****Correspondence:***
Jessica K. Lerch, Department of Neuroscience, Center for Brain and Spinal Cord Repair, The Ohio State University, 460 W 12th Ave., Columbus, OH 43210, USA
e-mail: jessica.lerch@osumc.edu

Neuropathic pain (NP) is caused by damage to the nervous system, resulting in dysfunction and aberrant pain. The cellular functions (e.g., peripheral neuron spinal cord innervation, neuronal excitability) associated with NP often develop over time and are likely associated with gene expression changes. Gene expression studies on the cells involved in NP (e.g., sensory dorsal root ganglion neurons) are publically available; the mining of these studies may enable the identification of novel targets and the subsequent development of therapies that are essential for improving quality of life for the millions of individuals suffering with NP. Here we analyzed a publically available microarray dataset (GSE30165) in order to identify new RNAs (e.g., messenger RNA (mRNA) isoforms and non-coding RNAs) underlying NP. GSE30165 profiled gene expression in dorsal root ganglion neurons (DRG) and in sciatic nerve (SN) after resection, a NP model. Gene ontological analysis shows enrichment for sensory and neuronal processes. Protein network analysis demonstrates DRG upregulated genes typical to an injury and NP response. Of the top changing genes, 34 and 36% are associated with more than one protein coding isoform in the DRG and SN, respectively. The majority of genes are receptor and enzymes. We identified 15 long non-coding RNAs (lncRNAs) targeting these genes in LNCipedia.org, an online comprehensive lncRNA database. These RNAs represent new therapeutic targets for preventing NP development and this approach demonstrates the feasibility of data reanalysis for their identification.

Keywords: gene expression, neuropathic pain, spinal cord injury, dorsal root ganglia, sciatic nerve, RNA

INTRODUCTION

The majority of patients with spinal cord injury (SCI) experience chronic pain, with a high percentage experiencing neuropathic pain (NP) (Siddall et al., 1999). NP develops concurrently with anatomical and physiological changes in the peripheral and central nervous system (PNS and CNS). For example, peripheral neuron innervation into the spinal dorsal horn (Nakamura and Myers, 2000) as well as both peripheral and central neurotransmitter expression and excitability change following injury (Chaplan et al., 1997; Fukuoka et al., 1998; Alexander et al., 2012). Identifying gene expression patterns in sensory neurons (i.e., dorsal root ganglion, DRG neurons) under normal and NP conditions is essential to understanding the genetic mechanisms behind the development of NP. Importantly, as the cells involved in NP are still alive, they are viable targets for small molecule or gene therapy approaches aimed at restoring normal function.

RNAs that do not code for a protein, or non-coding RNAs (ncRNAs; e.g., microRNAs: miRNAs and long ncRNAs: lncRNAs), are implicated in many biological and pathological processes such as cancer development, progression, and metastasis (Calin and Croce, 2006; Zhong et al., 2009; Gutschner and Diederichs, 2012; Ziats and Rennert, 2013), and genetic variations within ncRNA loci are increasingly associated with developmental disorders and disease states (Pasmant et al., 2011; Richardson et al., 2011; Zhang et al., 2012). Since RNA-regulated gene expression is increasingly involved in pathological conditions we wanted to understand RNA expression and diversity in the context of NP. Indeed evidence for the involvement of lnc and miRNAs in the development of NP is emerging although in its infancy. For example, *Kcna2 antisense* lncRNA is expressed in DRG neurons and causes or reduces NP through its ability to regulate the voltage-dependent potassium channel, *Kcna2*, impacting neuronal excitability (Zhao et al., 2013). A recent study examined miRNA expression along with gene expression in a sciatic nerve (SN) ligation model of NP (von Schack et al., 2011). The authors found 63 miRNAs changing expression; interestingly the majority (59) of miRNAs were down-regulated in the ipsilateral DRG one level above the injury (von Schack et al., 2011). It is likely that additional ncRNAs contribute to NP development after SCI but identification of these RNAs has remained challenging.

In addition to ncRNAs, messenger RNA (mRNA) isoforms drive distinct biological functions (Hong et al., 2008) and may underlie pathological conditions (Gerstin et al., 1998; Pertin et al., 2005; Dina et al., 2008; Kanzaki et al., 2012). For example, neuregulin-1 has three isoforms that undergo alternative expression regulation (Nrg1 I and II increase and Nrg1 III decreases)

after spinal nerve ligation in the rat, changes associated with mechanical sensitivity of the ipsilateral hind paw (Kanzaki et al., 2012). Protein kinase C isoform delta is linked to L-type calcium channel upregulation and may contribute to alcohol-induced peripheral neuropathy (Gerstin et al., 1998; Dina et al., 2008). These findings demonstrate that mRNA isoforms play an important biological role but the paucity of evidence for mRNA isoforms in critical biological roles may in part be due to lack of their complete identification.

Here we sought to identify additional mRNA isoforms and regulatory RNAs contributing to NP development. Multiple methods are available for understanding gene expression (e.g., microarray, RNA-seq) and many laboratories are applying these methods to various pathologies such as SCI and NP. The majority of SCI research is performed in *Rattus norvegicus* (rat) because the injury response and lesion formation are similar to human (Sroga et al., 2003). A search of the Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2005) database using "drg pain" or "drg NP" as terms produced over 200 results, with the majority of studies in rat using microarrays. We examined several datasets and chose GSE30165 because it examined global gene expression changes after SN resection in both the DRG and SN. We identified the differentially expressed rat genes and then converted them to their mouse homologs using a sequence based strategy, allowing us to identify the associated mRNA isoforms and regulatory RNAs. This strategy globally identifies possible new RNAs for targeting and provides a roadmap for the re-evaluation of already existing datasets.

MATERIALS AND METHODS

SCIATIC NERVE INJURY

This following procedural guideline was kindly provided by Dr. Bin Yu, Jiangsu Key Laboratory of Neuroregeneration, Nantong University, Nantong, China, the investigator who uploaded the results to the NCBI GEO Database. Briefly, male Sprague-Dawley rats (180–220 g), were anesthetized by an intraperitoneal injection of complex narcotics (85 mg/kg trichloroacetaldehyde monohydrate, 42 mg/kg magnesium sulfate, 17 mg/kg sodium pentobarbital), and the SN was exposed and lifted through an incision on the lateral aspect of the mid-thigh of the left hind limb. A 1 cm long segment of SN was then resected at the site just proximal to the division of tibial and common peroneal nerves, and the incision sites were then closed. To minimize discomfort and possible painful mechanical stimulation, the rats were housed in large cages with sawdust bedding after surgery. L4-6 DRG tissues and SN tissues (0.5 cm) were collected at different time points after injury, respectively. All the experimental procedures involving animals were conducted in accordance with Institutional Animal Care guidelines and ethically approved by the Administration Committee of Experimental Animals, Jiangsu Province, China.

GENE EXPRESSION ANALYSIS

Gene expression data and analysis was obtained from the NCBI NIH GEO, dataset GSE30165. Sample preparation was described in the dataset design description. Briefly, gene expression levels from L4-6 DRG tissues and proximal SN tissues (0.5 cm) were examined at 0 days, 1 day, 4 days, 7 days, and 14 days after SN resection. This dataset consisted of three samples each for the DRG and SN tissues, and gene expression data was available for all samples at each of the 5 times points. GEO2R was used to compare expression between sham and 1 day post-injury (dpi); sham and 4 dpi; sham and 7 dpi; and finally sham and 14 dpi for both the DRG and SN. GEO2R analyzes gene expression using GEOquery and the Linear Models of Microarray Analysis R package (limma) (Edgar et al., 2002; Gentleman et al., 2004; Smyth, 2004, 2005; Barrett et al., 2005; Davis and Meltzer, 2007). First, GEOquery formats the data into tables for R and then limma R applies the Benjamini and Hochberg False Discovery Rate (FDR) correction for multiple comparisons testing to determine the adjusted p-value, p-value, moderate t-statistic, log fold change, and the moderate F-statistic (Edgar et al., 2002; Barrett et al., 2005; Gentleman et al., 2004; Smyth, 2004; Davis and Meltzer, 2007). We determined the top 250 genes that changed significantly at each time point compared to baseline with an adjusted p-value of <0.05 in order to identify the genes that changed over the time-course following injury, and not to identify the most differentially expressed genes across the experiment. We looked at the top 250 differentially expressed genes in each comparison to focus our results to only the genes that changed the most at each time point. The final subset of genes from each comparison was restricted to only those with a fold change in either direction that was greater than 2 for the DRG and SN tissues separately. The final list of genes consisted of all that had at least one time point that showed a change with an adjusted $p < 0.05$ and a fold change of 2, resulting in the identification of 246 genes for the DRG dataset and 549 for the SN dataset. The values at each time point were normalized with respect to the average expression value over all time points for each gene. Heatmaps were generated using the bioinformatics toolbox in Matlab.

GENE ONTOLOGY ANALYSES

The final gene list after applying the cutoffs (adj. $p < 0.05$ and fold change of 2) was input into the DAVID Functional Annotation interface and submitted as a gene list selecting species *Rattus norvegicus* (Huang da et al., 2009a,b). Gene Ontology (GO) charts were created using the following options: thresholds: count 2, EASE 0.1; Benjamini correction, Number of records = 1000.

RAT TO MOUSE CONVERSION

The microarray probe sequences for the differentially expressed genes at different time points following nerve injury were extracted for both DRG and proximal SN tissues from the GEO, Agilent-014879 Whole Rat Genome Microarray 4x44K G4131F. The extracted sequences were then aligned against mouse reference (Ensembl), Mus_musculus.GRCm38.74.cdna.all.fa (Flicek et al., 2013, 2014) using BLAT (Kent, 2002), a fast spliced alignment program. BLAT was executed with blast8 as output and all other parameters set at default values. The alignment was done against mouse reference to identify the homologous sequences between the two rodent species. The aligned rat sequences were then annotated using mouse, Mus_musculus.GRCm38.74.gtf to associate the rat genes from the microarray data against the corresponding mouse homologs based on the alignment results, and then the gene biotypes were assigned based on the mouse

annotation provided by Ensembl (Hubbard et al., 2002). Since the rat annotations are not defined as thoroughly as the mouse (**Table 1**), mouse annotation was chosen to classify the gene biotypes that includes protein-coding and specific type of non-coding.

NETWORK ANALYSES

Protein interactions (Figure 2)

A very popular tool named STRING (V. 9.1, http://string-db.org/) was used for visualizing interactomes starting from identified differentially expressed entities (genes and transcripts) in both species. In particular the confidence and evidence STRING protein–protein interaction modes were applied.

In confidence view, stronger associations are represented by thicker lines, while in evidence view; different line colors represent the types of evidence for specific associations: expression, binding catalysis, and post-translational modification.

Expression interactions (Figure 3)

Mouse gene symbols returned from the rat to mouse conversion were uploaded to Ingenuity® Systems (www.ingenuity.com). Interactions were added using the Connect Tool. Molecules involved in depolarization and nociception were identified using the Overlay Tool. The RNAs with greater than 1 CDS and associated ncRNAs were added by hand.

RESULTS

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

We identified the top 250 IDs from the microarray dataset that met our cutoffs for a significant expression change (adjusted $p < 0.05$ and fold change >2; **Figures 1A,B**). There were 549 unique IDs corresponding to 366 rat genes with gene symbols in the SN and 246 unique IDs corresponding to 158 rat genes with gene symbols in the DRG (**Figure 1C**; Supplementary Tables 1, 2). 25 of the top changing were found in both samples, 18 of which had associated gene symbols (Supplementary Table 3). In the SN, a subset of genes decreased expression (Group 1, **Figure 1A**; Supplementary Table 1), while the bulk increased in expression (Group 2, **Figure 1A**; Supplementary Table 1). In the DRG, the majority of genes increased in expression (Group2, **Figure 1B**; Supplementary Table 2). These data indicate major gene expression changes in the SN and in the DRG after injury.

GENE ONTOLOGY ANALYSIS

A GO term enrichment analysis (Huang da et al., 2009a,b) was subsequently performed to gain a deeper understanding of these genes. GO enrichment analysis assigns general descriptions based on biological function, cellular component, and molecular function, to groups of genes. We isolated the up or down regulated genes (SN: Group 1–3, Supplementary Table 1; DRG: Group 1 and 2, Supplementary Table 2) and performed GO analysis using DAVID Bioinformatics Resource v6.7 (Huang da et al., 2009a,b). GO analysis on the down-regulated genes in the SN sample show the majority of biological processes are biosynthetic and catabolic functions while the majority of the up-regulated processes are related to the detection of stimuli and signaling responses (Supplementary Table 4). Not surprisingly, the majority of cellular components up- or down-regulated are associated with the cytoplasm and cellular membrane (Supplementary Table 4). The majority of molecular functions switch from ion binding (downregulated) to chemokine and enzymatic activities (upregulated; Supplementary Table 5). These data suggest a switch from neurotransmission and normal sensory functioning to immune response detection and receptor activation, consistent with a switch from normal sensory neurotransmission to an injury response in the SN. In the DRG sample, the majority of genes were upregulated after injury (**Figure 1B**). Most biological processes in the DRG upregulated genes fall into signaling pathways (e.g., G-protein, neuropeptide) or detection and reaction to stimuli (e.g., sensory perception of chemical stimulus, inflammatory response; Supplementary Table 5). In cellular component, the majority associated with the membrane, extracellular space, and

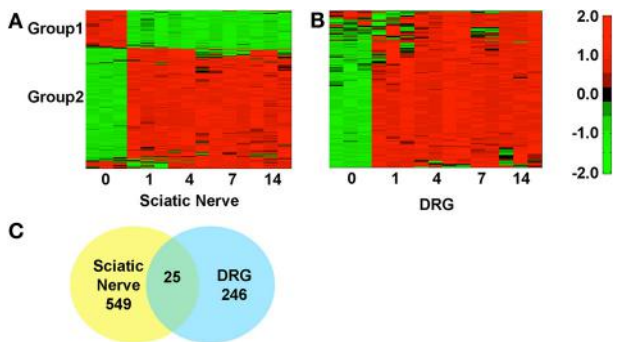


FIGURE 1 | Hierarchical clustering of normalized expression values at various days post-sciatic nerve resection in the sciatic nerve and in the DRG. The heatmaps reflect gene expression values normalized to the mean across all time points (day 0, 1, 4, 7, and 14 post-injury) for genes that met the cutoff in at least one time point ($p < 0.05$ and fold change >2). **(A)** In the SN there are two distinct groups, 1 and 2, which decrease (green) or increase (red) in expression, Supplementary Table 2. **(B)** In the DRG the majority of genes decrease in expression. **(C)** Some genes overlap (25) between SN and DRG but the majorities of changing genes were unique to each tissue.

Table 1 | The rat genome has fewer RNA annotations in all categories.

	Protein coding	Micro	Long non-coding	Small-nucleolar	Small-nuclear	Antisense
Mus musculus	22,740	2010	1795	1556	1387	1476
Rattus norvegicus	19,878	419	0	0	0	0

The number of protein coding, micro, long non-coding, small nucleolar, small nuclear, and antisense RNAs found in the Mus_musculus.GRCm38.74.gtf and Rattus_norvegicus.Rnor_5.0.74.gtf from the Ensemble Database.

nerve terminal (cellular component, Supplementary Table 6) and the molecular functions are associated with receptors, cytokines, or hormone activity (molecular function, Supplementary Table 5). These data suggest a major change in DRG gene expression in areas directly associated with NP development such as neurotransmission and receptor expression (Xu et al., 1993, 2007; Fukuoka et al., 1998; Sah et al., 2003; Pertin et al., 2005; Mika et al., 2008; Miller et al., 2009).

IDENTIFICATION OF ISOFORMS

During the analysis it was observed that many rat UniqueIDs were not associated with a gene name or symbol (Supplementary Tables 2, 3). Indeed the rat genome contains far fewer elements compared to the mouse (**Table 1**). This suggests that using the rat for gene array and/or RNA-seq experiments is problematic and could severely limit gene expression analysis interpretation. To address this problem and gain insight into gene expression and regulation we converted the rat genes (**Figure 1**) to their mouse homologs using a sequence based strategy (Methods; Supplementary Tables 6, 7). BLAT finds similar sequences of length 25 base pairs or greater. We set a homology threshold of 84% and higher to extract the potential homologs from the BLAT output using the default parameters. We retrieved the corresponding target mouse gene names from the BLAT output and used them for downstream analysis. Using this homology-based strategy we identified 455 corresponding mouse genes in SN and 167 in the DRG (Supplementary Tables 6, 7). These genes give rise to hundreds of isoforms and produce multiple protein isoforms (**Table 2**). Isoform switching [aka: alternative open reading frame (ORF) utilization], is one mechanism driving neural development (Ruusuvuori et al., 2004; Bani-Yaghoub et al., 2007) and contributing to disease states in the body (Periasamy and Kalyanasundaram, 2007). It could be a potential mechanism underlying NP development. We identified numerous differentially expressed genes whose isoforms differ at the level of the coding DNA sequence (CDS) leading to alternative ORFs (**Table 2**). Protein coding differences were most abundant in enzymes, ion-channels, transcription regulators, and G-protein coupled receptors (**Table 3**), all highly associated and implicated in NP.

NETWORK ANALYSIS AND ncRNA REGULATION PREDICTION

In large datasets relationships between differentially expressed genes are uncovered by examining protein-protein interactions. We used STRING (Franceschini et al., 2013), which utilizes both known and predicted protein associations to generate protein interaction networks. In DRG up-regulated genes, several direct protein interactions among molecules known to change expression after DRG neuron injury were uncovered. The most prominent group of interactions in this analysis was between the neuropeptides vasoactive intestinal peptide (VIP), its receptors (VIPR1/2), pituitary adenylate cyclase-activating polypeptide (ADCYAP1 aka PACAP), its receptor (ADCYAP1R1), and cholecystokinin (CCK) and its receptors (CCKAR, CCKBR; **Figure 2**). VIP, ADCYAP1, and CCK are upregulated in DRG after injury and are associated with NP (Nielsch and Keen, 1989; Xu et al., 1993; Ma and Bisby, 1998; Ohsawa et al., 2002). These observations support the involvement of these neuropeptides in NP development and support that this dataset is reflecting gene expression changes regulating NP. Interestingly, these neuropeptide receptors have multiple isoforms (Bokaei et al., 2006; Nachtergael et al., 2006), but to date no studies have examined their function in NP models.

The role of RNA isoforms and their contributions to neuronal development and pathology is slowly being elucidated (Gerstin

Table 2 | Differentially expressed genes have abundant transcript diversity.

	SN	DRG
Genes	445	167
Transcripts	1451	409
Transcripts with different CDS	162	36

Mouse transcript information was obtained from the Ensemble Mus_musculus.GRCm38.74.gtf. The number of genes, transcripts and transcript harboring changes in the coding DNA sequence (CDS) was identified.

Table 3 | Enzymes and transcription regulators are associated with the most protein coding isoforms in the SN and DRG, respectively.

	SN	DRG
Enzyme	49	3
G-protein coupled receptor	6	5
Ion channel	12	1
Kinase	5	1
Peptidase	7	2
Transcription regulator	2	8
Translation regulator	1	1
Transmembrane regulator	5	2
Transporter	1	1

Categories were assigned using the molecular annotations feature in Ingenuity® Systems, www.ingenuity.com.

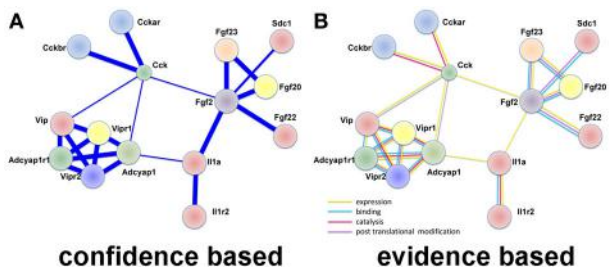


FIGURE 2 | Protein interaction network using LINC identifies neuropeptide interactions after injury in the DRG. (A) A confidence view of protein interactions. Thicker lines represent stronger associations. **(B)** An evidence view of protein interactions. Both observations support previous studies demonstrating increased expression of neuropeptides after injury. In particular, it is noticed the modular connectivity centered on VIP, ADCYAP1, and CCK, all appearing up-regulated in DRG after injury and associated with NP. Evidence is taken from reports in the literature. Yellow: expression; blue: binding; lilac: violet: catalysis; and post-translation modification. This observation supports previous studies demonstrating increased expression of neuropeptides after injury.

et al., 1998; Pertin et al., 2005; Dina et al., 2008; Hong et al., 2008; Kanzaki et al., 2012; Lerch et al., 2012b) but a full understanding of RNA isoform diversity is broadly lacking. To identify mRNAs with alternative CDS's with the potential to impact NP development we created a network of DRG enriched genes with the ability to directly regulate each other's expression (**Figure 3**). As expected, many genes have a role in neuronal depolarization (Jarvis et al., 1995; Beaudet et al., 2000) and nociception (Jeftinija et al., 1982; Mika et al., 2008; Belcheva et al., 2009), two properties of sensory neurons altered in NP states (Chaplan et al., 1997; Fukuoka et al., 1998; Alexander et al., 2012). We highlight genes with more than one CDS because alternative CDS's leads to changes in functional protein domains which alter cellular function.

LncRNAs have recently been demonstrated to regulate sensory neuronal excitability and NP (Zhao et al., 2013). To identify potential additional gene targets for regulation we searched a database of lncRNAs (Volders et al., 2013). The nomenclature for lncRNAs in this database makes searching straightforward. Transcripts overlapping one or more exons are named with the same gene symbol and therefore considered the same gene (Volders et al., 2013). Searching gene symbols identifies associated lncRNAs. We found 15 lncRNAs conserved between human and mouse in our dataset that corresponded to significantly changing genes (Supplementary Table 8). There were an additional 11 lncRNAs not identified as conserved across species (http://www.lncipedia.org/db/search). Given that lncRNAs have a high degree of evolutionary conservation (Qu and Adelson, 2012); it is possible these additional genes are regulated similarly in rats and mice (**Figure 3**). The genes identified with a potential lncRNAs fall into many categories such as enzymes (HSD3B2 and PDE6B), growth factors (FGF2), transmembrane receptors (CHRNA1, HLA-DRA, HLA-DRB1, IL1R2, and SEMA6A), and transcriptional regulators (NKZ6.2, SOX11, and STAT4). This demonstrates that lncRNA regulation of gene expression is likely not limited to one particular gene category or class of protein. These strategies highlight a way to reanalyze existing data and extend it to identify novel mRNA isoforms and regulatory RNAs to further our understanding of NP and can be extended to other disease datasets.

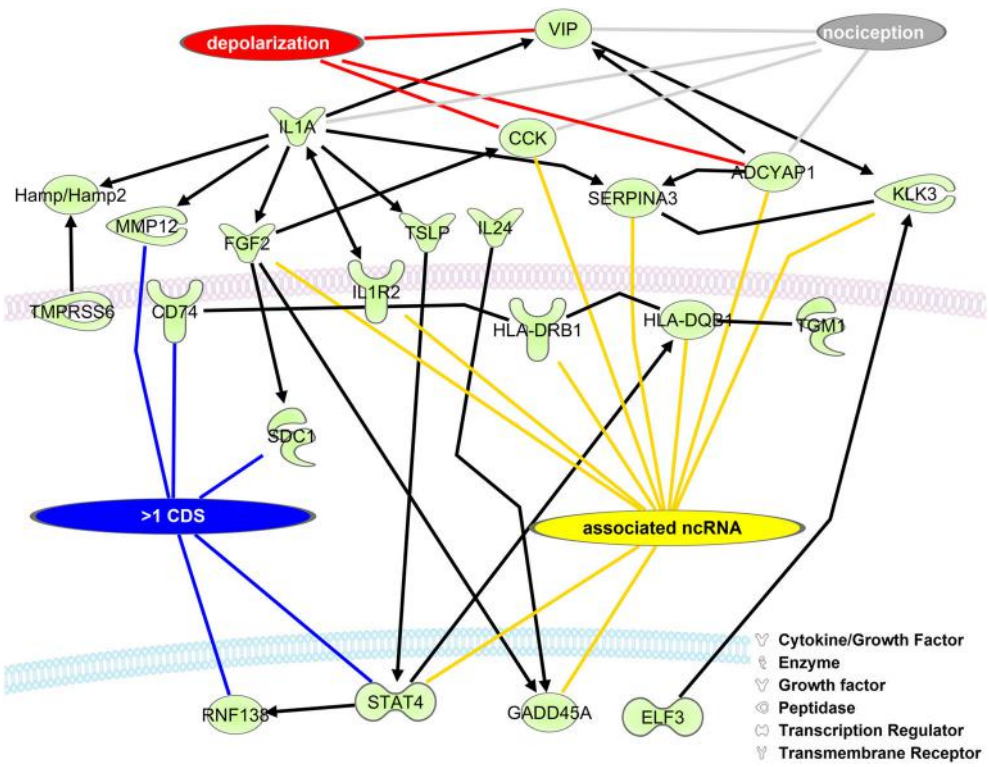


FIGURE 3 | Upregulated DRG neuronal network is associated with mRNA isoforms and ncRNAs. The list of mouse homolog DRG upregulated genes (Group 2, Supplementary Table 2) was put into a direct interaction network (Ingenuity® Systems, www.ingenuity.com). Genes having more than one CDS (blue line), an associated ncRNA (yellow line), involved in depolarization (red line), and/or nociception (gray line) are indicated. Black lines with arrows indicate expression activation. Straight black lines indicate protein-protein interaction. ADCYAP1, adenylate cyclase activating polypeptide 1; CCK, cholecystokinin; CD74, CD74 molecule, major histocompatibility complex, class II invariant chain; ELF3, E74-like factor 3; FGF2, fibroblast growth factor 2; GADD45A, growth arrest and DNA-damage-inducible, alpha; Hamp/Hamp2, hepcidin antimicrobial peptide; HLA-DQB1, major histocompatibility complex, class II, DQ beta 1; HLA-DRB1, major histocompatibility complex, class II, DR beta 1; IL1A, interleukin 1, alpha; IL1R2, interleukin 1 receptor, type II; IL24, interleukin 24; KLK3, kallikrein-related peptidase 3; MMP12, matrix metallopeptidase 12; RNF138, ring finger protein 138; E3 ubiquitin protein ligase,: syndecan 1; SERPINA3, serpin peptidase inhibitor, clade A, member 3; STAT4, signal transducer and activator of transcription 4; TGM1, transglutaminase 1; TMPRSS6, transmembrane protease, serine 6; TSLP, thymic stromal lymphopoietin; VIP, vasoactive intestinal peptide.

DISCUSSION

Millions of people worldwide, including the majority of SCI patients, experience NP. The prevalence of NP and the minimal availability of effective treatment options make the identification of the molecular pathways leading to NP development a high priority. The majority of studies examining gene expression changes in NP models use a microarray approach (except one study, GSE53768, released 01/07/2014 which used RNA-seq and was unpublished at the time paper submission). Therefore, the identification of all expressed RNAs (e.g., isoforms and regulatory RNAs) is lacking, omitting numerous potential therapeutic targets. To identify RNA isoforms and regulatory RNAs relevant to NP we examined differentially expressed genes from a publically available microarray study using a rat NP model (GSE30165). We identified over 200 genes significantly changing in DRG neurons and over 400 in the SN (**Figure 1**; Supplementary Tables 1–3). Differentially expressed genes in this dataset show GO enrichment for inflammatory processes, critical regulators and contributors to NP (Supplementary Tables 4, 5; Hulsebosch, 2008; Kigerl et al., 2009; Alexander et al., 2012).

One challenge of this dataset was the use of a rat model system. We suggest that genetic studies should be performed in mice given that the rat genome annotation is vastly incomplete (**Table 1**). Given the lack of annotation, our ability to identify mRNA isoforms and ncRNAs from the rat database was limited. Therefore, we retrieved differentially expressed rat RNA sequences (Supplementary Tables 1, 2), and took mouse RNAs with an 84% and greater homology to the rat sequences and then examined these sequences for RNA isoforms and potential regulatory RNAs. We identified 455 mouse genes in SN, 167 in DRG, and thousands of RNA isoforms for each gene (Supplementary Tables 6, 7). We created a network of the interacting up-regulated genes from the DRG dataset. Interestingly, in this dataset we identified 15 conserved lncRNAs that could regulate these transcripts in the rat or mouse (**Figure 3**, Supplementary Table 8). LncRNAs regulate protein coding gene expression by affecting DNA organization (e.g., defining chromatin domains; Rinn et al., 2007), transcription (Zhao et al., 2013), and/or post-transcription processing (Mercer et al., 2009). Most lncRNAs are associated with a decrease in their target's expression [e.g., HOTAIR's repression of the HoxD locus (Rinn et al., 2007); Kcna3 antisense repression of Kcna3 (Zhao et al., 2013)]. There is a single compelling example of an lncRNA regulating NP development. Kcna3 antisense expression increased after peripheral nerve injury, increased neuronal excitability, and when overexpressed induced NP pain symptoms (Zhao et al., 2013), a remarkable effect for a single lncRNA. One area of future investigation is to determine global lncRNA expression changes after SCI, because while Kcnc3 antisense expression increased, it is just as likely that some lncRNAs expression would decrease. In this study we found that SNI in the DRG led to a majority of genes increasing expression (**Figure 1**). Therefore, it is possible that SNI causes a reduction in the lncRNAs we identified (**Figure 3**) that contributed to their target gene expression increase (**Figure 1**). In addition, we hypothesize that these lncRNAs represent therapeutic targets since overexpressing them would repress their target genes and potentially reduce NP symptoms. For example, the increases in interleukin 1 receptor (IL1R), adenylate cyclase activating polypeptide 1 (ADCYAP1), and cholecystokinin (CCK) may be associated with a decrease in their associated lncRNAs (**Figure 3**). This interaction, if occurring, may contribute to their roles in nociception (**Figure 3**; IL1R through binding to IL1A and ADCYAP1 through VIP binding; Jeftinija et al., 1982; Xu et al., 1993; Mika et al., 2008). We acknowledge that while these are intriguing possibilities, all of these isoforms and lncRNAs require functional studies to test if they are viable candidates, but note that identification is the first step toward determining functional relevance.

NP is debilitating and in need of better therapeutic strategies. A multitude of well-controlled publically available data exists in the GEO database. We identified isoform diversity and potential ncRNAs through a data reanalysis using a straightforward bioinformatic approach. There is growing evidence that RNA isoforms and lncRNAs are important regulators of cellular function and contribute to pathological processes (Gerstin et al., 1998; Hong et al., 2008; Kanzaki et al., 2012; Lerch et al., 2012b). Future studies will employ RNA-seq enabling full scale detection of all RNAs within a cell type (Faghihi and Wahlestedt, 2009; Lerch et al., 2012a,b) giving a complete picture of gene expression but here we demonstrate a fast and economical way to find new targets underlying NP development.

ACKNOWLEDGMENTS

We thank Dr. Bin Yu from the Jiangsu Key Laboratory of Neuroregeneration at Nantong University Nantong, China for providing details about the DRG and SN injury and tissue processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00131/abstract

REFERENCES

Alexander, J. K., Cox, G. M., Tian, J. B., Zha, A. M., Wei, P., Kigerl, K. A., et al. (2012). Macrophage migration inhibitory factor (MIF) is essential for inflammatory and neuropathic pain and enhances pain in response to stress. *Exp. Neurol.* 236, 351–362. doi: 10.1016/j.expneurol.2012.04.018

Bani-Yaghoub, M., Kubu, C. J., Cowling, R., Rochira, J., Nikopoulos, G. N., Bellum, S., et al. (2007). A switch in numb isoforms is a critical step in cortical development. *Dev. Dyn.* 236, 696–705. doi: 10.1002/dvdy.21072

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., et al. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.* 33, D562–D566. doi: 10.1093/nar/gki022

Beaudet, M. M., Parsons, R. L., Braas, K. M., and May, V. (2000). Mechanisms mediating pituitary adenylate cyclase-activating polypeptide depolarization of rat sympathetic neurons. *J. Neurosci.* 20, 7353–7361.

Belcheva, I., Ivanova, M., Tashev, R., and Belcheva, S. (2009). Differential involvement of hippocampal vasoactive intestinal peptide in nociception of rats with a model of depression. *Peptides* 30, 1497–1501. doi: 10.1016/j.peptides.2009.05.015

Bokaei, P. B., Ma, X. Z., Byczynski, B., Keller, J., Sakac, D., Fahim, S., et al. (2006). Identification and characterization of five-transmembrane isoforms of human vasoactive intestinal peptide and pituitary adenylate cyclase-activating polypeptide receptors. *Genomics* 88, 791–800. doi: 10.1016/j.ygeno.2006.07.008

Calin, G. A., and Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nat. Rev. Cancer* 6, 857–866. doi: 10.1038/nrc1997

Chaplan, S. R., Malmberg, A. B., and Yaksh, T. L. (1997). Efficacy of spinal NMDA receptor antagonism in formalin hyperalgesia and nerve injury evoked allodynia in the rat. *J. Pharmacol. Exp. Ther.* 280, 829–838.

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

Dina, O. A., Khasar, S. G., Alessandri-Haber, N., Green, P. G., Messing, R. O., and Levine, J. D. (2008). Alcohol-induced stress in painful alcoholic neuropathy. *Eur. J. Neurosci*. 27, 83–92. doi: 10.1111/j.1460-9568.2007.05987.x

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 30, 207–210. doi: 10.1093/nar/30.1.207

Faghihi, M. A., and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol*. 10, 637–643. doi: 10.1038/nrm2738

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

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res*. 42, D749–D755. doi: 10.1093/nar/gkt1196

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 41, D808–D815. doi: 10.1093/nar/gks1094

Fukuoka, T., Tokunaga, A., Kondo, E., Miki, K., Tachibana, T., and Noguchi, K. (1998). Change in mRNAs for neuropeptides and the GABA(A) receptor in dorsal root ganglion neurons in a rat experimental neuropathic pain model. *Pain* 78, 13–26. doi: 10.1016/S0304-3959(98)00111-0

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

Gerstin, E. H. Jr., McMahon, T., Dadgar, J., and Messing, R. O. (1998). Protein kinase Cdelta mediates ethanol-induced up-regulation of L-type calcium channels. *J. Biol. Chem*. 273, 16409–16414. doi: 10.1074/jbc.273.26.16409

Gutschner, T., and Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol*. 9, 703–719. doi: 10.4161/rna.20481

Hong, E. J., McCord, A. E., and Greenberg, M. E. (2008). A biological function for the neuronal activity-dependent component of Bdnf transcription in the development of cortical inhibition. *Neuron* 60, 610–624. doi: 10.1016/j.neuron.2008.09.024

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 37, 1–13. doi: 10.1093/nar/gkn923

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

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002). The Ensembl genome database project. *Nucleic Acids Res*. 30, 38–41. doi: 10.1093/nar/30.1.38

Hulsebosch, C. E. (2008). Gliopathy ensures persistent inflammation and chronic pain after spinal cord injury. *Exp. Neurol*. 214, 6–9. doi: 10.1016/j.expneurol.2008.07.016

Jarvis, C. R., Van de Heijning, B. J., and Renaud, L. P. (1995). Cholecystokinin evokes vasopressin release from perfused hypothalamic-neurohypophyseal explants. *Regul. Pept*. 56, 131–137. doi: 10.1016/0167-0115(95)00005-V

Jeftinija, S., Murase, K., Nedeljkov, V., and Randic, M. (1982). Vasoactive intestinal polypeptide excites mammalian dorsal horn neurons both *in vivo* and *in vitro*. *Brain Res*. 243, 158–164. doi: 10.1016/0006-8993(82)91131-3

Kanzaki, H., Mizobuchi, S., Obata, N., Itano, Y., Kaku, R., Tomotsuka, N., et al. (2012). Expression changes of the neuregulin 1 isoforms in neuropathic pain model rats. *Neurosci. Lett*. 508, 78–83. doi: 10.1016/j.neulet.2011.12.023

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res*. 12, 656–664. doi: 10.1101/gr.229202.

Kigerl, K. A., Gensel, J. C., Ankeny, D. P., Alexander, J. K., Donnelly, D. J., and Popovich, P. G. (2009). Identification of two distinct macrophage subsets with divergent effects causing either neurotoxicity or regeneration in the injured mouse spinal cord. *J. Neurosci*. 29, 13435–13444. doi: 10.1523/JNEUROSCI.3257-09.2009

Lerch, J. K., Bixby, J. L., and Lemmon, V. P. (2012a). Isoform diversity and its importance for axon regeneration. *Neuropathology* 32, 420–431. doi: 10.1111/j.1440-1789.2011.01280.x

Lerch, J. K., Kuo, F., Motti, D., Morris, R., Bixby, J. L., and Lemmon, V. P. (2012b). Isoform diversity and regulation in peripheral and central neurons revealed through RNA-seq. *PLoS ONE* 7:e30417. doi: 10.1371/journal.pone.0030417

Ma, W., and Bisby, M. A. (1998). Partial and complete sciatic nerve injuries induce similar increases of neuropeptide Y and vasoactive intestinal peptide immunoreactivities in primary sensory neurons and their central projections. *Neuroscience* 86, 1217–1234. doi: 10.1016/S0306-4522(98)00068-2

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet*. 10, 155–159. doi: 10.1038/nrg2521

Mika, J., Korostynski, M., Kaminska, D., Wawrzczak-Bargiela, A., Osikowicz, M., Makuch, W., et al. (2008). Interleukin-1 alpha has antiallodynic and antihyperalgesic activities in a rat neuropathic pain model. *Pain* 138, 587–597. doi: 10.1016/j.pain.2008.02.015

Miller, R. J., Jung, H., Bhangoo, S. K., and White, F. A. (2009). Cytokine and chemokine regulation of sensory neuron function. *Handb. Exp. Pharmacol*. 194, 417–449. doi: 10.1007/978-3-540-79090-7_12

Nachtergael, I., Gaspard, N., Langlet, C., Robberecht, P., and Langer, I. (2006). Asn229 in the third helix of VPAC1 receptor is essential for receptor activation but not for receptor phosphorylation and internalization: comparison with Asn216 in VPAC2 receptor. *Cell. Signal*. 18, 2121–2130. doi: 10.1016/j.cellsig.2006.03.006

Nakamura, S. I., and Myers, R. R. (2000). Injury to dorsal root ganglia alters innervation of spinal cord dorsal horn lamina involved in nociception. *Spine (Phila Pa. 1976)* 25, 537–542. doi: 10.1097/00007632-200003010-00002

Nielsch, U., and Keen, P. (1989). Reciprocal regulation of tachykinin- and vasoactive intestinal peptide-gene expression in rat sensory neurones following cut and crush injury. *Brain Res*. 481, 25–30. doi: 10.1016/0006-8993(89)90481-2

Ohsawa, M., Brailoiu, G. C., Shiraki, M., Dun, N. J., Paul, K., and Tseng, L. F. (2002). Modulation of nociceptive transmission by pituitary adenylate cyclase activating polypeptide in the spinal cord of the mouse. *Pain* 100, 27–34. doi: 10.1016/S0304-3959(02)00207-5

Pasmant, E., Sabbagh, A., Vidaud, M., and Bieche, I. (2011). ANRIL, a long, non-coding RNA, is an unexpected major hotspot in GWAS. *FASEB J*. 25, 444–448. doi: 10.1096/fj.10-172452

Periasamy, M., and Kalyanasundaram, A. (2007). SERCA pump isoforms: their role in calcium transport and disease. *Muscle Nerve* 35, 430–442. doi: 10.1002/mus.20745

Pertin, M., Ji, R. R., Berta, T., Powell, A. J., Karchewski, L., Tate, S. N., et al. (2005). Upregulation of the voltage-gated sodium channel beta2 subunit in neuropathic pain models: characterization of expression in injured and non-injured primary sensory neurons. *J. Neurosci*. 25, 10970–10980. doi: 10.1523/JNEUROSCI.3066-05.2005

Qu, Z., and Adelson, D. L. (2012). Evolutionary conservation and functional roles of ncRNA. *Front. Genet*. 3:205. doi: 10.3389/fgene.2012.00205

Richardson, K., Lai, C. Q., Parnell, L. D., Lee, Y. C., and Ordovas, J. M. (2011). A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. *BMC Genomics* 12:504. doi: 10.1186/1471-2164-12-504

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323. doi: 10.1016/j.cell.2007.05.022

Ruusuvuori, E., Li, H., Huttu, K., Palva, J. M., Smirnov, S., Rivera, C., et al. (2004). Carbonic anhydrase isoform VII acts as a molecular switch in the development of synchronous gamma-frequency firing of hippocampal CA1 pyramidal cells. *J. Neurosci*. 24, 2699–2707. doi: 10.1523/JNEUROSCI.5176-03.2004

Sah, D. W., Ossipo, M. H., and Porreca, F. (2003). Neurotrophic factors as novel therapeutics for neuropathic pain. *Nat. Rev. Drug Discov*. 2, 460–472. doi: 10.1038/nrd1107

Siddall, P. J., Taylor, D. A., McClelland, J. M., Rutkowski, S. B., and Cousins, M. J. (1999). Pain report and the relationship of pain to physical factors in the first 6 months following spinal cord injury. *Pain* 81, 187–197. doi: 10.1016/S0304-3959(99)00023-8

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*. 3, Article3. doi: 10.2202/1544-6115.1027

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.

Sroga, J. M., Jones, T. B., Kigerl, K. A., McGaughy, V. M., and Popovich, P. G. (2003). Rats and mice exhibit distinct inflammatory reactions after spinal cord injury. *J. Comp. Neurol*. 462, 223–240. doi: 10.1002/cne.10736

Volders, P. J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res*. 41, D246–D251. doi: 10.1093/nar/gks915

von Schack, D., Agostino, M. J., Murray, B. S., Li, Y., Reddy, P. S., Chen, J., et al. (2011). Dynamic changes in the microRNA expression profile reveal multiple regulatory mechanisms in the spinal nerve ligation model of neuropathic pain. *PLoS ONE* 6:e17670. doi: 10.1371/journal.pone.0017670

Xu, J. T., Tu, H. Y., Xin, W. J., Liu, X. G., Zhang, G. H., and Zhai, C. H. (2007). Activation of phosphatidylinositol 3-kinase and protein kinase B/Akt in dorsal root ganglia and spinal cord contributes to the neuropathic pain induced by spinal nerve ligation in rats. *Exp. Neurol*. 206, 269–279. doi: 10.1016/j.expneurol.2007.05.029

Xu, X. J., Puke, M. J., Verge, V. M., Wiesenfeld-Hallin, Z., Hughes, J., and Hokfelt, T. (1993). Up-regulation of cholecystokinin in primary sensory neurons is associated with morphine insensitivity in experimental neuropathic pain in the rat. *Neurosci. Lett*. 152, 129–132. doi: 10.1016/0304-3940(93)90500-K

Zhang, W., Chen, Y., Liu, P., Chen, J., Song, L., Tang, Y., et al. (2012). Variants on chromosome 9p21.3 correlated with ANRIL expression contribute to stroke risk and recurrence in a large prospective stroke population. *Stroke* 43, 14–21. doi: 10.1161/STROKEAHA.111.625442

Zhao, X., Tang, Z., Zhang, H., Atianjoh, F. E., Zhao, J. Y., Liang, L., et al. (2013). A long noncoding RNA contributes to neuropathic pain by silencing Kcna2 in primary afferent neurons. *Nat. Neurosci*. 16, 1024–1031. doi: 10.1038/nn.3438

Zhong, J., Chuang, S. C., Bianchi, R., Zhao, W., Lee, H., Fenton, A. A., et al. (2009). BC1 regulation of metabotropic glutamate receptor-mediated neuronal excitability. *J. Neurosci*. 29, 9977–9986. doi: 10.1523/JNEUROSCI.3893-08.2009

Ziats, M. N., and Rennert, O. M. (2013). Aberrant expression of long noncoding RNAs in autistic brain. *J. Mol. Neurosci*. 49, 589–593. doi: 10.1007/s12031-012-9880-8

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: 11 February 2014; accepted: 24 April 2014; published online: 23 May 2014.
*Citation: Raju HB, Englander Z, Capobianco E, Tsinoremas NF and Lerch JK (2014) Identification of potential therapeutic targets in a model of neuropathic pain. Front. Genet. **5**:131. doi: 10.3389/fgene.2014.00131*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.
Copyright © 2014 Raju, Englander, Capobianco, Tsinoremas and Lerch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
GENETICS

REVIEW ARTICLE
published: 30 May 2014
doi: 10.3389/fgene.2014.00152

Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial

Nicolas Servant[1,2,3]*, Julien Roméjon[1,2,3], Pierre Gestraud[1,2,3], Philippe La Rosa[1,2,3], Georges Lucotte[1,2,3], Séverine Lair[1,2,3], Virginie Bernard[4], Bruno Zeitouni[1,2,3], Fanny Coffin[1,2,3], Gérôme Jules-Clément[1,2,3,4], Florent Yvon[1,2,3], Alban Lermine[1,2,3], Patrick Poullet[1,2,3], Stéphane Liva[1,2,3], Stuart Pook[1,2,3], Tatiana Popova[1,5], Camille Barette[1,2,3,6], François Prud'homme[1,2,3,6,7], Jean-Gabriel Dick[6], Maud Kamal[8], Christophe Le Tourneau[2,3,9], Emmanuel Barillot[1,2,3] and Philippe Hupé[1,2,3,10]*

[1] Institut Curie, Paris, France
[2] INSERM U900, Paris, France
[3] Mines ParisTech, Fontainebleau, France
[4] INSERM U932, Paris, France
[5] INSERM U830, Paris, France
[6] Institut Curie, Informatic Department, Paris, France
[7] Institut Curie, Sequencing Facility ICGex, Paris, France
[8] Institut Curie, Translational Research Department, Paris, France
[9] Department of Medical Oncology, Institut Curie, Paris, France
[10] CNRS UMR144, Paris, France

Edited by:
Enrico Capobianco, Center for Computational Science, University of Miami, USA

Reviewed by:
Enrico Capobianco, Center for Computational Science, University of Miami, USA
Bibhash Mukhopadhyay, Johnson & Johnson, USA

***Correspondence:**
Philippe Hupé, Plateforme de Bioinformatique, Unité 900: Institut Curie – INSERM – Mines ParisTech, UMR144: Institut Curie – CNRS, 26, rue d'Ulm, 75248 Paris Cedex 05, France
e-mail: philippe.hupe@curie.fr;
Nicolas Servant, Institut Curie – INSERM U900, Paris, France
e-mail: nicolas.servant@curie.fr

Precision medicine (PM) requires the delivery of individually adapted medical care based on the genetic characteristics of each patient and his/her tumor. The last decade witnessed the development of high-throughput technologies such as microarrays and next-generation sequencing which paved the way to PM in the field of oncology. While the cost of these technologies decreases, we are facing an exponential increase in the amount of data produced. Our ability to use this information in daily practice relies strongly on the availability of an efficient bioinformatics system that assists in the translation of knowledge from the bench towards molecular targeting and diagnosis. Clinical trials and routine diagnoses constitute different approaches, both requiring a strong bioinformatics environment capable of (i) warranting the integration and the traceability of data, (ii) ensuring the correct processing and analyses of genomic data, and (iii) applying well-defined and reproducible procedures for workflow management and decision-making. To address the issues, a seamless information system was developed at Institut Curie which facilitates the data integration and tracks in real-time the processing of individual samples. Moreover, computational pipelines were developed to identify reliably genomic alterations and mutations from the molecular profiles of each patient. After a rigorous quality control, a meaningful report is delivered to the clinicians and biologists for the therapeutic decision. The complete bioinformatics environment and the key points of its implementation are presented in the context of the SHIVA clinical trial, a multicentric randomized phase II trial comparing targeted therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer. The numerous challenges faced in practice during the setting up and the conduct of this trial are discussed as an illustration of PM application.

Keywords: precision medicine, clinical trial, bioinformatics, sequencing, oncology, SHIVA

INTRODUCTION

ERA OF PRECISION MEDICINE

Though physicians have always considered the individual characteristics of each of their patients, the term personalized medicine appeared recently to account for our new abilities to characterize each person biologically with genomic analysis, and to use this information to guide medical decision-making and deliver the best treatment to each patient. This concept is also referred to as genomic medicine, and other terms such as stratified medicine or targeted medicine are sometimes used interchangeably. A few years ago, the concept of P4 medicine was introduced with the idea of managing the patient's health instead of the patient's disease (Hood and Friend, 2011). As a matter of fact, the practice of medicine today is mainly reactive, i.e., the physician treats the patient's disease and little is done to prevent the occurrence of the disease. The P4 medicine considers a model of healthcare that is predictive (considering the genetic background of the individual and his/her environment), preventive (adapting lifestyle, taking prophylactic drugs), personalized (tailoring the treatment to the individual's unique features, such as the patient's genetic background, the tumor's genetic and epigenetic landscape, his/her life environment) and participatory (many options about healthcare

which require in-depth exchanges between the individual and his/her physician). P4 medicine therefore extends the concept of personalized medicine.

The term precision medicine (PM) is also frequently encountered in the literature to denote similar ideas, and generally refers to delivering the right drug at the right time to the right patient, by targeting specifically the molecular events that are responsible for the disease. We will use in this article the terminology PM defined as a customization of healthcare that takes into account individual differences among patients from prevention, diagnosis, prognosis, choice of the treatment and follow-up. PM combines the knowledge of the patient's characteristics with traditional medical records and environmental information to optimize health. PM does not only rely on genomic medicine but also integrates any other relevant information such as non-genomic biological data, clinical data, environmental parameters and the patient's lifestyle.

PM IN ONCOLOGY

In a special issue, the Journal of Clinical Oncology has focused on PM in oncology (Garraway et al., 2013) showing that this new era of medicine offers new perspectives to cure cancer. PM also raises numerous challenges including biobanking, bioinformatics and legal issues (Garraway et al., 2013; Meric-Bernstam et al., 2013; Overby and Tarczy-Hornoch, 2013; Suh et al., 2013). The intrinsic complexity of cancer and the variety of its forms (each tumor being genetically unique) designate this pathology as a prime target for PM approaches. Cancer is a disease caused by the accumulation of mutations occurring in critical genes (oncogenes and tumor-suppressor genes) and resulting in the alteration of key molecular pathways. Due to the genetic nature of cancer, the oncology research has largely benefited from the advances in high-throughput genomics technologies in order to decipher the molecular alterations involved in the tumorigenesis on one hand, and to help the clinician to tailor the therapy on the other (Tamborero et al., 2013). Molecular profiling based on genomics information from the tumoral DNA and constitutional DNA offers new insights into the prediction of the disease progression and the response to treatment for each individual patient. These approaches are based in particular on two dominant concepts: oncogene addiction and synthetic lethality. The first one, oncogene addiction, stipulates that some tumors rely on one particular oncogene for their survival and progression, and inhibiting this gene would therefore stop tumor growth; this is the magic bullet idea introduced by Paul Ehrlich in 1900. The second one, synthetic lethality, refers to the observation that the inactivation of a pair (or more) of genes might be lethal, whereas individual inactivation of any of these genes would not kill the cell. It offers an opportunity to selectively kill cancer cells, if they already present gene inactivation for one gene of the synthetic lethal pair, by targeting the second gene of the pair. A famous example is the synthetic lethality of *BRCA* and *PARP* genes, which is exploited by using *PARP* inhibitors for treating *BRCA* deficient breast cancer tumors. Both oncogene addiction and synthetic lethality are typical situations where targeted therapy should be the solution of choice.

The identification of genomic alterations used as biomarkers along with the emergence of molecularly targeted agents (MTAs) such as tyrosine-kinase inhibitors have promoted the development of PM in oncology. MTAs have proven their efficacy in some cancer subtypes and they provide new opportunities to treat the disease (see Dienstmann et al., 2013, for a review). The first MTA has been trastuzumab, which is a monoclonal antibody targeting the *ERBB2* receptor. This gene is amplified in 15–20% of patients with breast adenocarcinoma. Treating patients with locally advanced disease with trastuzumab for a year decreases by 50% the risk of recurrence (Piccart-Gebhart et al., 2005). Targeting the *BCR/ABL* fusion gene (i.e., the Philadelphia chromosome) with another MTA, imatinib, in patients with chronic myelogenous leukemia has dramatically improved their outcome (Druker et al., 2001). *BRAF(V600E)* mutation is frequently associated with melanoma, where it seems to play a critical role in the malignancy process and can be effectively treated using vemurafenib (Flaherty et al., 2010). *BRAF(V600E)* mutation has been also identified in multiple forms of advanced cancers such as colorectal or thyroid cancer (Cantwell-Dorris et al., 2011). It is generally accepted today that using MTA has great potential in the treatment of many types of cancer. Around 40 MTAs have been approved to date for the treatment of cancer and the development of new inhibitors is in progress. Developing new MTAs imply also to decipher new biomarkers among the large number of genomic alterations observed in tumors (mutations, amplifications, deletions, translocations, fusions and other structural variants). A large number of genomic alterations are passengers while very few are drivers. A subset of these drivers are actionable, i.e., have significant diagnosis, prognosis, or therapeutic implications in cancer, and a subset may also be druggable, i.e., targets for therapeutic development (Dancey et al., 2012). Classifying these genomic alterations into actionable and/or druggable is difficult and high-throughput screening techniques might help this classification. The possibility to search within each tumor the actionable/druggable alteration using high-throughput technologies opens the way to PM in the field of oncology.

HIGH-THROUGHPUT SCREENING TECHNOLOGIES FOR PM

During the last two decades, the advent of high-throughput technologies has allowed the genome-wide characterization of molecular profiles in tumors. Among the different techniques, the gene-expression microarrays have been widely used so far in particular to build signatures for diagnostic and prognostic purposes. These gene signatures are now proposed as clinical tools for some types of breast cancer, for example Agendia's 70-gene Agilent-based MammaPrint®, i.e., the Amsterdam Signature (van't Veer et al., 2002; van de Vijver et al., 2002), Veridex's 76-gene signature, i.e., the Rotterdam Signature (Wang et al., 2005; Foekens et al., 2006), Genomic Health's 21-gene RT-PCR-based Oncotype DX™ (Cobleigh et al., 2005; Hornberger et al., 2005) and a 41-gene expression set (Ahr et al., 2001; Molecular, Ahr et al., 2002). Ten years ago, next-generation sequencing (NGS) technology appeared. It has evolved so quickly that it is possible today to sequence a genome for a few thousand dollars

within a few days. Of note, the sequencing of the first human genome costed around 3 billion dollars and took more than 10 years to be completed in 2003. The ability to simultaneously sequence millions of short nucleic acid fragments in parallel in a very short time and at very competitive costs (Sboner et al., 2011) makes NGS a major tool in oncology (Tran et al., 2012). NGS will very likely replace microarrays in a near future both for research and clinical applications. The current NGS techniques allow the profiling of the transcriptome (RNA-seq), the genome (DNA-seq, exome-seq), the epigenome (bisulfite-seq), the identification of DNA-protein interactions (ChIP-seq) and the reconstruction of chromosome architecture (Hi-C). While some sequencing platforms are very suitable for research purposes, the long duration of runs as well as the cost of these instruments are clearly incompatible with a real-time application for clinical use (e.g., the HiSeq sequencer from Illumina which tends to become the reference for very high-throughput sequencing, requires approximately 11 days per instrument and per run to generate data). In response to these concerns, benchtop sequencers were introduced such as the MiSeqDx^{TM} from Illumina or the Ion Torrent^{TM} PGM from Life Technologies. Benchtop sequencers allow the sequencing of a few megabases in a couple of hours with a very high depth of coverage. Their relatively low cost and rapid turnaround time make them very suitable for clinical applications. In November 2013, the MiSeqDx^{TM} was the first sequencer obtaining clearance from the Food and Drug Administration for clinical use as this platform demonstrated its precision and reproducibility across instruments, users, days and reagent lots (Collins and Hamburg, 2013). Benchtop sequencers make it possible to sequence rapidly fractions of the genome (target-seq) like the coding regions or a subset of known genes or mutation hotspots. The target-seq offers the possibility to screen several hundred mutation hotspots located in tumor-suppressor genes and oncogenes using dedicated cancer panel kits. Thus, the target-seq techniques offer new opportunities for diagnosis and many laboratories are shifting from Sanger sequencing to NGS platforms in order to meet challenges in terms of throughput and turnaround time. As an example recent advances have been made in the screening of the *BRCA1* and *BRCA2* genes and the detection of germline mutation related to an increased risk of developing breast cancer (Bosdet et al., 2013; Tarabeux et al., 2013).

FRAMEWORK FOR PM IN ONCOLOGY

Precision medicine requires a strong interdisciplinary collaboration between several stakeholders covering a large continuum of expertise ranging from medical, clinical, biological, translational, technical, and biotechnological know-hows. **Figure 1** illustrates the different practitioners involved in the complex process, describes the data workflow starting from and coming back to the patient in order to tailor the therapy and shows the informatics and bioinformatics infrastructure supporting the workflow. To build the therapeutic decision, the most exhaustive data ranging from clinical to biological, environmental and family information (e.g., description of the tumor histology, list of previous treatments, family history, etc.) needs to be collected along a complex healthcare pathway. As the disease evolves, new experiments such as high-throughput screens (with microarray or NGS technologies for example) or biomarkers detection by immunohistochemistry (IHC) have to be performed to measure relevant biological information required to choose the best therapy. During the process, physicians (including different specialists such as surgeons, pathologists, radiation and medical oncologists, etc.), biologists, pharmacists, bioinformaticians, computational biologists, biostatisticians, informaticians, biobank managers, biotechnological platform managers, clinical research associates, and the technical staff will offer their expertise for the benefit of the patient. Different actors and cultures and a variety of miscellaneous constraints, including meeting the deadlines for results delivery, render the application of PM in daily clinical practice extremely challenging. Organizational aspects are therefore essential for the success of PM (Veltman et al., 2013). Downing et al. (2009) mentioned the importance of Electronic Health Record (EHR) and Clinical Decision Support (CDS) for care delivery due to the acceleration of knowledge discovery and its impact on the increasing number of possible clinical decisions. Development in CDS is required to handle the large heterogeneity of data and their complexity. The authors also pinpoint the fact that PM strongly depends on our ability to collect, disseminate and process complex information. Indeed, every stakeholder produces information during the healthcare pathway at different time points and in different places. The overall information needs to be gathered, integrated and summarized in a digest report to facilitate the therapeutic decision-making.

NEED FOR BIOINFORMATICS SOLUTIONS TO SUPPORT PM

The availability of high-throughput technologies dedicated to clinical applications makes it very attractive for cancer centers to use these new tools on a daily basis. However, establishing such a clinical facility is not a trivial task due to the aforementioned complexity of PM framework along with the overwhelming amount of data. Indeed, the field of oncology has entered the so-called big data era as the particle physics did several years ago. From the big data 4 V's perspective, data integration issue (i.e., merging heterogeneous data in a seamless information system) in oncology can be formulated as follows: a large *Volume* of patients' data is disseminated across a large *Variety* of databases which increase in size at a huge *Velocity*. In order to extract most of the hidden *Value* from these data we must face challenges at: (i) the technical level to develop a powerful computational architecture (software / hardware), (ii) the organizational and management levels to define the procedures to collect data with highest confidence, quality and traceability, and (iii) the scientific level to create sophisticated mathematical models to predict the evolution of the disease and risks to the patient. Obviously, an efficient informatics and bioinformatics architecture is definitely needed to support PM in order to record, manage and analyze all the information collected. The architecture must also permit the query and the easy retrieval of any data that might be useful for therapeutic decision in real-time thus allowing clinicians to propose the tailored therapy to the patient in the shortest delay. Therefore, bioinformatics is among the most important bottlenecks towards the routine

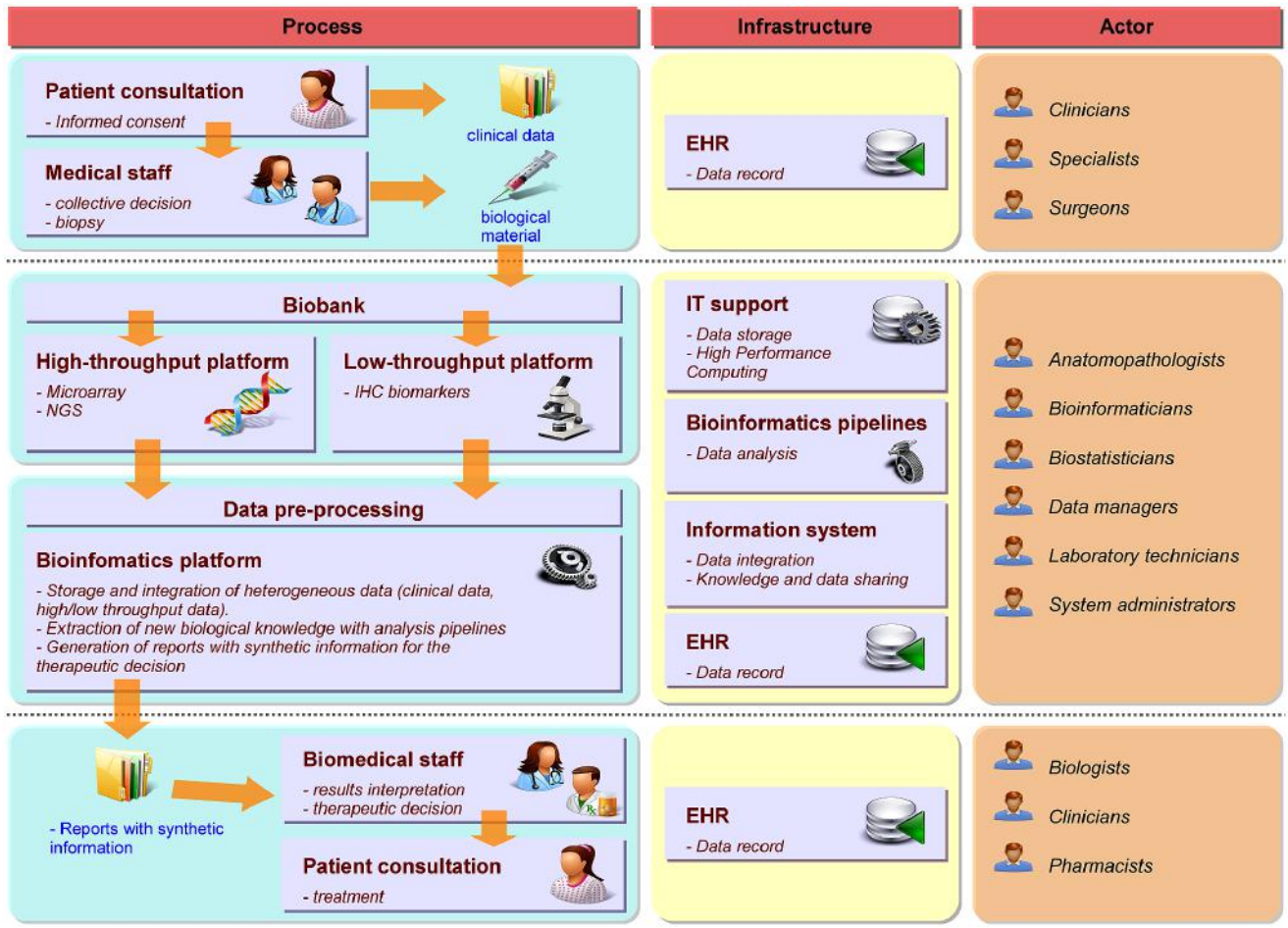


FIGURE 1 | Framework for PM in oncology. The left part describes the workflow and processes required for the decision-making from patient consultation to the therapeutic decision. The middle part focuses on the informatics and bioinformatics architecture required to support the different steps of the workflow. The right part indicates the different experts involved in each process.

application of PM and several challenges need to be faced to make it a reality (Fernald et al., 2011). First, the development of a seamless information system allowing data integration, data traceability, and knowledge sharing across the different stakeholders is mandatory. Second, bioinformatics pipelines need to be developed in order to provide relevant biological information from the high-throughput molecular profiles of the patient. Third, the architecture must warrant the reproducibility of the results.

If many recent publications point out the key role of the bioinformatics for PM today (see Simon and Roychowdhury, 2013 for a review), clinical trials usually do not detail the complete bioinformatics environment used in practice to assess the quality and the traceability of the generated data. Different software platforms such as transMART (Athey et al., 2013), G-DOC (Madhavan et al., 2011) or the cBio Cancer Genomics Portal (Cerami et al., 2012) have been recently developed to promote the data sharing and analysis of genomics data in translational research. Canuel et al. (2014) reviewed the different solutions available and compared their functionalities. One of the most interesting features of these platforms relies on their analytical functionalities. They provide ready-to-use tools through user-friendly interface offering interesting functionalities for data queries and user analysis. However, these different solutions do not address essential aspects which are offered by our system: first, often they handle a specific type of data; second they do not cover management and traceability of the data in real-time as long as they are generated by the different stakeholders; third they do not provide clinicians with a meaningful digest of the analyses, which they need to take clinical decisions.

In the next section, we will focus on the bioinformatics solutions implemented in order to tackle these challenges in the Institut Curie Bioinformatics platform in the context of the SHIVA clinical trial (Le Tourneau et al., 2012) initiated in October 2012 at Institut Curie (Paris, France). This trial provides a concrete and practical application of a PM project. First, we will describe the design of the SHIVA clinical trial. Second, the seamless information system we have implemented to manage data along with the bioinformatics pipelines used to deliver the results for the therapeutic

decision will be presented. Finally, the ongoing challenges will be listed.

DESIGN OF THE SHIVA CLINICAL TRIAL

SHIVA is a randomized proof-of-concept phase II trial comparing molecularly targeted therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer[1] (**Figure 2A**; Le Tourneau et al., 2012). Randomized trials in oncology are usually performed in a homogeneous population of patients with a specific tumor type and in a specific setting. In contrast, the goal of the SHIVA clinical trial is to bring the proof-of-concept that the prescription of molecularly targeted therapies

[1]http://clinicaltrials.gov/show/NCT01771458

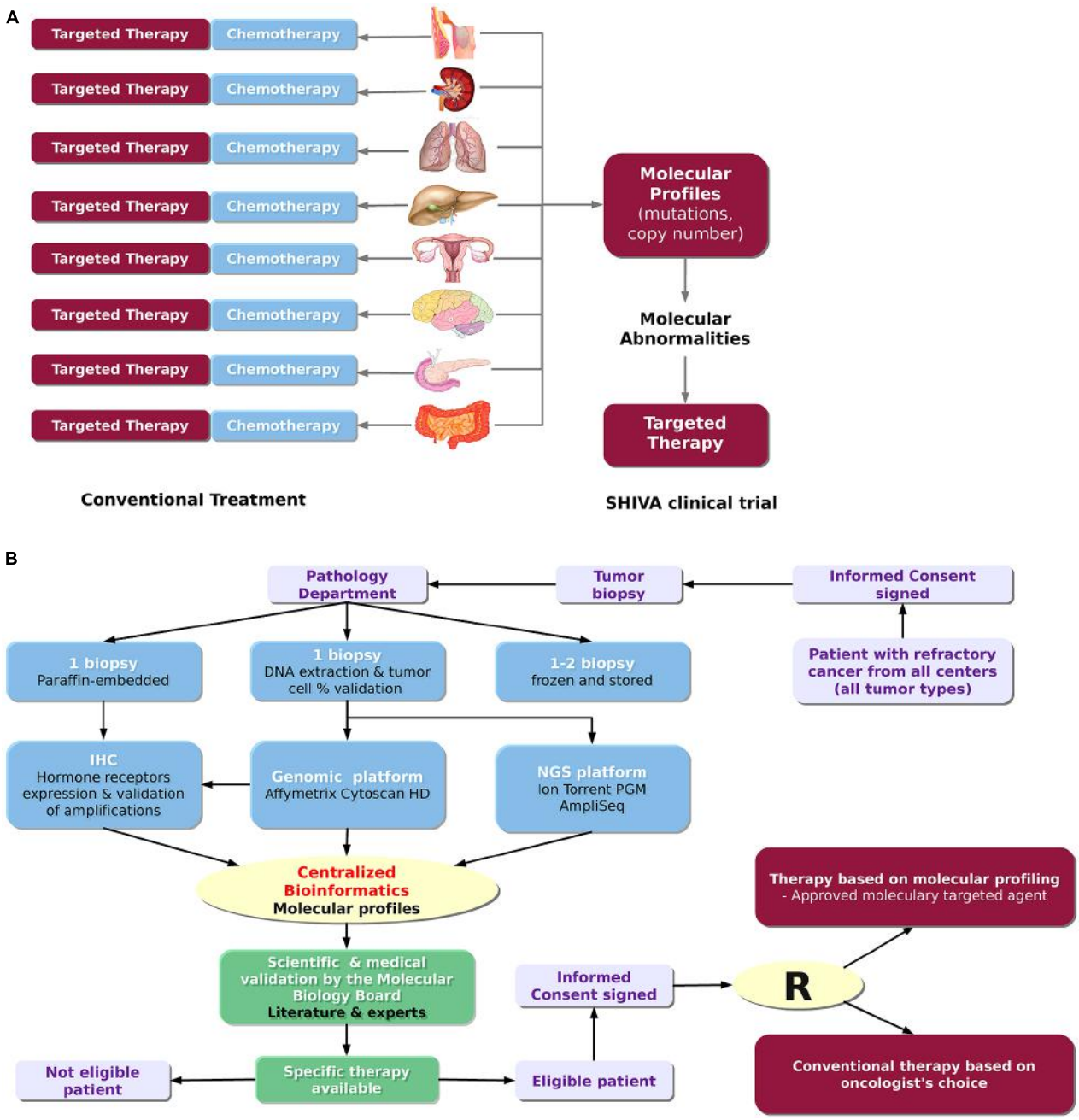


FIGURE 2 | The SHIVA clinical trial. (A) Rationale of the trial. The SHIVA trial aims at determining whether the prescription of molecularly targeted therapies based on tumor molecular abnormalities, independently of primary tumor location and histology, would improve the outcome of cancer patients. **(B)** Design of trial (adapted from Le Tourneau et al., 2012). The SHIVA clinical trial involves many different actors from the patient inclusion to the therapeutic decision. The whole process requires less than four weeks including 4 days for the bioinformatics treatment.

based on tumor molecular abnormalities, independently of primary tumor location and histology, would improve the outcome of cancer patients. Therefore, all tumor types are allowed in the trial (n.b. no more than 20% of patients with the same primary tumor location will be randomized). Both DNA copy number alterations and mutations in a subset of 76 genes are considered for the decision-making. These genes cover in particular three main signaling pathways: (1) the hormone receptors pathway, (2) the *PI3K/AKT/mTOR* pathway, and (3) the *MAP* kinase pathway. They include predictive biomarkers of efficacy of the MTAs as well as known biomarkers of resistance (e.g., *KRAS*). These predictive biomarkers had either been validated in the clinic (e.g., *ERBB2* amplification for anti-ERBB2 therapy, Piccart-Gebhart et al., 2005) or been supported by strong preclinical study (e.g., *PI3KCA* mutations for mTOR inhibitors, see Carew et al., 2011, for a review). Of note, not all of the 76 genes are of interest for the SHIVA trial but the whole panel includes mutations that might be of interest for non-randomized patients who may be eligible for clinical trials based on not yet approved MTAs. For each patient, a biopsy from the metastasis is performed and the molecular profiles are assessed using both the Cytoscan HD technology (Affymetrix) for the detection of DNA copy number alterations and loss of heterozygosity (LOH), and the Ion Torrent^{TM} PGM sequencing technology (Life Technology) for the detection of somatic mutations. IHC is used for the assessment of hormone receptor status, including estrogen, progesterone and androgen receptors, as well as for the validation of focal gene amplifications detected with Cytoscan HD in the following genes: *ALK*, *BRAF*, *EGFR*, *ERBB2*, *KIT*, *MET*, *PDGFRA*, *PDGFRB* and *PTEN*. Only samples which contain more than 30% of tumor cells are processed to control at best sample heterogeneity. Patients from seven hospitals in France can be included in the study. The establishment of the molecular profiles follows the process and the timelines described in **Figure 2B**. All the bioinformatics steps including data management and integration, molecular profile analyses and data coherence checking, are centralized at the Institut Curie bioinformatics platform. This centralization permits the analysis of the molecular data from the different hospitals using the same parameters therefore ensuring the reproducibility of results. The whole process was set up in real-time in order to have less than four weeks elapsed between the biopsy and the randomization, including 4 days for the bioinformatics treatment (**Figure 2B**). Thus, this trial represents a concrete application of PM. It highlights the real challenges and difficulties about the feasibility of such project in real-time. A committee of expert named the Molecular Biology Board (MBB) has been appointed. It consists of biologists, bioinformaticians and medical oncologists of each hospital. The MBB meets each week to decide what the best therapy is for each patient. Based on its scientific expertise and a literature review, the MBB has defined a set of rules taking into account the relevant molecular abnormalities identified in the tumor to decide which MTAs to choose (among a list of 11 drugs) to treat the patient. MTAs allowed in trial are only drugs that are approved for clinical use in France. In the next section, we will describe the bioinformatics solutions we have developed at Institut Curie to manage the data workflow for the SHIVA clinical trial.

BIOINFORMATICS ENVIRONMENT FOR THE SHIVA CLINICAL TRIAL

SEAMLESS INFORMATION SYSTEM

Precision medicine relies on a tight connection between many different stakeholders. As the choice of the therapy is based on a combination of different information levels including clinical data, high-throughput profiles (somatic mutations and DNA copy number alterations) and IHC data, all this information related to a given patient needs to be gathered in a seamless information system. Data integration is definitely required and bioinformatics plays a central role in setting up this infrastructure. To tackle this challenge, we have developed a seamless information system named KDI (Knowledge and Data Integration) described in **Figure 3**. The KDI system ensures information sharing, cross-software interoperability, automatic data extraction, and secure data transfer. In the context of the SHIVA clinical trial, high-throughput and IHC data are sent by the different biotechnological platforms to the bioinformatics platform using standardized procedures for transfer and synchronization. Data are then integrated into the KDI system within ad-hoc repositories and databases. Metadata describing the data are stored in the KDI core database such as the patient identifier, the type of data (e.g., mutation screening, clinical data, DNA copy number profile) and the technology used (e.g., Affymetrix microarray, Ion Torrent^{TM} PGM sequencing). Each type of data is then processed by dedicated bioinformatics pipelines in order to extract the relevant biological information such as the list of mutations and the list of amplifications/deletions. Therefore, the KDI core database acts as a hub allowing referencing all data through the use of web services. The KDI core database knows exhaustively which data is available for a given patient and where the raw and processed data are physically stored. It thus offers the possibility for clinicians to make queries through a web application and to extract the list of available information for a given patient. In addition, the system is also used to manage and perform automatic integrative analysis required for the therapeutic decision.

From a technical point of view, the KDI system consists of different modules dedicated to the storage, processing, analysis and visualization of each type of data (clinical, biological, microarray, NGS, etc.). High modularity associated with an efficient interoperability makes our system able to retrieve any relevant information. To facilitate the developments of these modules, we have retained a classical n-tiers architecture implemented with the JAVA/J2EE language. The core of each module of the KDI system can be presented as the association of different layers (**Figure 3B**).

Data layer

Data are stored in a relational database using the Entity-Attribute-Value (EAV) pattern. This conceptual modeling provides a data model plasticity required to handle the heterogeneity and the scalability of the variables of interest. Therefore, with EAV modeling, same concepts managed by different projects (with specific requirements by project) can be stored in a unique database without any modification of the data model. MySQL has been chosen as database provider for all web applications of the system.

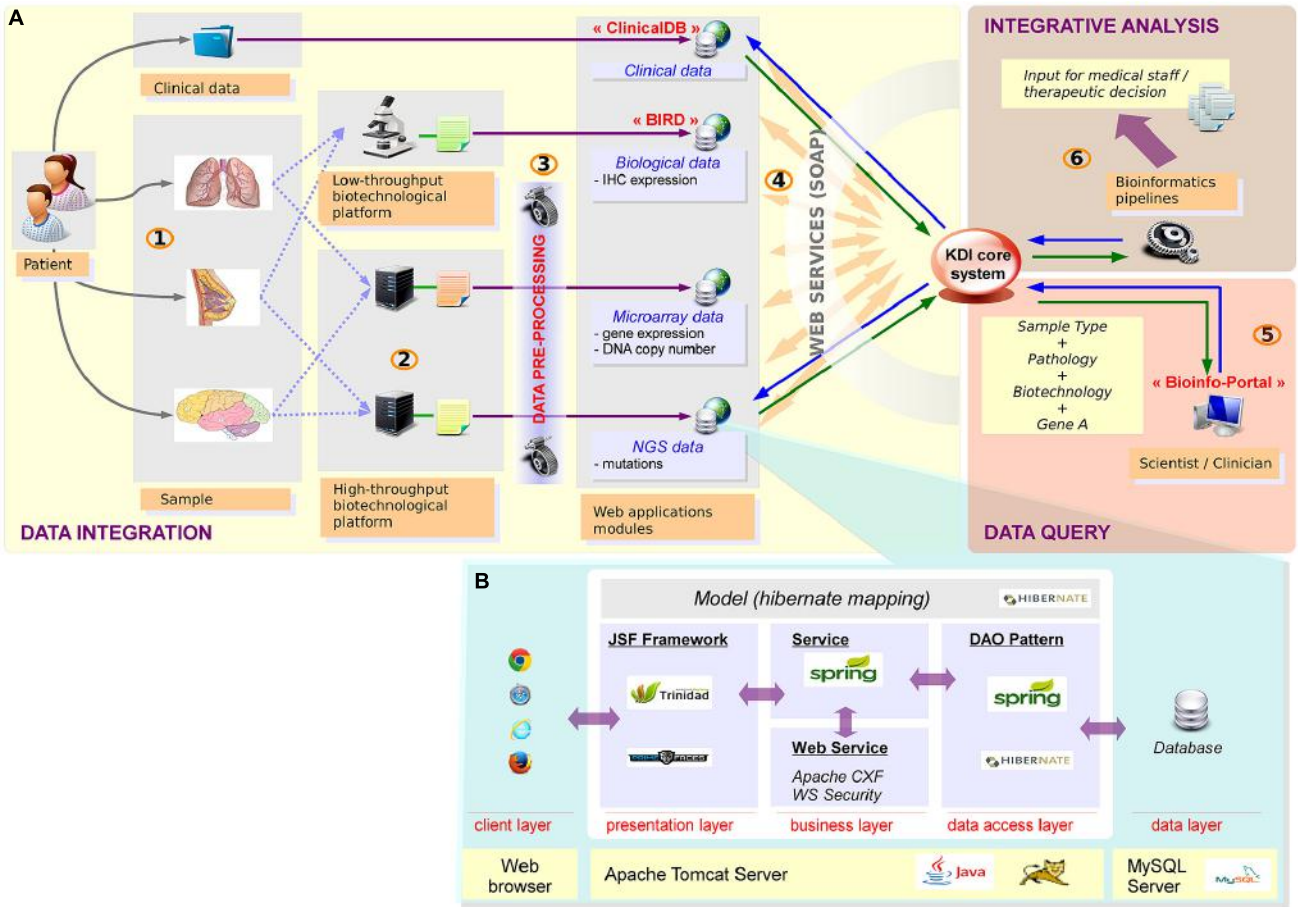


FIGURE 3 | Knowledge and data integration system. (A) Integration of heterogeneous data requires a seamless information system with high scalability, plasticity, reliability and interoperability. (1) The samples are first collected for a given patient. (2) The samples are then processed by the technological platforms (NGS, microarrays, immunochemistry, etc.) (3) Raw data are analyzed using dedicated bioinformatics pipelines. Results are stored in dedicated applications. (4) Sample data and the corresponding patient are referenced in the KDI core system. (5) Advanced research functionality enable multiple data queries. (6) Specific bioinformatics pipelines generate new integrative knowledge from heterogeneous sources of data. **(B)** Technical view of web applications: web applications are based on n-tiers architecture, developed with J2EE technologies.

Complementary solutions such as NoSQL databases are currently evaluated for particular requirements (ontologies storage, specific queries, etc.).

Data access layer

Data access is supported by the DAO (Data Access Object) pattern. By using HibernateDaoSupport superclass provided by Spring Framework, we promote the standardization of database access for all standard queries (findAll, findById, save, delete). Moreover, Hibernate mapping through JPA annotations associated with use of Hibernate Criteria provides a homogeneous frame for this critical layer. Database sessions and transactional aspects are also delegated to Spring Framework.

Business layer

Business core of our web applications has two main objectives: (i) provide structured data for presentation layer, and (ii) make data available for remote and secured access by other applications and technical users. Standard services are developed using core functionalities of Spring framework (Aspect-Oriented Programming - AOP, Inversion of Control - IoC, JavaBeans Factory). Web services are published (server side) and invoked (client side) through Apache CXF framework. To respect Web Services Security (WS-Security) standards, we use the Apache WSS4J project provided by CXF (with interceptors chain process) to set up a username token authentication on each web application in the system.

Front-end layer

Presentation layer is based on JSF (Java Server Faces) which is a component oriented framework for building user interfaces for web applications. To enrich the basic component set provided by JSF, we use additional component libraries such as Apache Trinidad and Primefaces. By this systematic approach for each user interface, we aim to build a visual identity, ergonomic, easily usable, for the whole information system. All data available within KDI can be browsed and retrieved from a user-friendly bioinformatics web portal.

Client layer

This layer represents the web browser through which end-users access KDI system.

DNA COPY NUMBER ANALYSIS PIPELINE

The use of the Affymetrix CytoscanHD microarray allows both the detection of DNA copy number alterations and the loss of heterozygosity events. The analysis workflow is presented in **Figure 4A**. Raw data are normalized with the Affymetrix Power Tools software package[2]. Then, the log R ratio is segmented in order to detect breakpoints and assign copy number status using Colibri (Rigaill, 2010) and GLAD (Hupé et al., 2004) software. A similar process is applied on the allele difference profile using the GAP software (Popova et al., 2009). Both profiles (DNA copy number and LOH) allow the estimation of absolute copy number for each probe taking into account the sample cellularity and tumor ploidy estimated by the GAP algorithm. Each gene status (normal, gained, amplified, lost, deleted, loss of heterozygosity) can then be assessed. Copy number alterations are defined as follows: deletion = 0 copy, loss = 1 copy, normal = 2 copies, gain = 3, 4 or 5 copies and amplification ≥ 6 copies for diploid tumor, and deletion = 0 copy, loss = 1 or

[2]http://www.affymetrix.com

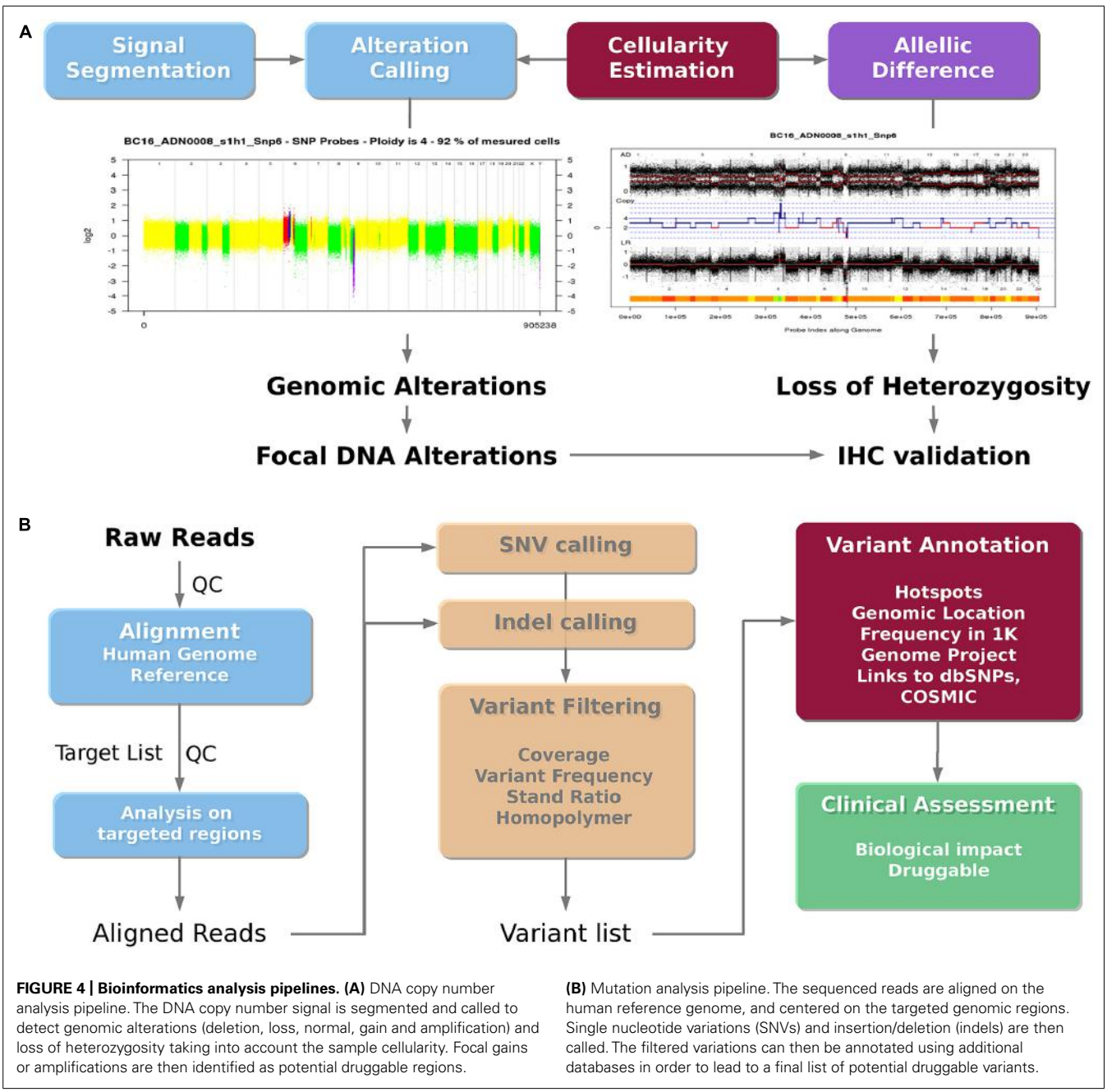


FIGURE 4 | Bioinformatics analysis pipelines. (A) DNA copy number analysis pipeline. The DNA copy number signal is segmented and called to detect genomic alterations (deletion, loss, normal, gain and amplification) and loss of heterozygosity taking into account the sample cellularity. Focal gains or amplifications are then identified as potential druggable regions. **(B)** Mutation analysis pipeline. The sequenced reads are aligned on the human reference genome, and centered on the targeted genomic regions. Single nucleotide variations (SNVs) and insertion/deletion (indels) are then called. The filtered variations can then be annotated using additional databases in order to lead to a final list of potential druggable variants.

2 copies, normal = 3 or 4 copies, gain = 5 or 6 copies and amplification ≥ 7 copies for tetraploid tumors. Additional steps in the analysis are performed to distinguish between large scale events such as chromosome arm gain and focal events targeting single oncogene or tumor-suppressor gene. Focal gains and amplifications are defined as genomic alterations with a size less than 10 Mb, and a copy number greater than the surrounding regions. In order to check whether a focal gain or an amplification of a size between 1 and 10 Mb induce a protein overexpression, a validation using IHC is performed. A report with the list of genes to be validated by IHC is automatically sent to the pathologists.

MUTATION ANALYSIS PIPELINE

The bioinformatics pipeline presented in the **Figure 4B** was applied to detect somatic mutations from the Ion Torrent^{TM} PGM sequencer using the Ampliseq^{TM} cancer panel. Ion Torrent^{TM} PGM raw reads are aligned on the reference human genome hg19 using the TMAP aligner (v0.3.7 Life Technologies). The best mapping score for each read is used to detect misalignment. The standalone package of the Torrent Variant Caller (v2.2 Life Technologies) is then used to call variants (SNVs and indels) from the mapped reads. In the context of clinical trial, variants have to be filtered to promote a high specificity, in order to avoid any false positive mutations. Thus, detected variants are filtered according to their frequency (≥4% for SNVs and 5% for indels), strand ratio (≥0.2), and reads coverage (≥30X for SNVs and 100X for indels). In addition, SNVs and mainly insertions and deletions detected in the context of a repeated region or a homopolymer are double checked. Homopolymer and repeated regions are prone to contain recurrent false positive, because of the limitation of the Ion Torrent^{TM} PGM technology. In most cases, the variant is discarded if also detected in other patients from the same sequencing run. Otherwise, variations specific to a sample, even within a repeat context, are reported. To facilitate the interpretation of individual patient data for clinical trials, the filtered list of variants is then annotated using the ANNOVAR software (Wang et al., 2010). Common polymorphisms found on more than 1% of the ESP or 1K Genome project population as well as recurrent and neutral variants on hotspots are reported. These variants do not present any therapeutic interest but are good internal controls to ensure the quality of the sequencing data. The Catalog of Somatic Mutation in Cancer (COSMIC) is used to annotate the mutations detected at a hotspot position. Non targeted mutations in genes covered by the panel, being non polymorphic nonsense, missense or indels are also reported, even if it may be difficult to know whether the alteration is involved in deregulating a particular pathway and whether it is clinically relevant. However, more stringent frequency filtering are applied for these cases (frequency ≥10% for SNVs and 15% for indels) leading to a higher specificity. Then, relevant mutations and variations are visualized using the IGV browser (v.2.0.35, Thorvaldsdóttir et al., 2013). The visualization remains an important step to assess the overall quality of the variant call, by taking into account the reads coverage, the error rate in the flanking region, the mutation position across the targeted region and across reads supporting them.

INTEGRATIVE ANALYSIS: THE REPORT FOR THE MOLECULAR BIOLOGY BOARD

The last step of the bioinformatics workflow is the production a technical report for the MBB. This task is crucial and must be complete and precise on one hand, and summarized on the other to allow a quick decision of the board. To answer this need, a report is generated for each patient. This report first presents the clinical information of the patient and the overall molecular profiles per gene, with the DNA copy number alterations, LOH status, and number of mutations (**Figures 5A,B**). This first section provides the MBB with a rapid overview of all detected alterations. If needed, the MBB can also have access to more detailed results, with graphical views of the copy number profiles for each gene, as well as the list of mutations with detailed annotation as previously described (**Figures 5C,D**). This name-blinded technical report is sent to the members of the MBB for scientific validation and prioritization of the identified molecular abnormalities.

SUMMARY OF THE DATA INTEGRATION WORKFLOW FOR THE SHIVA CLINICAL TRIAL WITHIN KDI

In the context of the SHIVA trial, the clinical data needed for the MBB are first imported in a dedicated module of the KDI system, named *ClinicalDB* (Clinical Database, **Figure 3A**). This step is performed weekly and updates the system by creating the patients recently included in the trial into the KDI core database. At the same time, an anonymous identifier is generated by the system for each new patient. Conversion between the different patient identifiers is guaranteed by the KDI core database and is accessible through the *Bioinfo-Portal* web application. Once available, the raw data generated by the biotechnological platforms are transferred to the bioinformatics platform for analysis (using rsync system). Bioinformatics pipelines (mutation and DNA copy number pipelines) process each molecular profile and are responsible for raw data storage and traceability within the KDI core database. The summarized results are structured in the *BIRD* (Biological Results Database) application. The last step of the data integration workflow is the generation of the bioinformatics reports. Two reports are required in the context of the SHIVA clinical trial at two different time points. A first report is generated by the system after the processing of the DNA copy number profile in order to request an IHC validation if needed. A second report is generated by the system and sent to the MBB for the final therapeutic decision. All reports, data and analysis results for each patient are gathered within the KDI modules (KDI core database, *ClinicalDB*, *BIRD*). All the information are available under controlled access for any member of the project through the KDI *Bioinfo-Portal*. In order to supervise the patients' process at each step of the whole bioinformatics workflow, an additional module of the system named the *Bioinfo-Board* application (**Figure 6**) has been developed. This web application aims to controlling, monitoring and checking the evolution and status of each SHIVA patient in real-time.

QUALITY MANAGEMENT

Offering a high quality service is most required in the context of a clinical application. The availability of all KDI's component and the reproducibility of the analyses is thus mandatory. To this

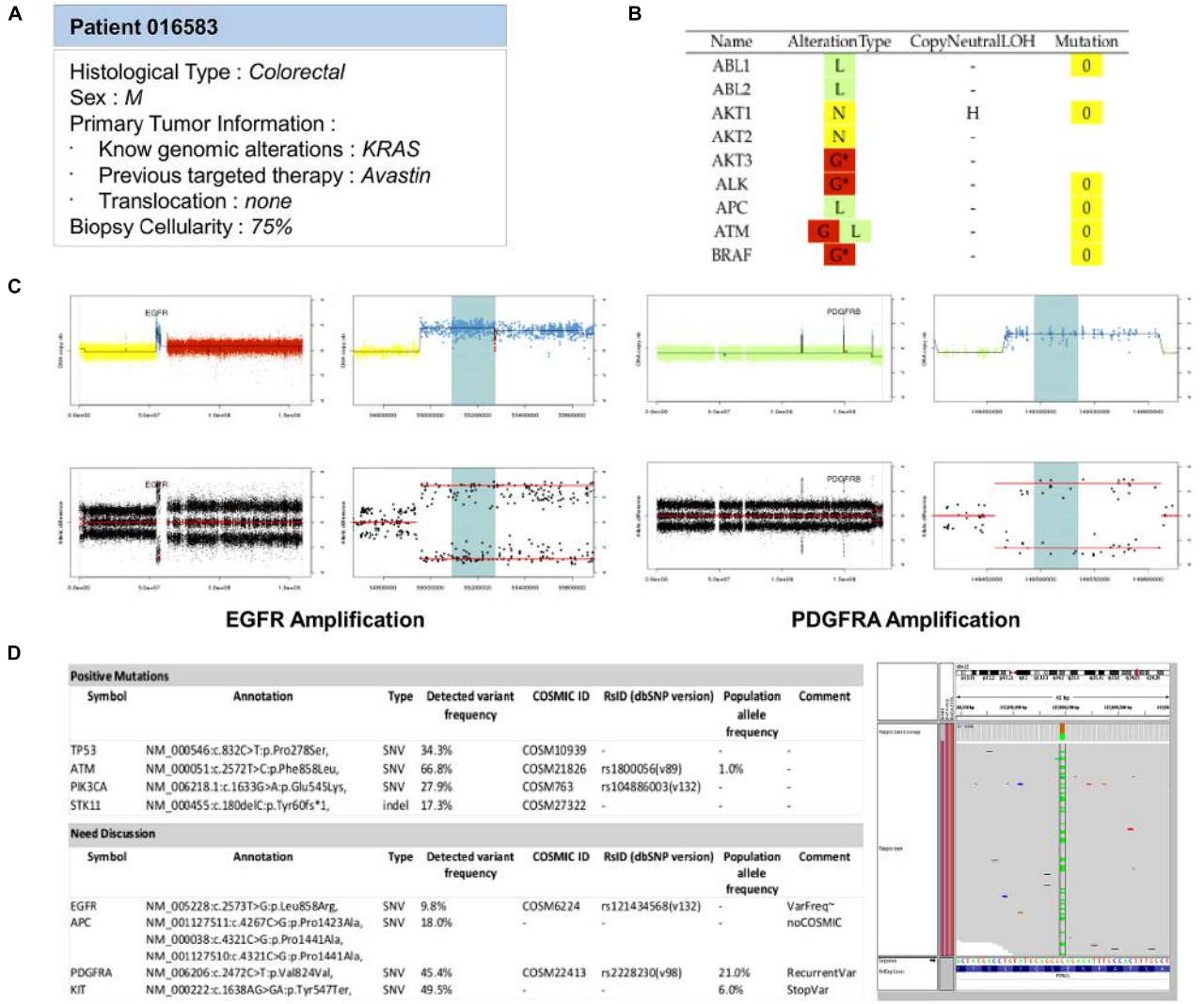


FIGURE 5 | Example of the main features reported in the MBB report. **(A)** Patient information. All available clinical and biological informations about the patient are reported to guide the therapeutic decision. **(B)** Summary of molecular abnormalities. DNA copy number and mutations statuses for all genes are reported in a main table in order to provide a quick overview of the potential targets. The DNA copy number (D = Deletion, L = Loss, N = Normal, G = Gain, A = Amplification) and the mutations statuses are color-coded to ease the interpretation. **(C)** DNA copy number status. For each gene, detailed informations of their DNA gene copy number and loss of heterozygosity statuses are provided. The profiles are shown for two patients carrying an amplification of EGFR and PDGFRA. **(D)** Mutation status. Mutations passing all the filters are reported as positives mutations. Others cases can be discussed. In the following example, one variant does not have any COSMIC ID and has a frequency lower than 10%. Another is annotated as a recurrent variant, or outside an hotspot region. An IGV screenshot of the genomic region can help in validating the variant.

aim, we have promoted a set of good practices for the software building process. First, the software development phase follows a strict frame with positive technical constraints, and a common methodology known and shared by each data manager and software developer. The configuration management is delegated to a SVN repository where all the source codes of KDI system are regularly committed. The unit testing is strongly recommended for all programming languages involved in the system (X-Unit) and part of our continuous integration server based on Jenkins software. This system allows to check weekly that all the tests parameterized for all applications are successfully passed. This control ensures that the analysis pipelines provide the expected results, identical to a reference analysis which is considered as a gold standard. Second, we pay attention to the availability of KDI system. All web applications are monitored with Nagios software in order to be able to detect in real-time any disorder on the system and therefore take immediately all necessary actions (log analysis, server restart, update configuration, etc.) to restore initial and nominal state if needed.

In order to reach this high quality service expectations, three different informatics environments (meaning three different instances of all applications, three different web servers, three different database servers and three dedicated file systems) have been set up. An update on any environment is always linked with

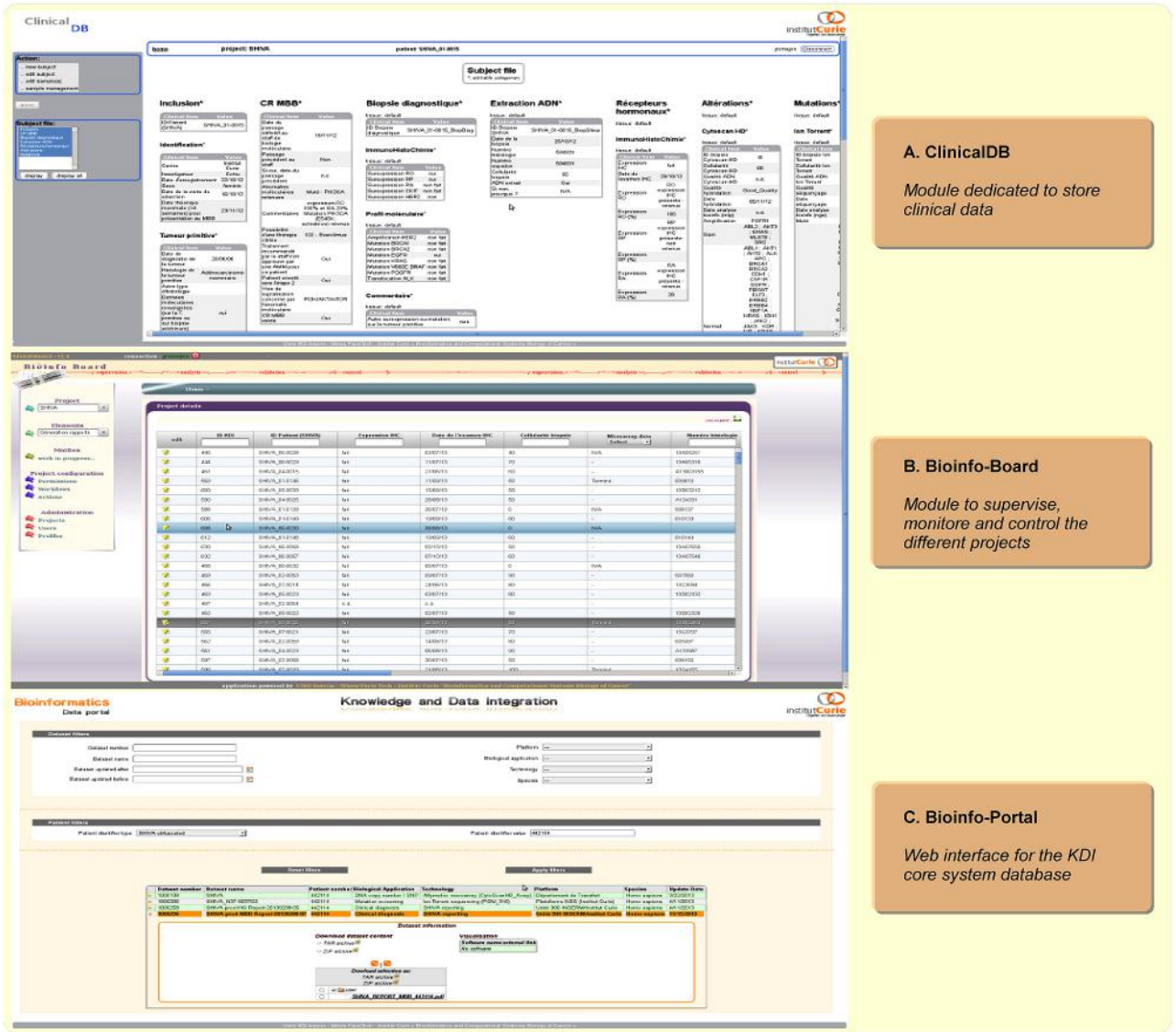


FIGURE 6 | Bioinformatics web applications screenshots. **(A)** *Clinical DB* application is dedicated to clinical data storage. **(B)** *Bioinfo-Board* application aims to manage, and monitor each patient data workflow in real-time. **(C)** *Bioinfo-Portal* application is dedicated to the KDI core system database. It allows the access to all exported data and information for a given patient/project, such as data or clinical report in the context of the SHIVA clinical trial.

a SVN revision. The first environment is the development (D-env) which is the place of the version currently in development. Each developer, after doing unit testing on his local workspace is allowed to install a new version of his components on this environment. It results that the D-env can be temporarily unstable and this is assumed. The second environment is the validation (V-env) which must be stable at every time. Integration testing is performed on this environment to validate the candidate release of the KDI system. The V-env can be seen as a pre-production environment. Only the persons in charge of the final installation are allowed to update the V-env. The third environment is the production (P-env) which is the instance of KDI system really used by the end-users. The updates of the P-env have to be planned, secured, and widely announced to avoid any inconvenience. During delivery periods, the environments of validation and production must be identical. These three environments permit to secure our delivery process with a high reliability and traceability.

FEASIBILITY OF THE SHIVA TRIAL

The presented bioinformatics environment is in use since October 2012 to manage and analyze the molecular profiles of the patients included in the SHIVA trial.

Results of the feasibility part of the project, focused on the first 100 patients were recently published (Le Tourneau et al., 2014). Among the first 100 patients, diagnostic confirmation and

IHC analyses for hormone receptors were performed in 92% of the patients. Genomic analyses were performed for 65 patients (68%). DNA copy number analyses met quality criteria in all the 65 patients, while a technical problem occurred in 2 patients for mutations analyses. Overall, 58 out of the 95 patients (61%) had a complete molecular profile. All patient data were integrated in the KDI system. The median timeframe for the bioinformatics analysis (DNA copy number, mutation profiles and MBB report) was 5 days. Median timeframe from tumor biopsy/resection to MBB was 26 days [range: 14–42]. To date, eight French cancer centers are participating to the SHIVA trial, and more than 700 patients were included. All data provided by the different centers are centralized at Institut Curie using the KDI system, and analyzed in routine.

ON-GOING CHALLENGES FOR PM

The solution we have developed to manage the SHIVA clinical trial provides a first step towards the routine application for PM. However, many challenges still need to be tackled and will require a lot of mutualization and harmonization efforts within the scientific community. The main on-going challenges are listed in what follows.

COMPUTATIONAL ARCHITECTURE

Precision medicine does not only require an efficient informatics infrastructure at the software level but also at the hardware level. Indeed, as NGS is now widely used for tumor profiling, data processing relies on an efficient High Performance Computing (HPC) infrastructure for data storage, transfer, computation and access control. So far, mainly targeted sequencing on a limited panel of genes (e.g., using Ion Torrent^{TM} PGM with AmpliSeq^{TM}) has been used and can be processed with relatively moderate computing resources. However, as sequencing cost keeps on decreasing, whole-exome or even whole-genome might be used soon thus requiring HPC infrastructure. According to Moore's law, Kryder's law and Butter's law, costs are halved every 18, 12, and 9 months for processor, storage and data transfer, respectively (Stein, 2010) while 5 months was the rule for sequencing costs during the period 2007-2011 period (source[3]). Thus, the difference between biotechnological and informatics capacities grows exponentially. Entering the era of big data in cancer research implies a breakthrough at the informatics level. First, the scalability of the infrastrure (Input/Ouput performance and computing power) is required to allow the management and analysis of ever-growing data. Second, bioinformaticians must be trained to the use of low-level programming languages for parallel computing such as Message Passing Interface (MPI), Open MultiProcessing (OpenMP), or MapReduce (Dean and Ghemawat, 2008) and to the algorithm analysis. Developing these new skills will be essential in order to improve the efficiency of software used in downstream analysis to deliver results as quick as possible to meet deadline expected in the clinical practice. Third, the configuration of job scheduler (such as Torque/PBS, OGE, or Slurm) must ensure that resources could be available and allocated to analyze in priority the data needed for decision-making in clinic. This also implies a redundancy of the hardware components to ensure their availability. Resources and new know-how are definitely needed to handle NGS data and PM. Importantly, the question of which data and how long the data must be stored is an important issue. We can anticipate that at some point, the storage capacity will be lower than the amount of data generated meaning that data will have to be analyzed on-the-fly to extract the relevant information and reduce the volume.

EXCHANGE STANDARDS AND ONTOLOGY

The large heterogeneity of the data that are collected along the healthcare pathway hampered their exchange and their comparison. Therefore, it is crucial to describe all the data that are generated with controlled vocabularies also called ontologies. Ontologies offer a formal representation of knowledge with definition of the relevant semantic attributes, their hierarchy and their relationship using a well-defined logic. Importantly, not only one single ontology can pretend to describe all the knowledge in a field but different ontologies (see[4]) are necessary to cover different entities of interest such as the gene (Gene Ontology), the disease (Disease Ontology) and the sequence (Sequence Ontology). Semantic Web standards promoted by World Wide Web Consortium (W3C) make it possible to link knowledge and data together so they can be queried and retrieved. To this aim, the Resource Description Framework (RDF) data format along with SPARQL query language provide the technical framework to describe, share, interact and query semantic data. While the technical solutions exist to support data exchange and linking, the definition of ontology, their choice and their use in practice for healthcare and biomedical data is still an issue. In order to tackle these challenges and to promote the use of standards and ontologies in the biomedical field, many European initiatives supported by the European Community (FP6 and FP7 programs) are involved in the definition and harmonization of standards:

1. *SemanticHEALTH* FP6[5] focused on semantic interoperability issues of electronic health systems and infrastructures and provided a number of relevant definitions, standards, and application domains for semantic interoperability (Stroetman et al., 2009).
2. *SemanticHealthNet* FP7[6] develops a scalable and sustainable pan-European organizational and governance process for the semantic interoperability of clinical and biomedical knowledge, to ensure that EHR systems are optimized for patient care, public health and clinical research across healthcare systems and institutions.
3. *p-medicine* FP7[7] aims at developing new tools, data sharing and integration systems, IT infrastructure and Virtual Physiological Human (VPH) models to accelerate PM for the benefit of the patient.

Moreover, the European effort BioMedBridges supported by the European Strategy Forum on Research Infrastructures (ESFRI)

[3]http://www.genome.gov/sequencingcosts/

[4]http://bioportal.bioontology.org/

[5]http://www.semantichealth.org

[6]http://www.semantichealthnet.eu

[7]http://p-medicine.eu

aim to construct the data and service bridges needed to connect emerging biomedical sciences research infrastructures. Among the infrastructures concerned let us mention:

1. the European Infrastructure for translational medicine: EATRIS supports the development of biomedical discoveries for novel preventive, diagnostic or therapeutic products up to clinical proof of concept.
2. the Biobanking and Biomolecular Resources Research Infrastructure: BBMRI will form an interface between biological specimens and data and top-level biological and medical research.
3. the European Clinical Research Infrastructures Network: ECRIN supports multinational clinical research projects in Europe.
4. ELIXIR: (it aims to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society.

DEVELOPMENT OF SUSTAINABLE BIOINFORMATICS ANALYSIS PIPELINES

Maintaining an efficient bioinformatics workflow in the context of PM is today challenging because of the frequent updates of the computational solutions either installed on the sequencing machine or provided as standalone applications. These frequent updates are mainly due to the rapid evolution of the sequencing and microarray technologies but remain a major issue to ensure the operability of the bioinformatics pipelines and their reproducibility. As a consequence, any update requires that each bioinformatics pipeline is validated to warrant it provides a very high specificity and sensitivity. Indeed, any changes in the data format or in the analysis methods can have critical consequences on the downstream analysis and results. Moreover, many different methods are currently available to analyze NGS data but no consensus or standard computational tools exist so far. For instance, detecting germline or somatic mutations can be achieved using different bioinformatics algorithms, tools and filters. Choosing the most efficient algorithm is not an easy task and a feasibility phase is mandatory to define which algorithms and parameters to apply for a dedicated question.

SAMPLE QUALITY CONTROL

The use of high-throughput technology in a clinical context also offers new challenges in the development of cutting edge statistical methods and algorithms dedicated to the field. As an example, the integration of heterogeneous molecular profiles provided by microarrays and sequencing assays could be used to define a patient genotype signature, to improve molecular profile accuracy and to ensure that the generated data come from the same biological samples and patient. The intersection of genotype variations available through the SNPs arrays technology could thus be intersected with the genotype information extracted from next-generation sequencing. However, this type of quality control requires the sequencing of a large DNA region to ensure that a sufficient number of polymorphism is covered. In the same way, the biopsy cellularity can also be estimated using both microarrays and sequencing assays (Larson and Fridley, 2013) in order to correlate the tumor purity from both profiles and detect intra-tumor heterogeneity.

DEVELOPMENT OF DEDICATED COMPUTATIONAL AND MATHEMATICAL METHODS - TOWARDS SYSTEM MEDICINE

Clinical trials for PM rely so far on a very limited number of biomarkers used for the therapeutic decision (see Simon and Polley, 2013 for a review). Typically from one up to less than 50 biomarkers are used for PM in currently on-going clinical trials worldwide. Moreover, the decision is based on a univariate decision rule meaning that a possible interaction between biomarkers is not considered which certainly explains part of the limited efficacy of targeted therapies even in the presence of their targets. For example, Prahallad et al. (2012) showed that vemurafenib is highly effective in the treatment of melanoma in patients with *BRAF(V600E)* mutation while colon cancer patients harboring the same *BRAF(V600E)* mutation have a very limited response to this drug. They found that *BRAF* normally exerts a negative feedback regulation of *EGFR*. Therefore *BRAF* inhibition causes a rapid feedback activation of *EGFR*, which enhances cell proliferation. As melanoma cells express low levels of *EGFR* they are not subject to this feedback activation in contrast to colon cancer. Thus, they propose that these patients might benefit from combined therapy consisting of *BRAF* and *EGFR* inhibitors. This example highlights the fact that considering interactions between biomarkers and combining different therapies together can dramatically strengthen the efficiency of PM. Also it clearly shows that elucidating the reasons behind treatment escape and proposing backup therapeutic strategies would benefit greatly from the knowledge and modeling of the cell regulatory network rewiring. Therefore, computational systems biology approaches, based on mathematical models of the cell regulatory network rewiring, are definitely needed to deepen our understanding of the cancer cell and to improve current decision rules. Systems biology and systems medicine are two disciplines which open the road to PM. Machine learning techniques will also be very useful to develop prediction rules to predict outcome and response to treatment. We can imagine that online machine learning techniques could be used to refine and optimize decision rules as long as new data and knowledge are generated. The key defining characteristic of online learning is that soon after the prediction is made, the true label of the instance is discovered. This information can then be used to refine the prediction hypothesis used by the algorithm. In the case of cancer, every day, for several patients, information is collected: survival, response to therapy, molecular profiles, pathological complete response, etc. This information could be used to retrain the classifier on the available data. In addition to these data-driven approaches, knowledge-based approaches must be developed to capitalize on the large amount of knowledge that is present in the scientific and medical literature to build efficient decision rules. IBM has developed a supercomputer named Watson (the name of IBM's founder) able to understand question in natural language and to extract relevant information from the literature. Watson supercomputer is currently used at the Memorial Sloan-Kettering (New-York, USA) to help for diagnosis in lung cancer.

SEQUENCING THE GENOME AND BEYOND

Available NGS techniques expand from sequencing panels based on a couple of genes to whole-exome and whole-genome sequencing. Even if the whole-exome and whole-genome sequencing are currently used in cancer research, and can be seen as the future of the clinical investigation, their use in routine clinical practice is much more difficult, mainly because the average depth of coverage is much lower than for targeted genes sequencing complicating mutations detection. However, these applications offer new ways to explore DNA copy number and structural variations and can thus be used as an alternative to the current microarray technologies. In addition, the current sequencing capabilities also offer new opportunities to develop gene/transcript expression and epigenomics biomarkers in clinic. For instance, the detection of *BRCA1/BRCA2* isoforms and their quantification using RNA-seq approach would be an interesting complementary approach to mutations screening. In the same way, DNA methylation, histone modifications, small non-coding regulatory RNAs, or nucleosome remodeling regulate many biological processes involved in tumorigenesis. More recently, evidence that genetic and epigenetic mechanisms are related events in cancer has emerged. Alteration in epigenetic mechanisms can lead to somatic mutations, as well as somatic mutations in epigenetic regulators can lead to an altered epigenome (You and Jones, 2012; Timp and Feinberg, 2013). If drug discovery in cancer epigenetics had been held back due to concern about specificity and toxicity, it remains an active field of investigation (see Dawson and Kouzarides, 2012, for a review). The application of these new fields in clinic raises the question of combined therapies. Combination of targeted therapy with chemotherapy or with other targeted therapies is challenging because of increased toxicity. Solutions include the use of lower doses of drugs which might not be relevant if the biologically active dose is not reached and the use of drugs in a sequential manner although the relevance of this approach still needs to be demonstrated. For instance, it is likely that the combination of standard chemotherapy together with drugs against mutated proteins and epigenetics drugs offer synergetic benefits and increase therapeutic efficacy. Integrative analysis considering the multidimensional nature of the cancer (genome, proteome, epigenome, kinome, etc.) is therefore a major challenge to unravel the complexity of the disease and identify the most efficient treatments. To this aim, we will have to capitalize on large collection of public datasets such as data from The Cancer Genome Atlas (TCGA[8], Kandoth et al., 2013) or International Cancer Genome Consortium (ICGC[9]) and also pathway databases for gene regulatory network, signaling pathway, metabolic pathway, Protein-Protein Interaction network and protein-compound network (e.g., DIP, HPRD, KEGG, Reactome to name only a few). The TCGA has initiated a pan-cancer analysis project (Cancer Genome Atlas Research Network et al., 2013) on the first 12 tumor types profiled by the consortium where the goal is to characterize molecular alterations and their functional impact across tumor type in order to promote the development of new therapies to fight cancer.

[8]http://cancergenome.nih.gov/

[9]http://icgc.org/

CONCLUSION

We have developed a seamless information system named KDI that fully supports the essential bioinformatics requirements for PM. The system allows management and analysis of clinical information, classical biological data as well as high-throughput molecular profiles. It can deliver in real-time information to be used by the medical and biological staff for therapeutic decision-making. KDI makes it possible to share information and communicate reports and results across numerous stakeholders, representing a large continuum of expertise from medical, clinical, biological, translational, technical and biotechnological know-hows. The system relies on state-of-the-art informatic technologies allowing cross-software interoperability, automatic data extraction, quality control and secure data transfer. KDI has been successfully used in the framework of the SHIVA clinical trial for more than 18 months. KDI is also currently used for other clinical trials supported by European Union consortia covering cancer (RAIDs - Rational molecular Assessments and Innovative Drugs selection in cervival cancer) and non-cancer applications (MAARS - Microbes in Allergy and Autoimmunity Related to the Skin). This demonstrates the potentiality and flexibility of our system to support PM covering all its requirements ranging from data management, data traceability, data analysis, query, and visualization.

The evolution of sequencing technologies has expanded the frontiers of genomics in both biology and clinical environments. The sequencing field will continue to evolve rapidly, offering lower costs and increased speeds. On-going developments in the sequencing technology, such as an ultrafast sequencer like nanopore technology, will improve performance and miniaturization, thus offering new tools to improve prevention, diagnosis, prognosis, choice of the treatment and follow-up for patients in oncology. To promote PM in daily clinical routine, flexible bioinformatics systems like KDI are definitely required for enabling efficient sharing of information in real-time, and rapid data processing needed for therapeutic decisions. KDI also provides the infrastructure for developing and integrating into the clinical decision process new integrative analysis methods with sophisticated mathematical models, representing the multidimensional nature of cancer to propose new biomarkers and to develop new therapies to fight cancer.

AUTHOR CONTRIBUTIONS

Nicolas Servant and Philippe Hupé coordinated the bioinformatics developments to support the SHIVA clinical trial. Philippe Hupé and Emmanuel Barillot coordinated the development for the seamless information system. Julien Roméjon, Philippe La Rosa, Georges Lucotte, Stéphane Liva, Alban Lermine, Virginie Bernard, Nicolas Servant managed the data and developed the bioinformatics pipelines for the SHIVA clinical trial. Virginie Bernard and Bruno Zeitouni developed the mutation pipeline for Ion Torrent^{TM} PGM. Pierre Gestraud, Philippe Hupé, Georges Lucotte, and Tatiana Popova developed the DNA copy number pipeline. Pierre Gestraud and Fanny Coffin developed the bioinformatics report for the molecular biology board. Philippe Hupé, Gérôme Jules-Clément, Florent Yvon, Patrick Poullet, Stéphane Liva, Alban Lermine, Stéphane Liva, Stuart Pook, Georges Lucotte,

Philippe La Rosa, Camille Barette, and Julien Roméjon developed the seamless information system KDI. Camille Barette, François Prud'homme, Jean-Gabriel Dick managed the informatics infrastructure. Christophe Le Tourneau is heading the Phase I Program as well as the Head and Neck Clinic at the Institut Curie (Paris, France). Christophe Le Tourneau is the principal investigator of the SHIVA randomized personalized medicine trial. Maud Kamal is the scientific coordinator of the SHIVA trial. Nicolas Servant, Philippe Hupé, Emmanuel Barillot, Pierre Gestraud, Maud Kamal, Christophe Le Tourneau and Julien Roméjon wrote the article.

ACKNOWLEDGMENTS

We would like to greatly thank Dr. Ivan Bièche, Dr. Céline Callens, Dr. Olivier Delattre, David Gentien, Dr. Thomas Rio-Frio, Dr. Gaelle Pierron, Dr. Etienne Rouleau, Dr. Xavier Poaletti, Dr. Claude Houdayer, Dr. Marc-Henri Stern, Eric Voirin, and Dr. Dominique Stoppa-Lyonnet from the Institut Curie (Paris France) for their fruitful advices. This work is supported by the grant ANR-10-EQPX-03 from the Agence Nationale de la Recherche (Investissements d'avenir) and grant INCa-DGOS-4654 SiRIC (Site de Recherche Intégré contre le Cancer) from the Institut National du Cancer. High-throughput sequencing has been performed by the NGS platform of the Institut Curie, supported by the grants ANR-10-EQPX-03 and ANR10-INBS-09-08 from the Agence Nationale de la Recherche (investissements d'avenir) and by the Canceropôle Ile-de-France. This project has received fundings from the European Union's Seventh Programme FP7-HEALTH under grant agreements No 304810 (RAIDS) and 261366 (MAARS).

REFERENCES

Ahr, A., Holtrich, U., Solbach, C., Scharl, A., Strebhardt, K., Karn, T., et al. (2001). Molecular classification of breast cancer patients by gene expression profiling. *J. Pathol.* 195, 312–320. doi: 10.1002/path.955

Ahr, A., Karn, T., Solbach, C., Seiter, T., Strebhardt, K., Holtrich, U., et al. (2002). Identification of high risk breast-cancer patients by gene expression profiling. *Lancet* 359, 131–132. doi: 10.1016/S0140-6736(02)07337-3

Athey B. D., Braxenthaler, M., Haas, M., and Guo, Y. (2013). tranSMART: an open source, and community-driven informatics, and data sharing platform for clinical, and translational research. *AMIA Summits Transl. Sci. Proc.* 2013, 6–8.

Bosdet, I. E., Docking, T. R., Butterfield, Y. S., Mungall, A. J., Zeng, T., Coope, R. J., et al. (2013). A clinically validated diagnostic second-generation sequencing assay for detection of hereditary BRCA1 and BRCA2 mutations. *J. Mol. Diagn.* 15, 796–809. doi: 10.1016/j.jmoldx.2013.07.004

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Cantwell-Dorris, E. R., O'Leary, J. J., and Sheils, O. M. (2011). BRAFV600E: implications for carcinogenesis and molecular therapy. *Mol. Cancer Ther.* 10, 385–394. doi: 10.1158/1535-7163.MCT-10-0799

Canuel, V., Rance, B., Avillach, P., Degoulet, P., and Burgun, A. (2014). Translational research platforms integrating clinical, and omics data: a review of publicly available solutions. *Brief. Bioinformat.* doi: 10.1093/bib/bbu006

Carew, J. S., Kelly, K. R., and Nawrocki, S. T. (2011). Mechanisms of mTOR inhibitor resistance in cancer therapy. *Target Oncol.* 6, 17–27. doi: 10.1007/s11523-011-0167-8

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Cobleigh, M. A., Tabesh, B., Bitterman, P., Baker, J., Cronin, M., Liu, M. L., et al. (2005). Tumor gene expression and prognosis in breast cancer patients with 10 or more positive lymph nodes. *Clin. Cancer Res.* 11, 8623–8631. doi: 10.1158/1078-0432.CCR-05-0735

Collins F. S., and Hamburg, M. A. (2013). First FDA authorization for next-generation sequencer. *N. Engl. J. Med.* 369, 2369–2371. doi: 10.1056/NEJMp1314561

Dancey, J. E., Bedard, P. L., Onetto, N., and Hudson, T. J. (2012). The genetic basis for cancer treatment decisions. *Cell* 148, 409-420. doi: 10.1016/j.cell.2012.01.014

Dawson, M. A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. *Cell* 150, 12–27. doi: 10.1016/j.cell.2012.06.013

Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 107–113. doi: 10.1145/1327452.1327492

Dienstmann, R., Rodon, J., Barretina, J., and Tabernero, J. (2013). Genomic medicine frontier in human solid tumors: prospects and challenges. *J. Clin. Oncol.* 31, 1874–1884. doi: 10.1145/1327452.1327492

Downing, G. J., Boyle, S. N., Brinner, K. M., and Osheroff, J. A. (2009). Information management to enable personalized medicine: stakeholder roles in building clinical decision support. *BMC Med. Inform. Decis. Mak.* 9:44. doi: 10.1186/1472-6947-9-44

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., et al. (2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* 344, 1031–1037. doi: 10.1056/NEJM200104053441401

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., and Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics* 27, 1741–1748. doi: 10.1093/bioinformatics/btr295

Flaherty, K. T., Puzanov, I., Kim, K. B., Ribas, A., McArthur, G. A., Sosman, J. A., et al. (2010). Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* 363, 809–819. doi: 10.1056/NEJMoa1002011

Foekens, J. A., Atkins, D., Zhang, Y., Sweep, F. C., Harbeck, N., Paradiso, A., et al. (2006). Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J. Clin. Oncol.* 24, 1665–1671. doi: 10.1200/JCO.2005.03.9115

Garraway, L. A., Verweij, J., and Ballman, K. V. (2013). Precision oncology: an overview. *J. Clin. Oncol.* 31, 1803–1805. doi: 10.1200/JCO.2013.49.4799

Hood, L., and Friend, S. H. (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat. Rev. Clin. Oncol.* 8, 184–187. doi: 10.1038/nrclinonc.2010.227

Hornberger, J., Cosler, L. E., and Lyman, G. H. (2005). Economic analysis of targeting chemotherapy using a 21-gene RT-PCR assay in lymph-node-negative, estrogen-receptor-positive, early-stage breast cancer. *Am. J. Manag. Care* 11, 313–324.

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 20, 3413–3422. doi: 10.1093/bioinformatics/bth418

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Larson, N. B., and Fridley, B. L. (2013). PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* 29, 1888–1889. doi: 10.1093/bioinformatics/btt293

Le Tourneau, C., Kamal, M., Trédan, O., Delord, J.-P., Campone, M., Goncalves, A., et al. (2012). Designs and challenges for personalized medicine studies in oncology: focus on the SHIVA trial. *Target Oncol.* 7, 253–265. doi: 10.1007/s11523-012-0237-6

Le Tourneau, C., Paoletti, X., Servant, N., Bieche, I., Gentien, D., Rio Frio, T., et al. (2014). Randomized proof-of-concept phase II trial comparing targeted therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer: results of the feasibility part of the SHIVA trial. *Br. J. Cancer* 1, 8. doi: 10.1038/bjc.2014.211

Madhavan, S., Gusev, Y., and Harris, M. A. (2011). G-CODE: enabling systems medicine through innovative informatics. *Genome Biol.* 12(Suppl. 1), P38. doi: 10.1186/gb-2011-12-s1-p38

Meric-Bernstam, F. F., Farhangfar, C., Mendelsohn, J., and Mills, G. B. (2013). Building a personalized medicine infrastructure at a major cancer center. *J. Clin. Oncol.* 31, 1849–1857. doi: 10.1200/JCO.2012.45.3043

Overby, C. L., and Tarczy-Hornoch, P. (2013). Personalized medicine: challenges and opportunities for translational bioinformatics. *Per. Med.* 10, 453–462. doi: 10.2217/pme.13.30

Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., et al. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N. Engl. J. Med.* 353, 1659–1672. doi: 10.1056/NEJMoa052306

Popova, T., Manié, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., and Stern, M. H. (2009). Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol.* 10:R128. doi: 10.1186/gb-2009-10-11-r128

Prahallad, A., Sun, C., Huang, S., Nicolantonio, F. D., Salazar, R., Zecchin, D., et al. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* 483, 100–103. doi: 10.1038/nature10868

Rigaill G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *Computation* arXiv, 10040887v1.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biol.* 12:125. doi: 10.1186/gb-2011-12-8-125

Simon, R., and Polley, E. (2013). Clinical trials for precision oncology using next-generation sequencing. *Personal. Med.* 10, 485–495. doi: 10.2217/pme.13.36

Simon, R., and Roychowdhury, S. (2013). Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* 12, 358–369. doi: 10.1038/nrd3979

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* 11:207. doi: 10.1186/gb-2010-11-5-207

Stroetman, V. N., Kalra, D., Lewalle, P., Rector, A., Rodrigues, J. M., Stroetman, K. A., et al. (2009). *Semantic interoperability for better health and safer healthcare.* Brussels: The European Commission. doi: 10.2759/38514

Suh, K. S., Sarojini, S., Youssif, M., Nalley, K., Milinovikj, N., Elloumi, F., et al. (2013). Tissue banking, bioinformatics, and electronic medical records: the front-end requirements for personalized medicine. *J. Oncol.* 2013, 368751. doi: 10.1155/2013/368751

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* 3:2650. doi: 10.1038/srep02650

Tarabeux, J., Zeitouni, B., Moncoutier, V., Tenreiro, H., Abidallah, K., Lair, S., et al. (2013). Streamlined ion torrent PGM-based diagnostics: BRCA1 and BRCA2 genes as a model. *Eur. J. Hum. Genet.* 22, 535–541. doi: 10.1038/ejhg.2013.181

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi:10.1093/bib/bbs017

Timp, W., and Feinberg, A. P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* 13, 497–510. doi: 10.1038/nrc3486

Tran, B., Dancey, J. E., Kamel-Reid, S., McPherson, J. D., Bedard, P. L., Brown, A. M. K., et al. (2012). Cancer genomics: technology, discovery, and translation. *J. Clin. Oncol.* 30, 647–660. doi: 10.1200/JCO.2011.39.2316

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999–2009. doi: 10.1056/NEJMoa021967

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancers. *Nature* 415, 530–536. doi: 10.1038/415530a

Veltman, J. A., Cuppen, E., and Vrijenhoek, T. (2013). Challenges for implementing next-generation sequencing-based genome diagnostics: it's also the people, not just the machines. *Personal. Med.* 10, 473–484. doi: 10.2217/pme.13.41

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet* 365, 671–679. doi: 10.1016/S0140-6736(05)17947-1

You, J. S., and Jones, P. A. (2012). Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* 22, 9–20. doi: 10.1016/j.ccr.2012.06.008

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: 02 December 2013; accepted: 08 May 2014; published online: 30 May 2014.
*Citation: Servant N, Roméjon J, Gestraud P, La Rosa P, Lucotte G, Lair S, Bernard V, Zeitouni B, Coffin F, Jules-Clément G, Yvon F, Lermine A, Poullet P, Liva S, Pook S, Popova T, Barette C, Prud'homme F, Dick J-G, Kamal M, Le Tourneau C, Barillot E and Hupé P (2014) Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. Front. Genet. **5**:152. doi: 10.3389/fgene.2014.00152*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.
Copyright © 2014 Servant, Roméjon, Gestraud, La Rosa, Lucotte, Lair, Bernard, Zeitouni, Coffin, Jules-Clément, Yvon, Lermine, Poullet, Liva, Pook, Popova, Barette, Prud'homme, Dick, Kamal, Le Tourneau, Barillot and Hupé. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
GENETICS

MINI REVIEW ARTICLE
published: 04 June 2014
doi: 10.3389/fgene.2014.00160

Targeting molecular networks for drug research

José P. Pinto[1], Rui S. R. Machado[1], Joana M. Xavier[1] and Matthias E. Futschik[1,2] *

[1] *SysBioLab, Centre for Molecular and Structural Biomedicine, Universidade do Algarve, Faro, Portugal*
[2] *Centre of Marine Sciences, Universidade do Algarve, Faro, Portugal*

Edited by:
Enrico Capobianco, Center for Computational Science – University of Miami, USA

Reviewed by:
Subha Madhavan, Georgetown University, USA
Alexey Goltsov, University of Abertay Dundee, UK
Zahraa Naji Sabra, American University of Beirut, Lebanon

***Correspondence:**
Matthias E. Futschik, SysBioLab, Centre for Molecular and Structural Biomedicine, Campus de Gambelas, Universidade do Algarve, CBME, FCT, Ed. 8, 8005-139 Faro, Portugal
e-mail: mfutschik@ualg.pt

The study of molecular networks has recently moved into the limelight of biomedical research. While it has certainly provided us with plenty of new insights into cellular mechanisms, the challenge now is how to modify or even restructure these networks. This is especially true for human diseases, which can be regarded as manifestations of distorted states of molecular networks. Of the possible interventions for altering networks, the use of drugs is presently the most feasible. In this mini-review, we present and discuss some exemplary approaches of how analysis of molecular interaction networks can contribute to pharmacology (e.g., by identifying new drug targets or prediction of drug side effects), as well as list pointers to relevant resources and software to guide future research. We also outline recent progress in the use of drugs for *in vitro* reprogramming of cells, which constitutes an example *par excellence* for altering molecular interaction networks with drugs.

Keywords: networks, molecular interactions, drugs, diseases, stem cells

INTRODUCTION

Over the last decade, we have witnessed impressive technological advances in the field of molecular biology. Many of them have brought us an incredible wealth of molecular data. Initially, it was hoped that large data-driven projects such as the Human Genome Project would readily pave the way for the development of new effective therapies in biomedicine. Unfortunately, the translation of these molecular data into biomedical breakthroughs has been dauntingly slow. Why is this so?

One reason for this "bottleneck" is that biological processes are highly interconnected, so their manipulation is a formidable challenge. In addition, major human diseases, such as cancer, type II diabetes, and hypertension, are genetically complex. Hence, a direct correspondence between causative genotype and disease phenotype, as observed in Mendelian disorders, is frequently obscure. Instead, these diseases are multi-factorial and seem to result from interplay between multiple genes and environmental factors, each having a relatively small effect, with few (if any) being prerequisites for the disease to occur (Manolio, 2010). This view is supported by several other lines of investigations that underline how important it is to regard causative genes not as isolated entities, but as integral parts of molecular networks or pathways (Badano and Katsanis, 2002; Oti and Brunner, 2007).

MOLECULAR NETWORKS: DATA AND ANALYSIS

In recognition of the importance of molecular networks, researchers from different fields have begun to study them intensely through computational and experimental means. Their underlying premise has been that changes to cellular networks determine many phenotypic variations, and that such changes can be provoked, not only by alterations to a gene product's abundance, but also through perturbations of its interactions.

The intensified interest in molecular networks has resulted in systematic gathering of interaction data for biomolecules, as well as the development of computational approaches for the analysis of biological networks. Nowadays, a large number of publicly accessible databases contain various types of molecular interaction data[1]. Networks derived from these resources frequently contain only a specific type of molecular interaction such a protein–protein or protein–DNA interactions. Based on the type of included interaction, we distinguish between different types of interaction networks. Currently, the major types are protein–protein interaction (PPI), gene regulatory and metabolic networks. These networks are often visually represented as simple graphs, with nodes or vertices denoting molecules, and links or edges denoting interactions between them. While such drastic simplification neglects many characteristics of individual components, it facilitates the analysis and modeling of large cellular networks. Furthermore, we can profit from the rich repertoire of mathematical tools and concepts already developed in graph theory.

The most basic characteristic of a node in a graph is its *degree*, i.e., the number of edges attached to it. In many biological networks, the majority of nodes have a low degree, and only a few nodes have a high degree. These highly connected nodes are known as hubs, and are important for the integrity of the network (Albert, 2005). Another important concept in graph theory is *modularity*. A module is commonly regarded as a set of nodes that are more densely connected with each other than with other nodes in the network (Pinto, 2012). These two concepts are illustrated for biological networks in **Figure 1A**. Modularity has also been suggested to contribute to *robustness* of molecular systems (Hartwell et al., 1999). In fact, robustness of molecular processes seems to result

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

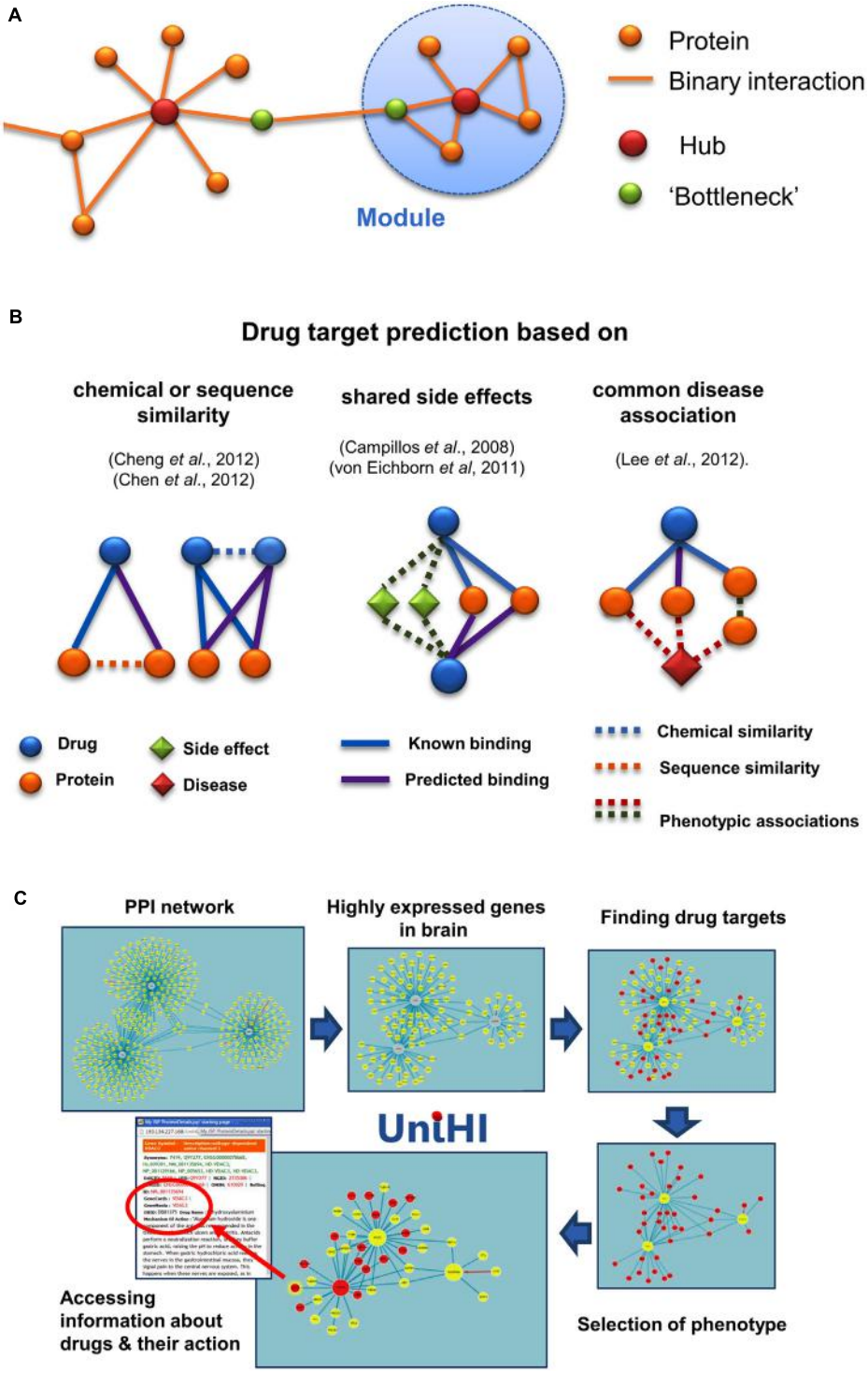


FIGURE 1 | (A) Illustration of basic concepts in the analysis of molecular networks. Hubs are defined by their large number of interactions, whereas "bottleneck" proteins link densely connected sub-networks or modules. Both types of nodes provide prominent targets for interventions, aimed at changing the network structure and integrity. **(B)** Approaches for network-based drug targeting and repositioning. Different types of heterogeneous bipartite or tripartite networks have been used in the literature to identify new targets for drugs. **(C)** Network-oriented pharmacology in the UniHI environment. After querying for molecular interactions for central proteins, UniHI derives tissue and phenotype-specific networks, which can be scrutinized for known drug targets. In the example shown, an interaction network with *GADD45A*, *SNCA*, *PARK2* as central proteins was retrieved and filtered using gene expression data from the brain. Additional filtering steps, using drug–target data and phenotypic information ("nervous system phenotype") from knock-out mice, generated a compact network of drug targets with potential relevance for neurological disorders. Information regarding the drugs and their mode of action can be interactively accessed within the displayed network.

directly from the structure of the underlying networks. Besides redundant genetic components, compensatory network structures such as alternative metabolic or signaling pathways can buffer the failure of single parts (Wagner, 2005). This feature of networks is a crucial aspect to be considered, when we want to design effective interventions in their functioning.

Prime examples of popular and freely available software for network analysis are R/Bioconductor[2] or Cytoscape[3]. While these are powerful and versatile tools, their use requires expertise in both data handling and processing. Alternatives are given by several on-line resources, which provide integrated and annotated data together with applications for analysis and visualization. For instance, our Unified Human Interactome (UniHI)[4] database stores a large number of molecular interactions for the human genome, together with other types of information, and includes tools for the interactive analysis of retrieved interaction networks (Chaurasia et al., 2007; Kalathur et al., 2014). Especially for researchers less acquainted with network analysis, such integrative platforms offer convenient gateways to a wealth of interaction data.

DRUGS AND THEIR TARGETS

Pharmaceutical drugs are a common means to modify the activity of biomolecules, making them prime candidates for altering activity and structure of molecular networks as well. The targets of drugs can be proteins, peptides or nucleic acids, whose activities can be modulated. Drugs can be sub-divided into at least three different classes: (i) chemical compounds with low molecular weight (typically referred to as small molecules) that target enzymes, receptors, transcription factors or ion channels; (ii) biologics (such as antibodies or recombinant proteins) that target extracellular proteins and transmembrane receptor; and (iii) nucleic acids that target messenger RNA by interference (Gashaw et al., 2011). Notably, small molecules are still by far the most common type of drugs, and are frequently associated with low costs and easy (i.e., oral) delivery. However, the number of proteins, which can be targeted by small molecules, appears to be fairly limited (Overington et al., 2006).

Ideally, drug targets should have: (i) a proven role in the pathophysiology of a disease; (ii) little impact on physiological (health) conditions when modulated; and (iii) a favorable prediction for potential side effects (Gashaw et al., 2011). To fulfill the later criterion, highly selective targeting is generally considered to be a desirable trait. To target multiple proteins, as is frequently required for treatment of complex diseases, it is therefore necessary to combine multiple drugs. Especially for cancer, combinatorial drug therapy has become a standard practice, minimizing the risk of drug resistance. However, kinase inhibitors, which target multiple pathways simultaneously, have shown efficacy in the treatment of different cancers (Al-Lazikani et al., 2012). Thus, it has been argued that multiple-target drugs might be a more favorable option, since detrimental drug–drug interactions can be avoided, and optimal dosage can be more easily determined (Hopkins, 2008).

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

[3]http://www.cytoscape.org

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

NETWORK-BASED APPROACHES FOR DRUG RESEARCH

IDENTIFICATION OF DRUG TARGETS

The identification of drug targets is a crucial, but laborious task in biomedical research. Nowadays, *in silico* methods can assist greatly. Conventional *in silico* methods for drug target prediction are typically receptor- or ligand-based models. Whereas receptor-based methods start with a known structure of the target, and employ docking to assess drug binding (Luo et al., 2011); ligand-based methods involve the comparison of drugs with known ligands of the target protein. A successful example of the latter method on a genomic scale is the study by Keiser et al. (2009), in which a large number of new potential targets for existing drugs were found based on chemical similarity with known ligands.

More recently, network-based methods have complemented the computational toolbox for drug target identification. They are especially helpful, if the three-dimensional structure of the target is unknown. Network-based methods are motivated by the observation that the general biological importance of a protein is at least partially linked to its location in relevant PPI networks. For instance, essential genes tend to correspond to hubs or central nodes in many PPI networks; although, in practice, such conclusions might be compromised by prevalent inspection biases (Futschik et al., 2007; Barabási et al., 2011). Consequently, drugs should target central nodes, when a lethal effect is intended, as it is the case, for example, in the treatment of cancer cells or pathogens (**Figure 1A**). In contrast, if a molecular process needs be adjusted, it might be preferable to target neighbors of central nodes (Csermely et al., 2013). This approach is consistent with observations that targets of approved drugs tend to have more connections on average than most proteins, but fewer connections than for those proteins that correspond to essential genes (Yildirim et al., 2007).

In addition to degree as a basic centrality measure, other more sophisticated local metrics, including bridging centrality and graphlet degree, have been proposed for the identification of drug targets in PPI networks (Hwang et al., 2008; Milenkoviæ et al., 2011). Alternatively, global network-based analyses can be used to provide cues for follow-up investigations. For example, a systematic review of major signaling pathways led to the conclusion that proteins involved in cross-talk between pathways, represent promising targets for drug (Korcsmáros et al., 2010).

While the study of the topology of PPI networks provides a valuable, general indication about the likelihood of finding drug targets; more specific predictions can be determined by evaluating local heterogeneous networks (**Figure 1B**). One of the first steps in this direction was taken in the work of Yamanishi et al. (2008), who transformed a bipartite network (in which two types of nodes form a network) of drugs and their known targets into a high dimensional composite "pharmacological feature space", where interacting drugs and targets were close to each other. New chemicals or targets could be mapped into this feature space, and drug–target interactions were predicted based on their spatial proximity. A simpler approach, based on diffusion of scores within the local bipartite network neighborhood, has recently been proposed. This approach outperformed predictions

based on interference using either chemical similarity of drugs, or sequence similarity of targets (Cheng et al., 2012). Although several of its predicted new targets of known drugs were successfully validated, a drawback of this simpler method is that it cannot be applied to novel drugs. This limitation can be overcome through integration of the drug–target network with drug–drug (based on chemical similarity) and target–target (based on sequence similarity) networks. In the study by Cheng et al. (2012), random walks on these integrated heterogeneous networks were simulated to connect drugs with potential targets. Using drug–drug connections, new drugs, for which no target is yet known, can be linked to proteins via drugs that have known targets.

Furthermore, the use of expression responses appears to assist in the process of drug target identification. Starting with a network of functional associations between proteins, Laenen et al. (2013) evaluated whether differential gene expression upon drug treatment can pinpoint the protein targeted by a drug. Strikingly, while the expression changes of the target itself was only moderately informative, integration of differential expression observed in the target's network neighborhood resulted in a drastic increase in prediction accuracy. However, it remains to be assessed, whether it is generally the case that expression of genes functionally related to a target is altered by its corresponding drug.

REPOSITIONING OF DRUGS

Closely related to drug target identification is the task of drug repositioning, i.e., finding new therapeutic uses for existing drugs (Tobinick, 2009). Since drug repositioning is based on known drugs, it provides an attractive shortcut to the lengthy development of new drugs. While the above mentioned approaches for drug target identification also can be applied to drug repositioning, several methods and software have been exclusively developed for this task. For instance, Mathur and Dinakarpandian (2011) proposed new possible disease–drug relationships through the analysis of affected biological processes. After identifying processes defined in Gene Ontology that were enriched by genes associated with a particular disease, drugs were linked to these processes, if they targeted central proteins of the PPI network representing these processes. Through comparing predicted disease–drug relationships with ones that had been reported in clinical trials, they found a statistically significant overlap. A similar, but more direct approach has been implemented in the PharmDB database, which integrates binary linkages between drug, proteins, and diseases (Lee et al., 2012). New targets of existing drugs are inferred using a method called Shared Neighborhood Scoring, which evaluates weighted connections between drug and disease nodes via their associated proteins in a tripartite network composite. An alternative software tool, which combines structural models with analysis of interaction profiles, is DRAR-CPI (Luo et al., 2011). This web-server compares the binding behavior of a candidate drug with a set of pre-determined drug–target interactions using a docking approach. Similar interaction profiles can indicate shared targets and common clinical application. The number of included reference targets for docking, however, is limited.

It is important to note, that the use of networks as computational tools is not necessary constrained to the representation of actual molecular interactions, but can be used to represent any kind of defined similarities or association between distinct entities. For instance, Iorio et al. (2010) derived a drug–drug network, where links between drugs indicated similar expression changes upon treatment; they exploited it both for drug target prediction, as well as repositioning.

ANALYSIS OF SIDE EFFECTS

Physiological side effects can be caused by binding of drugs to proteins ("off-targets"), in addition to their intended targets. As side effects are crucial factors in therapeutic applications, their accurate prediction is of eminent importance to avoid failure in drug trials. Notably, systematic recording of side effects represents a broad phenotying on the level of the human organism, providing valuable holistic information on the action of drugs. A unique resource, with this objective, is the SIDER database, which accumulates reported side effects for almost 1000 marketed drugs (Kuhn et al., 2010). Using this database, Mizutani et al. (2012) correlated a drug's side effects with the proteins it binds to. For this, side effects and bound proteins were represented as binary profiles and statistically associated using a modified version of canonical correlation analysis. The obtained correlation was used subsequently for the prediction of side effects, by evaluating the proteins that the drug binds to. Remarkably, it is equally possible to predict a drug's target based on its side effects. This relationship was originally explored by Campillos et al. (2008); they identified new targets of known drugs based on the similarity of their side effects with those of other drugs. There is now a database, which has implemented this approach, called PROMISCUOUS (von Eichborn et al., 2011). It enables the interactive exploration of an integrated network of drug, protein, and side effect nodes, and can be used to gain new insight into the drug's mode of action. Finally, side effects can also be indicative for drug–drug interactions, which are frequently of clinical relevance. It was recently shown that two drugs tend to interact, if their targets are in close proximity in a PPI network, or if they have similar side effects (Huang et al., 2013). Moreover, combining information on physical interaction of drug targets and recorded side effects improves the prediction accuracy for drug–drug interactions.

In **Table 1**, we provide a selection of publicly available databases and computational resources, which may be useful for the reader to initiate their own investigations in the field of network-based pharmacology.

NEW HORIZONS: *IN VITRO* REPROGRAMMING OF CELLS USING SMALL MOLECULES

In the network-based approaches described above, drugs mainly act within small sub-networks in order to "fix" or interfere with particular processes. This contrasts with their recent use in stem cell biology, where small molecules have been used to re-wire entire cellular networks. Their main object in this context is to convert (or reprogram) somatic cells, specific to an individual, into stem cells. These cells may eventually provide a personalized supply of tissue to replenish cells lost in

Table 1 | Publically available resource for network-based drug targeting and repositioning.

Resource	URL	Description	Reference
DRAR-CPI	http://cpi.bio-x.cn/drar/	Web server that derives and compares the interaction profile of a inputted drug with those of a library of drugs	Luo et al. (2011)
DrugBank	http://www.drugbank.ca/	Database containing detailed information for approved or experimental drugs and their targets	Knox et al. (2011)
DvD	http://www.ebi.ac.uk/saezrodriguez/DVD/	Add-on software packages for R and Cytoscape for drug repurposing using gene expression data	Pacini et al. (2013)
Mantra	http://mantra.tigem.it/	Computational on-line tool for analyzing the mode of action of a drug using its induced gene expression	Iorio et al. (2010)
PROMISCUOUS	http://bioinformatics.charite.de/promiscuous	Database for drug repositioning based on integrated PPI, drug–protein interactions, and side effects	von Eichborn et al. (2011)
SIDER	http://sideeffects.embl.de	Database containing side effects of marketed drugs	Kuhn et al. (2010)
Stitch	http://stitch.embl.de/	Database accumulating a large number of interactions between chemicals and proteins for various organisms	Kuhn et al. (2012)
UniHI	http://www.unihi.org	Web-based platform integrating human molecular interactions, gene expression, phenotypes, and drug target information (**Figure 1C**)	Kalathur et al. (2014)

degenerative diseases. Pioneering work led by Yamanaka showed that such conversion is possible through forced expression of merely four transcription factors using viral vectors (Takahashi and Yamanaka, 2006). The original combination of transcription factors used by Yamanaka comprises Octamer-binding transcription factor 4 (Oct4), Sex-determining region Y-box 2 (Sox2), Kruppel-like factor 4 (Klf4), and v-myc avian myelocytomatosis viral oncogene homolog (c-Myc). However, this approach suffers from low efficiency. Furthermore, the viral integration of exogenous transcription factors, in particular of oncogenes, such as Klf4 and c-Myc, is unlikely to offer a viable therapeutic option. Thus, efforts have been made by various groups to find small molecules that can boost reprogramming efficiency, as well as replace virally transduced transcription factors.

Two main classes of small molecules have been identified so far: (i) molecules that facilitate chromatin remodeling by inhibition of, e.g., histone deacetylase, and thereby increase the plasticity of cells (Huangfu et al., 2008); and (ii) molecules that block signaling events that induce differentiation. Examples of the latter class are inhibitors of extracellular signal-regulated kinases (ERKs) and glycogen synthase kinase 3 (GSK3; Silva et al., 2008). By combining these two classes of small molecules, it is even possible to replace all four transcription factors (Hou et al., 2013). A remaining challenge, however, is to determine the underlying molecular processes of chemically induced pluripotency. So far, only rudimentary models, which lack mechanistic details, have been proposed for the activation of key transcription factors by the applied molecules (Hou et al., 2013). Here computational methods for "reverse engineering" of gene regulatory networks can be very helpful. These methods aim to infer regulatory interactions from observed gene expression patterns and comprise a diverse set of statistical approaches such as regression, analysis of correlation or mutational information or Bayesian networks (Marbach et al., 2012). Usually, their application requires a large set of genome-wide expression measurements and might not scale up very well to the complexity of regulatory networks in higher eukaryotes. Nevertheless, a recent study identified successfully a novel regulator of stem cell differentiation through reverse engineering of gene regulatory networks from microarray expression data (De Cegli et al., 2013). We anticipate that such approaches as well as systems biology in general will help to establish a rational basis for creating chemically induced pluripotency.

PERSPECTIVES

Our review highlights several applications of molecular networks, in which they act as versatile interfaces between phenotypes and drugs. While these applications demonstrate the utility of network-based analyses, several major challenges still exist. Firstly, the quality and coverage of interaction data need to be improved and consolidated. Many interaction data sets suffer from both detection and selection biases, which limit their use (Futschik et al., 2007). Published drug target data also appear to be compromised by their low reproducibility (Prinz et al., 2011). Secondly, condition-specific networks need to be constructed, reflecting the dynamics of molecular processes, in contrast to the static nature of current models. In this way, it will be possible to study the effects of external and internal stimuli on network structure and function. Finally, the vast majority of available drugs target network nodes, disrupting the general activity of a specific biomolecule. Only a small number of drugs are directed towards specific interactions (Wells and McClendon, 2007). Such "link-directed" drugs, however, can provide a more precise means to modulate molecular networks.

In summary, network-based analyses offer new ways of studying targets and effects of drugs. Although challenges lie ahead, network models promise to be powerful and versatile tools in our quest to better understand and control molecular systems in health and disease.

ACKNOWLEDGMENTS

This work was supported by the Portuguese Fundação para a Ciência e a Tecnologia (BIA-GEN/116519/2010, SFRH/BPD/96890/2013, IF/00881/2013 and PEst-OE/EQB/LA0023/2013). We would like to thank Trudi Semeniuk for critically reading our manuscript and the reviewers for their valuable suggestions.

REFERENCES

Albert, R. (2005). Scale-free networks in cell biology. *J. Cell Sci.* 118, 4947–4957. doi: 10.1242/jcs.02714

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* 30, 679–692. doi: 10.1038/nbt.2284

Badano, J. L., and Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* 3, 779–789. doi: 10.1038/nrg910

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

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321, 263–266. doi: 10.1126/science.1158140

Chaurasia, G., Iqbal, Y., Hänig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.* 35(Suppl. 1), D590–D594. doi: 10.1093/nar/gkl817

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8:e1002503. doi: 10.1371/journal.pcbi.1002503

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* 138, 333–408. doi: 10.1016/j.pharmthera.2013.01.016

De Cegli, R., Iacobacci, S., Flore, G., Gambardella, G., Mao, L., Cutillo, L., et al. (2013). Reverse engineering a mouse embryonic stem cell-specific transcriptional network reveals a new modulator of neuronal differentiation. *Nucleic Acids Res.* 41, 711–726. doi: 10.1093/nar/gks1136

Futschik, M. E., Chaurasia, G., and Herzel, H. (2007). Comparison of human protein–protein interaction maps. *Bioinformatics* 23, 605–611. doi: 10.1093/bioinformatics/btl683

Gashaw, I., Ellinghaus, P., Sommer, A., and Asadullah, K. (2011). What makes a good drug target? *Drug Discov. Today* 16, 1037–1043. doi: 10.1016/j.drudis.2011.09.007

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature* 402(Suppl.), C47–C52. doi: 10.1038/35011540

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690. doi: 10.1038/nchembio.118

Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., et al. (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* 341, 651–654. doi: 10.1126/science.1239278

Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H., and Han, J.-D. J. (2013). Systematic prediction of pharmacodynamic drug–drug interactions through protein–protein-interaction network. *PLoS Comput. Biol.* 9:e1002998. doi: 10.1371/journal.pcbi.1002998

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A. E., et al. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat. biotechnol.* 26, 795–797. doi: 10.1038/nbt1418

Hwang, W. C., Zhang, A., and Ramanathan, M. (2008). Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin. Pharmacol. Ther.* 84, 563–572. doi: 10.1038/clpt.2008.129

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14621–14626. doi: 10.1073/pnas.1000138107

Kalathur, R. K., Pinto, J. P., Hernández-Prieto, M. A., Machado, R. S., Almeida, D., Chaurasia, G., et al. (2014). UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. *Nucleic Acids Res.* 42, D408–D414. doi: 10.1093/nar/gkt1100

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., et al. (2009). Predicting new molecular targets for known drugs. *Nature* 462, 175–181. doi: 10.1038/nature08506

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 39(Suppl. 1), D1035–D1041. doi: 10.1093/nar/gkq1126

Korcsmáros, T., Farkas, I. J., Szalay, M. S., Rovó, P., Fazekas, D., Spiró, Z., et al. (2010). Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics* 26, 2042–2050. doi: 10.1093/bioinformatics/btq310

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 6, 343. doi: 10.1038/msb.2009.98

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., and Bork, P. (2012). STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res.* 40, D876–D880. doi: 10.1093/nar/gkr1011

Laenen, G., Thorrez, L., Börnigen, D., and Moreau, Y. (2013). Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol. Biosyst.* 9, 1676–1685. doi: 10.1039/C3MB25438K

Lee, H. S., Bae, T., Lee, J. H., Kim, D. G., Oh, Y. S., Jang, Y., et al. (2012). Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Syst. Biol.* 6:80. doi: 10.1186/1752-0509-6–80

Luo, H., Chen, J., Shi, L., Mikailov, M., Zhu, H., Wang, K., et al. (2011). DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. *Nucleic Acids Res.* 39(Suppl. 2), W492–W498. doi: 10.1093/nar/gkr299

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176. doi: 10.1056/NEJMra0905980

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Mathur, S., and Dinakarpandian, D. (2011). Drug repositioning using disease associated biological processes and network analysis of drug targets. *AMIA Annu. Symp. Proc.* 2011, 305–311.

Milenkoviæ, T., Memiševiæ, V., Bonato, A., and Pržulj, N. (2011). Dominating biological networks. *PLoS ONE* 6:e23016. doi: 10.1371/journal.pone.0023016

Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y. (2012). Relating drug–protein interaction network with drug side effects. *Bioinformatics* 28, i522–i528. doi: 10.1093/bioinformatics/bts383

Oti, M., and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clin. Genet.* 71, 1–11. doi: 10.1111/j.1399-0004.2006.00708.x

Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nat. Rev. Drug Discov.* 5, 993–996. doi: 10.1038/nrd2199

Pacini, C., Iorio, F., Gonçalves E., Iskar, M., Klabunde, T., Bork, P., et al. (2013). DvD: an R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* 29, 132–134. doi: 10.1093/bioinformatics/bts656

Pinto, J. P. (2012). *Computational Tools for Large-Scale Biological Network Analysis.* Ph.D. thesis, University of Minho, Braga.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10, 712–712. doi: 10.1038/nrd3439-c1

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T. W., and Smith, A. (2008). Promotion of reprogramming to ground state pluripotency by signal inhibition. *PLoS Biol.* 6:e253. doi: 10.1371/journal.pbio.0060253

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676. doi: 10.1016/j.cell.2006.07.024

Tobinick, E. L. (2009). The value of drug repositioning in the current pharmaceutical market. *Drug News Perspect.* 22, 119–125. doi: 10.1358/dnp.2009.22.2.1343228

von Eichborn, J., Murgueitio, M. S., Dunkel, M., Koerner, S., Bourne, P. E., and Preissner, R. (2011). PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.* 39(Suppl. 1), D1060–D1066. doi: 10.1093/nar/gkq1037

Wagner, A. (2005). Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays* 27, 176–188. doi: 10.1002/bies.20170

Wells, J. A., and McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 450, 1001–1009. doi: 10.1038/nature06526

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240. doi: 10.1093/bioinformatics/btn162

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabási, A. L., and Vidal, M. (2007). Drug–target network. *Nat. Biotechnol.* 25, 1119–1126. doi: 10.1038/nbt1338

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: 05 April 2014; accepted: 14 May 2014; published online: 04 June 2014.
*Citation: Pinto JP, Machado RSR, Xavier JM and Futschik ME (2014) Targeting molecular networks for drug research. Front. Genet. **5**:160. doi: 10.3389/fgene.2014.00160*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.
Copyright © 2014 Pinto, Machado, Xavier and Futschik. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
GENETICS

METHODS ARTICLE
published: 28 July 2014
doi: 10.3389/fgene.2014.00237

Development and *in silico* evaluation of large-scale metabolite identification methods using functional group detection for metabolomics

Joshua M. Mitchell, Teresa W.-M. Fan, Andrew N. Lane and Hunter N. B. Moseley*

Department of Molecular and Cellular Biochemistry, Markey Cancer Center, University of Kentucky, Lexington, KY, USA

Edited by:
Enrico Capobianco, University of Miami, USA

Reviewed by:
Vangelis Simeonidis, University of Luxembourg, Luxembourg
Reza M. Salek, European Bioinformatics Institute, UK

***Correspondence:**
Hunter N. B. Moseley, Department of Molecular and Cellular Biochemistry, Markey Cancer Center, University of Kentucky, CC434 Roach Building, 800 Rose Street, Lexington, KY 40536-0093, USA
e-mail: hunter.moseley@uky.edu

Large-scale identification of metabolites is key to elucidating and modeling metabolism at the systems level. Advances in metabolomics technologies, particularly ultra-high resolution mass spectrometry (MS) enable comprehensive and rapid analysis of metabolites. However, a significant barrier to meaningful data interpretation is the identification of a wide range of metabolites including unknowns and the determination of their role(s) in various metabolic networks. Chemoselective (CS) probes to tag metabolite functional groups combined with high mass accuracy provide additional structural constraints for metabolite identification and quantification. We have developed a novel algorithm, Chemically Aware Substructure Search (CASS) that efficiently detects functional groups within existing metabolite databases, allowing for combined molecular formula and functional group (from CS tagging) queries to aid in metabolite identification without *a priori* knowledge. Analysis of the isomeric compounds in both Human Metabolome Database (HMDB) and KEGG Ligand demonstrated a high percentage of isomeric molecular formulae (43 and 28%, respectively), indicating the necessity for techniques such as CS-tagging. Furthermore, these two databases have only moderate overlap in molecular formulae. Thus, it is prudent to use multiple databases in metabolite assignment, since each major metabolite database represents different portions of metabolism within the biosphere. *In silico* analysis of various CS-tagging strategies under different conditions for adduct formation demonstrate that combined FT-MS derived molecular formulae and CS-tagging can uniquely identify up to 71% of KEGG and 37% of the combined KEGG/HMDB database vs. 41 and 17%, respectively without adduct formation. This difference between database isomer disambiguation highlights the strength of CS-tagging for non-lipid metabolite identification. However, unique identification of complex lipids still needs additional information.

Keywords: metabolomics, chemical adduct, chemoselection, Fourier transform mass spectrometry, isotope-edited NMR, common subgraph isomorphism, graph theory, functional group resolved metabolite databases

INTRODUCTION

Metabolomics is the comprehensive study of metabolomes, which comprise the entirety of metabolites interconverted by networks of chemical reactions in living systems that make life possible and can be regarded as the functional readout of the genome and proteome (Kaddurah-Daouk et al., 2008; Le et al., 2012). Most of these chemical reactions are catalyzed by protein enzymes that interconvert a vast array of metabolites in complex networks.

Metabolites are bioorganic compounds that range widely in size and chemical complexity from small compounds with a few atoms (e.g., glycerol, $C_3H_8O_3$) to more complex structures consisting of hundreds of atoms and multiple functionalities (e.g., monosialotetrahexosyl ganglioside $C_{77}H_{139}N_3O_{31}$). The ability to identify and quantify a wide range of metabolites is the first step in a systematic elucidation and modeling of metabolic networks. The next important step is the ability to track individual atoms of various metabolites through the metabolic network using isotopically enriched tracers (e.g., ^{13}C, ^{15}N, and/or ^{2}H labeled precursors) coupled with stable isotope-resolved metabolomics (SIRM), from which metabolic networks can be robustly reconstructed (Fan et al., 2009, 2010, 2011, 2012; Moseley et al., 2011; Le et al., 2012). From such studies, we can acquire system biochemical insights across a broad spectrum of biological and biomedical problems (Lane et al., 2011; Ramautar et al., 2013; Armitage and Barbas, 2014; Wood, 2014; Zhang et al., 2014).

Despite the increasing interest in studying the metabolomes of different organisms, the systematic detection, identification, and quantification of metabolites, i.e., metabolomics, remains a challenge, which limits meaningful interpretation of metabolic data. Metabolomics employs numerous analytical techniques for elucidating metabolite structures and quantification, principally mass spectrometry (MS), and nuclear magnetic resonance (NMR). These complementary structure-based techniques afford a wider coverage of metabolites and versatility of structure

determination, particularly in terms of isotopic enrichment patterns of metabolites in SIRM studies. For example, NMR is excellently suited for determining different position(s) of ^{13}C label(s) in given metabolites (i.e., isotopomers) whereas MS readily provides the number of ^{13}C atoms in a metabolite (i.e., isotopologs). Both types of structural information are required for robust reconstruction of metabolic pathways (Fan et al., 2012). The combination of NMR with high resolution high sensitivity FT-MS makes it possible to obtain molecular formulae of a large number of metabolites as well as isotopomer and isotopolog distributions (Pan and Raftery, 2007; Fan and Lane, 2008; Lane et al., 2008; Fan et al., 2012; Lorkiewicz et al., 2012).

The high volume of data produced by these instruments requires computational approaches for automated assignment of the spectra and to analyze the data in an accurate, meaningful, and timely fashion (Goodacre et al., 2004). Furthermore, the exceptionally high resolution and sensitivity of FT-MS allows for the detection of metabolites that have not yet been characterized, complicating peak assignment and analysis (Kind and Fiehn, 2006). Despite the extremely high resolution and mass accuracy of ultra-high resolution mass spectrometers, assigning a unique formula to most peaks remains a non-trivial problem. Only by utilizing isotope abundance and isotopolog data, which eliminates >95% of possible peak-formula mappings, can assignment of a peak to a unique formula or a small set of formulae be achieved (Kind and Fiehn, 2006). However, this approach fails when dealing with isotopically enriched metabolites in SIRM studies, where the natural abundance distribution no longer holds. The many more detectable mass isotopologs arising from each labeled metabolite demand even higher mass resolution and accuracy for isotope-resolved molecular formula determination, thereby making the existing assignment algorithms error-prone.

An equally difficult problem arises when the molecular formulae must be mapped to specific metabolites. This is typically done by referencing a database of interest and searching for entries that match the computed mass and/or formula for the mass peak of interest. For human metabolomics research, the Human Metabolome Database (HMDB) is a growing source for human-specific metabolite data (Wishart et al., 2009, 2013). The HMDB currently contains 40,427 entries for compounds observed in the human metabolome. In addition to the HMDB, the KEGG Ligand database also contains a large number of metabolite entries. Although not uniquely focused on human metabolism, the KEGG database currently contains 16,396 metabolic entries from a variety of species (Goto et al., 2002) and additionally numerous drug compound entries. The compounds from other species not yet observed in humans may provide possible hints as to the identity of observed, uncharacterized human metabolites or metabolites present in human tissue that derive, from external sources, like essential amino acids, sucrose, bacterial and plant products. For both databases, the entries are stored as variants of the MDL Molfile (.mol) format, a standard format for storing the chemical structure, atoms, bonds, ionization state, and stereochemical information needed to represent any given molecule (Dalby et al., 1992). However, database searching is ambiguous, as often any given formula can correspond to more than one entry. For example, using the MOLGEN isomer generator and the formula $C_{15}H_{12}O_7$, 788,000 distinct structures are generated (even with restrictions on allowed functional groups) (Benecke et al., 1995; Kind and Fiehn, 2006). Fortunately, MOLGEN represents all *possible* structures, not just those that exist in known metabolic networks. Nevertheless, the presence of isomers, known as mass isomers in MS, greatly complicates the use of metabolite databases for metabolite assignment by MS. To overcome this difficulty, additional information must be obtained to accurately assign metabolite mass spectra. Tandem MS is often used to obtain chemical substructure of a given metabolite via its fragmentation pattern. Unfortunately, the data produced by tandem-MS requires very complicated, predictive algorithms for metabolite assignment and differences in fragmentation patterns generated by different instruments, in algorithms used for data analysis, and in data interpretation hampers the reproducibility and accuracy of these methods (Nesvizskii et al., 2007).

Chemoselective adduct formation, i.e., CS-tagging, of metabolite functional groups, with subsequent detection by ultra-high resolution FT-MS and/or NMR provides additional sources of chemical structure information that could facilitate the unique assignment of metabolites. Isotopically enriched reagents can be designed to react with particular functional groups present in metabolites, such as carboxylate (Ye et al., 2009), carbonyl (Fu et al., 2011; Mattingly et al., 2012), amino (Guo and Li, 2009), and sulfhydryl (Gori et al., 2014). **Figure 1A** shows the carbonyl-selective aminooxy reagent for simultaneous MS and NMR chemical editing. The adducts formed can be selected by isotope editing techniques by NMR or in high resolution MS, and the tag further provides enhanced sensitivity for MS (cf. **Figure 1B**). The subset of metabolites that react must therefore contain the particular functional group (**Figure 1C**), which when combined with stable isotope labeling of the aminooxy reagent and detection by high mass accuracy and isotope edited NMR shift data can often identify the metabolites uniquely, especially resolving isomeric structures (cf. workflow in **Figure 1D**). This CS-tagging approach provides information that directly relates to chemical substructure, and can be combined with accurate mass and fragmentation patterns from tandem-MS methods. However, in order to efficiently use functional group composition information along with molecular formulae, metabolite databases with functional groups delineated are needed.

Identifying functional groups in existing metabolite databases provides a convenient way of creating such a functional group-resolved metabolite database. Fundamentally, this problem requires the identification of metabolite substructures that are identical to functional groups of interest and storing this information in a well-organized manner as part of each metabolite entry. CheckMol is a publically available program which can determine the presence and number of over 240 different functional groups in molfile files (Haider, 2010b). Since its introduction in 2003, CheckMol has remained the industry standard for detecting functional groups within chemical structures and is a component in several chemoinformatics packages. Although CheckMol is a powerful and reliable tool, it does not use a generalized method for searching for each functional group; rather the method used

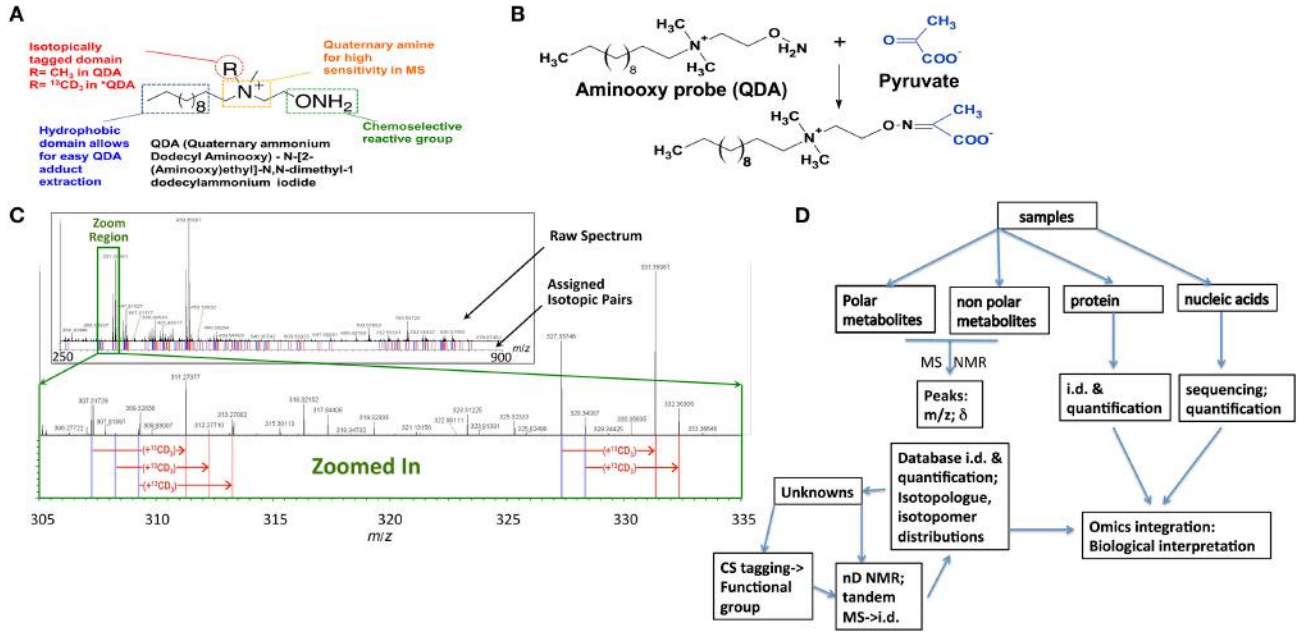


FIGURE 1 | Chemoselective tagging and large-scale detection of carbonyl-containing metabolites by FT-ICR-MS. (A) A chemoselective (CS) probe (QDA) for tagging carbonyl-containing compounds was designed to contain three different functionalities, i.e., an aminooxy group for specific reaction with carbonyls, a quaternary ammonium group for enhancing MS detection, and a hydrophobic domain for partitioning CS adducts into organic solvents to remove ionic matrix interferences. In addition, the R group in QDA can be isotopically labeled (such as $^{13}CD_3$ in *QDA) to facilitate automated assignment of carbonylated metabolites (Mattingly et al., 2012). **(B)** Reaction between pyruvate and QDA to form the oxime ether adduct is depicted. **(C)** A crude polar extract of human lung adenocarcinoma A549 cells was reacted with an equal mixture of QDA and *QDA before analysis by FT-ICR-MS. The top FT-ICR-MS spectrum shows an expanded spectral region from *m/z* 250 to 900, while the inset below (as depicted by the green box) more clearly illustrates the companion peaks of QDA and *QDA adducts. Blue and red lines (below the spectrum) respectively denote the *m/z* values of unlabeled and $^{13}CD_3$-labeled QDA derivatives, giving a "bar-code" profile of the adduct pairs (adapted from Mattingly et al., 2012). **(D)** Flow diagram from sample through to data integration. Samples from various sources are extracted into separate fractions of polar and non-polar metabolites and proteins (single step, Fan, 2012) and nucleic acids, either genomic DNA or mRNA (Fan et al., 2012). The metabolites are either separated by chromatography for MS (e.g., GC-MS, LC-MS) or subjected to direct infusion high resolution MS and NMR (Lane et al., 2008, 2009). The knowns are identified by comparison with standard databases (Fan and Lane, 2008; Lane et al., 2008). The unknowns may be identified using additional experiments including tandem MS and multidimensional NMR, especially if isotope enriched and where editing techniques can be used. Alternatively, the samples can be reacted with functional-group-specific reagents that introduce a tag that can be edited for by NMR, and imparts increased sensitivity in MS. This additional information is used to narrow down the possibilities in the database searches as described in the text.

for each functional group is unique and hard-coded. In order to add a new functional group to the list of functional groups searched for by CheckMol, a new method must be written in Pascal and then incorporated into the proper region in CheckMol, without introducing errors (Feldman et al., 2005).

To develop a tool that can search for a user-defined set of functional groups using a generalized strategy that does not require code modification, a natural choice is to abstract a molecule as a graph, in which the atoms are nodes and the bonds are vertices. The problem of detecting similarity between structures then is analogous to that of finding regions of similarity between the two graphs, called isomorphisms. This is the well-documented maximum common subgraph isomorphism (MCSI) problem in graph theory, for which several algorithms already exist, such as the Ullmann Algorithm (Ullmann, 1976). Also, graph theoretical approaches are widely used in chemoinformatics, notably to evaluate the structural similarity between compounds (Hattori et al., 2010) and to aid in the assignment of MS data (Hummel et al., 2010).

The Ullmann algorithm in its original form is unsuitable for our application as it implements a time-consuming brute force method for finding isomorphisms and lacks optimizations for isomorphism search in the context of chemical structures (Raymond and Willett, 2002). We have now implemented a novel algorithm loosely based on Ullmann's for finding subisomorphisms in database compounds that are completely isomorphic with a specific functional group. Our algorithm, called Chemically Aware Substructure Search (CASS), solves the subgraph isomorphism problem, which is NP-complete in computational complexity, but not NP-hard as in the case for MCSI. CASS utilizes a short-circuiting method to greatly accelerate the search for isomorphisms as well as a set of optimizations based on chemical structural rules. Although metabolite molfile files are readily available from KEGG and the HMDB, there is no database of functional group molfile files. We have hand crafted a database of 210 functional group molfile files using JChem which includes most of the functional groups searched for by CheckMol (Csizmadia, 2000). By applying our tools to both

KEGG Compound and HMDB, we have constructed a functional group-resolved database that combines the two databases into SQLite (Owens, 2006) relational tables. This database can be queried using the formulae detected by FT-MS along with CS-tagging to aid in the assignment of metabolites. Furthermore, additional chemical substructure information derived from either MS-MS analysis or NMR can be readily incorporated into the analysis by simply adding additional substructure molfile files for query.

MATERIALS AND METHODS

DATABASE ACCESS

Although both the HMDB and KEGG databases are publically accessible from web interfaces, local copies of the databases were needed for our analyses. The HMDB database was downloaded directly as a single SDfile (.sdf) file (i.e., flat file of concatenated molfile files with additional structured information) from the HMDB website. Like many sources of molfile files, the most recent versions of the HMDB contain additional structural and chemical information in each molfile file that is not specified in the original V3000 molfile file specification; therefore we developed a Perl script to handle these standard deviations from the molfile file specifications and create a specification compliant version. As the KEGG Ligand database is not available for download in any consolidated format, we developed a Python program that takes advantage of the KEGG REST interface to download molfile files (or kcf files) for each entry in the database and then concatenate them into a local copy of the KEGG database. The molfile files for KEGG entries do not contain database IDs nor compound names; these were collected from the KEGG database via its REST interface and added to the appropriate molfile file by our Python program.

Because we could not find a current functional group database that fit our particular design criteria (ability to specify both wild-carded and contextual atoms) (Kotera et al., 2008; Haider, 2010a,b; Eustis, 2011), we created one from scratch. To provide the same functionality as the existing CheckMol program, the list of functional groups detected by CheckMol was a natural starting point. For each functional group, the structure of the functional group was drawn by hand in JChem and the structures saved as molfile files. The molfile format designates each atom as a particular element. Therefore, we have developed a new nomenclature for describing these conditions. To designate that a particular atom could be one of several element types, the element type is designated as a list of possible element types separated by "|" while an "!" before an element type specifies the element type can be any element except the specified one (**Figure 2**). For example, "H|O|N" as an element type would specify that the atom could be hydrogen, oxygen or nitrogen while "!H" specifies that the element type can be any element type other than hydrogen. These descriptive facilities are more powerful than simple wild-carded "*" descriptive facilities available in other chemoinformatics tools (Daylight Chemical Information Systems, 2008).

Furthermore, to allow searching for a specific chemical substructure (e.g., −C=O or carbonyl) in particular chemical contexts (e.g., aldehydes or ketones), a way to designate atoms as "contextual" was added. Contextual atoms are designated with an asterisk after the element type and must be matched for a chemical substructure but are not considered as part of the substructure. For example, "C*" indicates a required element type of carbon that is not counted as part of the chemical substructure (**Figure 2**), for example, to distinguish between ketone and aldehyde carbonyls. To identify ketone carbonyls exclusively, the two carbon atoms bonded to the ketone carbonyl carbon atom are designated as contextual and therefore must be matched for the ketone carbonyl to be recognized but are not considered as part of the ketone carbonyl substructure. As a result, the carbonyl of an aldehyde, which is bonded to C and H, would not be recognized. The ability to designate contextual atoms in our chemical substructure descriptions is one of the main differences from previously published chemoinformatics toolkits that have substructure detection facilities. For example, while SMARTS allows for wild-carded atom designation (Daylight Chemical Information Systems, 2008) it does not allow for the designation of contextual atoms. This ability allows CASS to cleanly determine which atoms overlap between functional groups.

The functional group molfile files were concatenated to form a flat database similar to the downloaded copies of KEGG and the HMDB. Since flat files themselves provide no efficient means of searching for a particular entry and therefore must be parsed in their entirety, SQLite versions of these flat database files were created, to enable indexed entry retrieval. SQLite retains the simplicity and portability of flat files while offering the ability to search for entries in an efficient manner. Additionally, tools written in Perl were created to add a new entry to a SQLite database from a molfile file, return a particular molfile file from a database, and to check if a given entry exists in the database.

MOLFILE PARSERS

While molfile files accurately store chemical structures in a human-readable format, the structure of the molfile file format does not lend itself to computer manipulation and thus a more computer friendly internal format was needed. Toward this end, a molfile file parser was developed to convert molfile files into an internal representation shared among all of the programs. Due to differences between the formatting of KEGG and HMDB molfile files, different parsing methods are required for each database molfile file. Our parser can handle the molfile file variants used in both KEGG and HMDB as well as the proprietary.kcf format used in KEGG. This parser also handles the modified molfile file format used in our functional group molfile files via a parameter passed to the parser.

Regardless of the origin of the input molfile file, the final data structure generated by the parser is the same, a "molecule" object consisting of multiple data members representing different constituents and properties of a molecule. For each atom in the molfile file, an "atom" object data member is created that contains the element type, number of bonds to the atom, the sum of the bond order of all bonds to the atom and the index of the atom, which is its order in the list of atoms in the molfile file. Similarly, each bond has a corresponding "bond" object data member containing the indices of the two atoms it bonds and the order of the bond. Additionally, the molecule object contains the compound's database ID and name, a mathematical representation of

A mol file for Acyl Fluoride functional group

```
Beginning of SDF File FGroup032 Acyl Fluoride
  Mrv 0541 07231219352D

  4  3  0  0  0  0            999 v2000
    0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 !H  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 F  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  2  0  0  0  0
  1  4  1  0  0  0  0
M END
```

B mol file for Acyl Halide functional group

```
Beginning of SDF File FGroup033 Acyl Halide
  Mrv 0541 07231219352D

  4  3  0  0  0  0            999 v2000
    0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 !H  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 Cl|F|Br|I|At  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  2  0  0  0  0
  1  4  1  0  0  0  0
M END
```

C mol file for Carbonyl functional group

```
Beginning of SDF File FGroupX01 Carbonyl
  Mrv 0541 07231219352D

  2  1  0  0  0  0            999 v2000
    0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
M END
```

D mol file for Ketone Carbonyl functional group

```
Beginning of SDF File FGroupX02 Ketone Carbonyl
  Mrv 0541 07231219352D

  2  1  0  0  0  0            999 v2000
    0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C*  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C*  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
M END
```

FIGURE 2 | Example functional group molfile files. (A) The functional group molfile file for acyl fluoride demonstrates the use of the !X element type. The !H element type for atom 2 designates that it can be validly mapped to any non-hydrogen element type atom. **(B)** Similar to acyl fluoride, acyl halide uses the !H to designate a non-hydrogen element type. Additionally, since the halogen component of an acyl halide can be any halogen, the element type for the halogen atom is designated using the X|Y element type. **(C)** A typical functional molfile file. (**D**) The ketone carbonyl functional group uses contextual atoms to prevent matching of the molfile files to carbonyl-containing moieties that are not ketones.

its bonded structure, and optionally, a string representation of the molfile file from which it was generated.

In many database molfile files, implicit hydrogens are often excluded to reduce the size of the files. These implicit hydrogens must be added to the internal representation of each compound as the hydrogens could be included in a functional group of interest. We used standard molecular connectivity and valence methods to add the missing hydrogens (Weininger, 1988). This procedure does not account for pH or pK in these calculations and hydrogens are added to produce non-charged molecules unless the molfile file specifies otherwise (i.e., species that barely exist in practice). This procedure was validated by comparing known formulae for database compounds to computed formulae following hydrogen addition. Owing to the deviation of KEGG and HMDB molfile files from the molfile file standard, preexisting packages for manipulating molfile files could not be used and our own tool had to be created. These new tools add a variety of features in addition to adding implicit hydrogens and they support non-standard molfile files and KEGG compound files (.kcf), a molfile file derived file format used throughout the KEGG database.

ADJACENCY MATRIX REPRESENTATIONS

In order to use the graph theory algorithms in our substructure search program, numerical representations of each database's chemical structure are needed. The two common options for storing graph-like structures are adjacency lists and adjacency matrices. Although the list representation requires less memory than an adjacency matrix, matrices allow for direct testing of isomorphisms using very quick matrix comparisons and multiplication. In an adjacency matrix, each row and column corresponds to a specific node in a graph, or in this instance, an atom in a molecule (see **Figure 3**). The assignment of row or column to atom is done using the index of the atom. Row and column N is mapped to the atom with index N, therefore the first row and column both represent the first atom, the second row and column

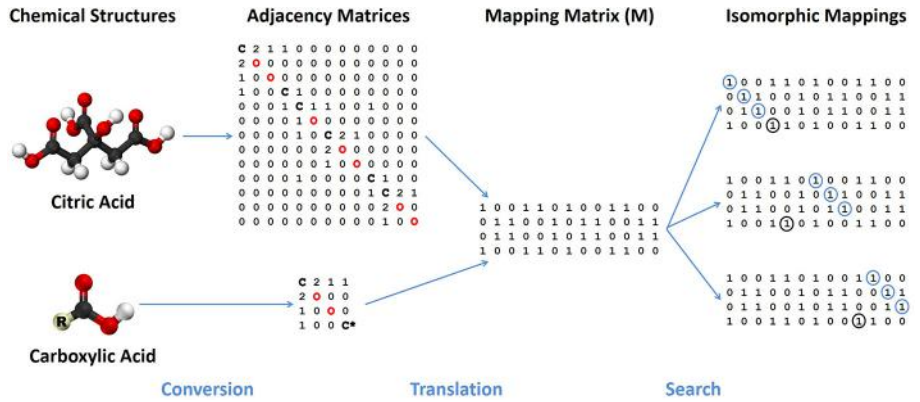


FIGURE 3 | Flowchart of major steps and representations needed by a chemical substructure search algorithm. Chemical structures of a specific compound and functional group are treated as colored graphs that are converted into adjacency matrices with elements identified along the diagonal. A mapping matrix (**M**) represents possible translations of specific atoms between the compound and functional group. The mapping matrix is searched for specific isomorphic mappings of the functional group mapped completely onto the compound chemical structure.

the second atom and so on. The value for an element (*i,j*) of the adjacency matrix corresponds to the presence or absence of a vertex connecting the two nodes, which correspond to chemical bonds between atoms. If the value of *i,j* is zero, no bond exists between the atoms.

To construct an adjacency matrix for a molecule with N atoms, our program first creates an N × N square matrix (**A**) with all $\mathbf{A}_{i,j}$ equal to zero. This saves a significant amount of time constructing the matrix as the entire matrix object is initialized at once and memory is already allocated for it. Second, as molecular graphs are often sparse (i.e., the number of possible vertices is much smaller than the maximum possible number of vertices), most of the values of $\mathbf{A}_{i,j}$ will be equal to zero. Thirdly, such a matrix can be created very efficiently utilizing functional programming methods which are heavily optimized in Perl. Not all values of $\mathbf{A}_{i,j}$ can remain zero, so for each bond object, the indices of the bonded atoms are retrieved along with the bond order and the corresponding values of $\mathbf{A}_{i,j}$ and $\mathbf{A}_{j,I}$ (as bonds are mutual) are set equal to the bond order. For example after processing, a double bond between atoms 2 and 4, $\mathbf{A}_{2,4} = 2$ and $\mathbf{A}_{4,2} = 2$. Once the adjacency matrix is constructed, they are stored as an object data member within the molecule object.

SUBSTRUCTURE SEARCHING

After the adjacency matrices for both the database compounds and the functional groups are constructed, our algorithm searches for isomorphic functional group substructures within the database compounds. The starting point for our algorithmic development was the Ullman algorithm (Ullmann, 1976). Owing to the presence of numerous "goto" statements in the original pseudocode, we converted this pseudo code into a control flow diagram (**Figures 4A,B**) and then into a modern control flow pseudocode representation (**Figures 5A,B**). We then deviated significantly from this new pseudocode representation during the development of our algorithm.

Given two graphs G_A and G_B representing the structure of a database molecule A and a functional group or a generic substructure query B and their corresponding adjacency matrices $\mathbf{A}_A$ and $\mathbf{A}_B$, the first step in both algorithms is the creation of a mapping matrix **M** (see **Figure 3**) with dimension b × a, where a and b are the number of atoms in A and B, respectively. It should be noted that a > b, as B must have fewer atoms than A, in order to be a subgraph of A. Each element of **M** is then assigned a value of 1 or 0. If $\mathbf{M}_{i,j} = 1$, the atom with index i in B can be "validly mapped" to the atom with index j in A and if $\mathbf{M}_{i,j} = 0$, no valid mapping can exist between the two atoms. In the traditional Ullmann algorithm, the definition of a valid mapping was determined by the number of vertices to the two nodes, i.e., valid mappings can only exist when the number of vertices to the jth point in A is greater than or equal to the degree of the ith point of B. Thus, the number of vertices "colors" the node and valid mappings are only allowed between nodes with the same or appropriate "color." Expressed in chemical terms, the jth atom in A must have an equal or greater number of bonds as the ith atom in B. By expanding the parameters that constitute a valid mapping, the total number of possible mappings that have to be tested can be minimized. In our algorithm, the element types of the two atoms are compared as well, set the corresponding $\mathbf{M}_{i,j}$ equal to zero ($\mathbf{M}_{i,j} = 0$), if the element types do not match. Here our expanded element types used in the functional group molfile files is important, as !X could map to an atom not element X ($\mathbf{M}_{i,j} = 1$), X|Y|Z could map to an atom of element type X, Y, or Z ($\mathbf{M}_{i,j} = 1$). As we are searching for complete instances of the functional group B as a substructure of A, every atom in B must have at least one valid potential mapping to an element in A. Therefore, if an entire row of **M** contains zeroes, no isomorphism can exist for that functional group-database compound pair as there is an atom with no possible valid mapping.

Since **M** represents simultaneously all possible mappings, not individual mappings of functional group atoms to database compound atoms, **M** must be searched to find specific mapping matrices **M'** for each mapping of all functional group atoms to particular database atoms (**Figure 3**). Thus, a comprehensive search of **M** enumerates all **M'** and the computational speed

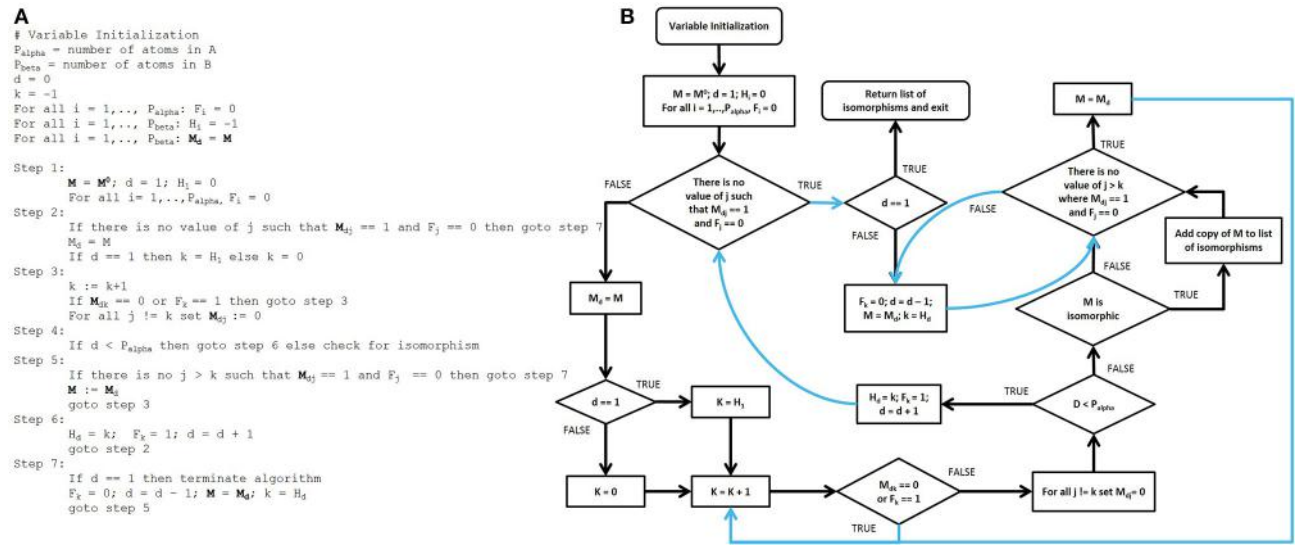


FIGURE 4 | Pseudocode and control flow diagram of the original Ullmann simple enumeration algorithm. (A) The original pseudocode for the Ullmann simple enumeration algorithm. The symbol ":=" was used in the original algorithm to denote assignment (e.g., d := d + 1 means "set d equal to d + 1"). **(B)** Control flow diagram generated from the original published 7-step Ullmann enumeration algorithm pseudocode due to the presence of numerous "goto" statements. Although "goto" statements still exist in some programming language, their use are highly discouraged in modern programming style in order to prevent errors and to improve both the readability and the maintainability of computer programs. The "goto" statements are represented by blue lines in the control flow structure. The variable SKIP2B was added and maps to no variable in the original algorithm, it is for control flow purposes only.

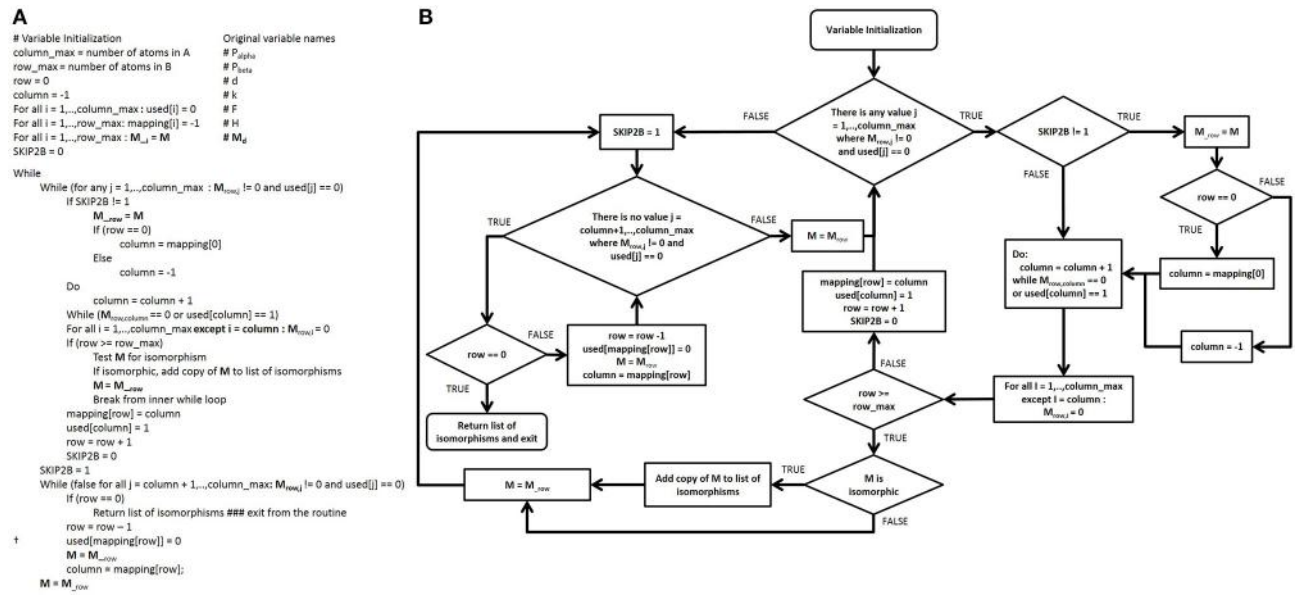


FIGURE 5 | Modernized pseudocode and control flow diagram of the Ullmann algorithm. (A) Modern Ullmann algorithm pseudocode produced from the control flow diagram with variables renamed to promote readability of the pseudocode. While converting the control flow diagram to the pseudocode, we noticed a typographical error in original published algorithmic pseudocode (line 2, step 7, in the simple enumeration algorithm on page 33 of the original publication), which we assume has been either overlooked or ignored since its publication. This line is marked † in the pseudocode above. **(B)** Control flow diagram of the modernized Ullmann algorithm.

of searching **M** is highly correlated to the number of "1" elements in **M**, which we call the "possible node mapping count" ($m = \sum \mathbf{M}_{i,j}$). Now, the Ullmann algorithm directly searches **M** in a depth-first manner; this involves copying and modifying large two-dimensional matrices frequently to enumerate all **M'**. Our algorithm avoids these costly operations by keeping track of the enumeration process with two one-dimensional integer vectors, **v** and **u**. $|\mathbf{v}|$ is equal to the number of atoms in B and **v** records

which atoms in B are mapped to atoms in A at any stage of the enumeration. The index of the element in **v** corresponds to the index of the atom of B and the value of $\mathbf{v}[i]$ the index of the atom in A to which it is mapped; so the value $\mathbf{v}[2] = 3$ denotes that atom two in B is currently mapped to atom three in A. Before any value $\mathbf{v}[i] = j$ is assigned, we check that $\mathbf{M}_{i,j} = 1$, so that the mapping stored in **v** is potentially valid. To denote an unmatched atom in B, the corresponding element of **v** is set equal to -1. Since **v** stores the same information as **M'** in the Ullmann algorithm, we can skip explicitly calculating **M'** all together saving both time and memory. In circumstances where knowing the existence of a valid mapping is sufficient, once a valid mapping is detected the algorithm can return the valid mapping and terminate. When applicable, this short-circuiting has the potential to substantially improve performance when the number of possible valid mappings is very large or the likelihood of finding a valid mapping early in the enumeration process is high (see **Figure 7A**). $|\mathbf{u}|$ is equal to the number of atoms in A and the elements in **u** indicate if a corresponding atom in A has been used in previously detected valid isomorphisms and should therefore be excluded from further enumeration. The index of a value in **u** represents the atom with the same index in A and the value of $\mathbf{u}[i]$ is either zero or one, representing if the column is non-excluded or excluded, respectively. The pseudocode for our enumeration method is shown in **Figure 6**.

Each **M'** generated by the Ullmann algorithm contains only one "1" per row and represents a particular mapping of the atoms, which must be checked to confirm if it is a valid isomorphism. The Ullmann algorithm checks for isomorphism by comparing a matrix **C** to $\mathbf{A_B}$, where $\mathbf{C} = \mathbf{M'}(\mathbf{M'A_A})^{\mathrm{T}}$., An isomorphism is found if it is true that $(\forall i, \forall j)$ where $(\mathbf{A}_{\mathrm{Bi,j}} = 1)$ then $(\mathbf{C}_{\mathrm{i,j}} = 1)$. In our algorithm, we circumvent the calculation of **C** by directly comparing $\mathbf{A}_{\mathrm{A}}$ and $\mathbf{A}_{\mathrm{B}}$ using the information stored in **v**. If $(\forall\ 0 \leq i \leq |\mathbf{v}|, \forall\ 0 \leq j \leq |\mathbf{v}|)\ (\mathbf{A}_{\mathrm{B\,i,j}} = \mathbf{A}_{\mathrm{Av[i],v[j]}})$, then **v** represents a valid isomorphism and a copy of **v** denoted as **v'** is stored in a list of isomorphisms. Once an atom in the functional group has been discovered in a valid isomorphism, the corresponding element in **u** is set to one to exclude that atom from additional enumeration. Additionally, values in **u** can be given as input to the enumerator to prevent mappings to those atoms. This is useful in excluding database compound atoms from searches or to import information concerning previously detected chemical substructure.

After all functional group-database compound pairs are checked for potential isomorphisms, it must be determined if these isomorphisms overlap one another or are subgraphs of one another. These conditions can be determined quickly by comparing the saved **v'** from each identified isomorphism. Consider two functional group isomorphisms E and F and their corresponding mapping vectors $\mathbf{v}_{\mathrm{E}}$ and $\mathbf{v}_{\mathrm{F}}$. First corresponding sets are constructed from each vector and the values with indices corresponding to context-only atoms are removed: $\mathrm{V_E} = \{\mathbf{v}_{\mathrm{E}}[i] | i \text{ is not the index of a context only atom in E}\}$ and $\mathrm{V_F} = \{\mathbf{v}_{\mathrm{F}}[i] | i \text{ is not the index of a context only atom in F}\}$. With these sets constructed, the vertices shared by E and F is simply the set $\mathrm{O} = \mathrm{V_E} \cap \mathrm{V_F}$ and the relationship between E and F can be

A

```
# Variable Initialization
column_max = number of atoms in A
row_max = number of atoms in B
row = 0
For all i = 0,..,row_max : mapping[i] = -1
For all i = 0,..,column_max : used[i] = 0
For all i in excluded_list : used[i] = 1
M⁰ is the mapping matrix between A and B.

While (row >= 0)
    Do
        mapping[row] = mapping[row] + 1
    While (mapping[row] <= column_max and (M⁰ row, mapping[row] == 0 or used[mapping[row]] == 1))
    If (mapping[row] > column_max)
        Next
    If (row == row_max)
        Test mapping for isomorphism
        Add a copy of mapping to list of isomorphisms if mapping is isomorphic
†       Return copy of mapping and exit algorithm if short-circuiting enabled.
Δ   Else If (row == 0 or mapping[0,..,row] contains an invalid mapping)
        used[mapping[row]] = 1
        row = row + 1
Continue
    While (row >= 0 and mapping[row] >= column_max)
        mapping[row] = -1;
        row = row – 1
        If (row >= 0 and mapping[row] not in excluded_list)
            used[mapping[row]] = 0
```

B

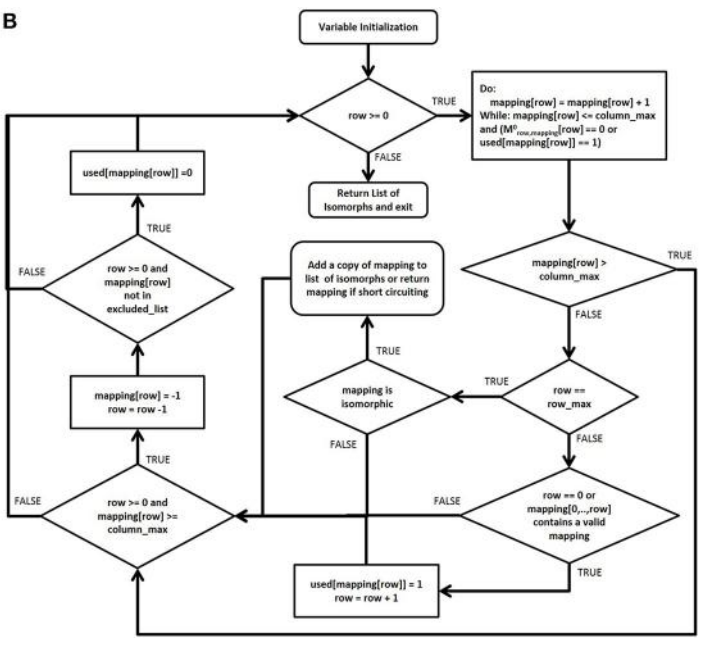


FIGURE 6 | Pseudocode and control flow diagram of the CASS algorithm. (A) Pseudocode for CASS. In this algorithm, matrices are neither copied nor modified, saving considerable computational time and memory. Also, the control flow is cleaner than in the modernized Ullmann algorithm pseudocode. Additionally, our short-circuiting method in the line marked † allows the algorithm to terminate early when a valid mapping has been identified. This allows for time savings when knowing that a single valid mapping exists is sufficient and the number of valid mappings is not needed (e.g., stereoisomerism detection). Also, partial invalid mappings allow additional short-circuiting to take place in the line marked Δ, since the algorithm is finding subgraphs in graph A that are isomorphic to graph B. **(B)** A control flow diagram of CASS.

determined by comparing O to V_E and V_F. If O = Ø, the two sets are disjoint and therefore E and F do not overlap. If |O|= $|V_E| = |V_F|$, the indices shared are identical and E and F represent mirror images of the same substructure. If $|O| = |V_E|$ and O != Ø then E completely overlaps with F and E is a subgraph of F. If $|O| = |V_F|$ and O != Ø then F completely overlaps with and is a subgraph of E. Else, $|O| < |V_F|$ and $|O| < |V_E|$ and O != Ø E and F overlap but neither is a subgraph of the other. This allows the program to differentiate functional groups that exist as a subgraph of other functional groups from those that do not and allow for proper counting of functional groups that are mirror images. Functional groups that are determined to be a subgraph of another functional group (conditions 2 and 3) have "subgraph" appended to their name. Functional groups that are overlapping but neither is a subgraph of the other (condition 4), both functional groups have "overlapping" appended to their name. For example, the hydroxyl group of a carboxylic acid would be designated a "subgraph-hydroxyl" while the carboxylic acid would be designated simply as "carboxylic acid." Additionally mirror image functional groups such as anhydrides, match twice, and this must be accounted for in order to arrive at the proper number of instances of such substructures. This comparison is conducted for all functional group pairings and once complete, the name and number of functional groups is appended to the molecular formula to generate an "extended formula." For example, if only ketones were searched for, the extended formula for acetone would be $C_3H_6O_1Ketone_1$. Functional groups can also be marked as "super" functional groups. These groups are excluded from the subgraph and inclusive designations and are used for functional groups that match a large number of other functional groups in the database or are a subgraph of many other functional groups. Alkyl halide is such a "super" functional group as it matches alkyl chloride, fluoride, iodide, and bromide; if not marked super, all instances of alkyl chloride for instances would be overlapping with alkyl halide.

In addition to searching for functional groups, CASS can also be configured to search for potential stereoisomerism between database compounds. First, all database compound pairings between database entries with the same molecular formula or extended molecular formula are identified. Searching by extended formula can greatly decrease the number of non-stereoisomeric pairing that must be tested as stereoisomers will contain the same functional groups in addition to having the same formula while other types of isomers may not. When searching for stereoisomers the same process as used for functional groups is utilized except that compounds A and B are the database compounds being tested. As our adjacency matrices do not store stereochemical information and oftentimes database molfile files only have 2 dimensional coordinates for the atoms, we do not utilize 3 dimensional coordinates in making this analysis, only the knowledge that two compounds have the same connectivity between their atoms. Implicit and explicit hydrogens can be omitted during this search to improve performance, since confirming two structures as stereoisomorphic is very time consuming, especially for large molecular graphs, where a large number of "bad mappings" must to be tested. Therefore, we had to expand our "node coloring" scheme. Thus, we included the "color" of bonded atoms to create a complex "patterned color" for an atom. This scheme can be recursively applied to include larger shells of bonded atoms. We refer to our initial coloring scheme as "element coloring" and then each shell of atoms included as "1-bond coloring," "2-bond coloring," etc. This improved node coloring scheme greatly reduces the size of *m* (**Figure 7A**), making detection of stereoisomers of large compounds tractable. Still, duplicate entries or duplicate structures cannot be distinguished using this method; although, it is reasonable to assume that the percentage of duplicate entries within any one database is very small and that stereoisomers identified by this method represent true stereoisomers. While this advanced node coloring scheme is straightforward to apply for stereoisomer analysis, it is harder to apply to functional group searching, due to boundary conditions for nodes with edges outside the functional group.

STORING FUNCTIONAL GROUP INFORMATION

Although CASS finds functional groups in relatively short time, it is undesirable to repeat the calculations every time the data needs to be accessed. To prevent repeating costly calculations, the functional group data for the database entries is stored as a SQLite database. All of the database entries are stored in one table with their molecular formulae, extended formulae, molfile files, text representations of the atom and bond objects and the number of each functional group present, including separate entries for overlapping and subgraph functional groups. Additionally the functional group molfile files and a list of the functional group names used when the database was created is stored as a separate table in the SQLite database. This is to allow for maximum portability and flexibility as everything needed to add a new compound entry is available in the SQLite database. With the appropriate program, a pre-existing functional group resolved database, and the molfile file for a new compound entry, the additional entry can be added with the functional groups stored in the SQLite database without reconstructing the entire database. However, if the list of functional groups is changed, the database must be reconstructed as the number of overlapping and subgraph functional groups may change. This SQLite format allows rapid and efficient searching for database compounds with certain properties including molecular formula and/or functional group composition, matching our standard use-case involving such information derived from CS-tagging and acquired by FT-MS.

CS-TAGGING STRATEGY ANALYSIS

After using CASS to determine the number of each functional group within all database entries, the functional group identified databases were analyzed to determine which combinations of functional groups under what conditions allows for the best disambiguation of isomeric database compounds. A specific CS-tagging strategy is represented by a set of functional group adducts and its' "performance" is measured by the number of non-isomeric extended formulae obtained from the database using the percent of non-isomeric compounds from the combined database as the base line.

As the number of functional groups in our functional group database is too large to test all permutations of all possible

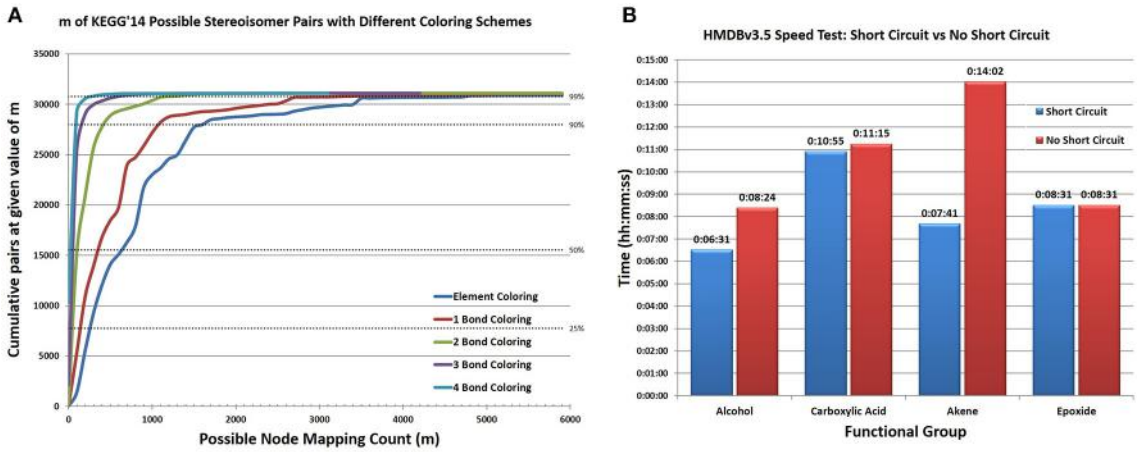


FIGURE 7 | Representative results of significant algorithmic improvements in CASS. (A) With element coloring only, the average number of mappings for each possible stereoisomer pair ranges from very small to very large, with 50% of pairs having an m greater than 700 and a maximum m of 25,219 (not shown in figure). Coloring using the element types connected within N bonds reduces m for almost all pairs. With 1 Bond coloring, 50% have an m greater than 400. A substantial improvement occurs between 1 bond and 2 bond coloring, with 50% having an m greater than 100 and a maximum m less than 6300. Incremental improvements in m occur for most compounds when $N > 2$. Using coloring schemes based on adjacent atoms greatly improves the speed of finding possible stereoisomerism with a negligible time investment to color each atom. This allows CASS to efficiently check for stereoisomerism between very large structures. **(B)** By searching for the four functional groups in the entirety of the HMDBv3.5 the time savings observed by enabling short circuiting mirrors the improvement seen in our trial set. Alcohols and alkenes show a large improvement while epoxides and carboxylic acids do not. This likely relates to the relative frequency of the functional groups. Very common functional groups show the greatest improvement as it is very likely to find an instance of the group early in the enumeration.

functional groups, strategies were generated iteratively assuming that functional group inclusion will have an additive effect on strategy performance. Therefore, strategies performing above a certain cutoff are expanded to include an additional functional group while poorly performing strategies are eliminated. In the first iteration, all one-functional group strategies are generated and the top 50 best performing strategies kept. For all iterations $i > 1$, the strategies from $i - 1$ are expanded to generate all pairings of each parent strategy with each functional group detected in the database to generate new child strategies. The performance of each child strategy is compared to the performance of the parent strategy; if the performance difference does not exceed a user-specified limit, the child strategy is removed. The top Y best performing non-redundant child strategies are then kept and passed into the next iteration. This process continues until the specified number of iterations is met or until an iteration generates no new child strategies above the performance cutoff.

As two functional groups (A and B) can perform synergistically, where in strategy [A,B] provides a greater disambiguation of isomeric compounds than the performance of [A] plus the performance of [B] would predict. Therefore, for effective strategy searching X and Y must be sufficiently large to allow poor performing strategies a chance to be paired with a synergistic functional group. Additionally, functional group adducts may not form stoichiometrically in all circumstances and the ideal strategy should take this into account. Therefore, strategy analysis can be performed in one of three modes: stoichiometrically where adduct formation can determine the precise number of functional groups, non-stoichiometrically where adduct formation can only determine whether a group is present and pseudostoichiometrically where adduct formation can determine if there is one or two instances of a functional group precisely but it cannot distinguish among 3 or more instances.

Furthermore, the number of instances of each functional group can be determined in a number of manners as we detect overlapping and subgraphs of each functional group. The strategy analysis was ran considering distinct functional groups only, distinct + overlapping, distinct + subgraph, distinct + subgraph + overlapping, distinct + subgraph + overlapping + super, and super functional groups only. Distinct only represents the functional groups likely to be detected by the most specific of adduct forming compounds, while other permutations allow us to consider the detection of functional groups in more permissive contexts. The increase in percent distinguishable compounds using the strategies generated by our analysis can guide researchers in both using commercially available adducts and guide development of new adducts.

COMPUTATIONAL PLATFORMS AND LIBRARIES

All timed analyses were done on three identical machines with dual Xeon X5650 processors @ 2.67 GHz and 24 GB of 1333 MHz ECC memory running Fedora 18 "Spherical Cow." All three algorithms were implemented in Perl 5.16.3 and SQLite v3.7.13 with DBI 1.631 was used in all programs interacting with a SQLite database.

RESULTS

ALGORITHM PERFORMANCE

CASS outperforms the older Ullmann algorithm significantly when searching for functional groups within molfile files. The

older Ullmann algorithm takes a prohibitively long amount of time for all but the most trivial analyses, while our algorithm readily performs in applications utilizing large numbers of molfile files. The number of atoms for a set of representative database compounds and functional groups was determined as was the possible node mapping count (m) for each functional-group/database-compound pair (**Tables 1**, **2**). The relationship between m and algorithm performance becomes apparent in **Figures 8–10**. **Figures 8**, **9**, based on Tables S1, S2 in Supplementary Material, visualize the obvious differences in performance between the Ullmann algorithm and CASS with no short-circuiting, in identifying four common functional groups in ten molfile files. The non-linear behavior of the Ullmann algorithm as shown in **Figure 8** is clearly unsuitable for our functional group searching. The pseudo-linear behavior of our new algorithm as shown in **Figure 9** is stable for values of m up to 150 and remains sufficiently fast for large values of m during functional group searching, making CASS tractable for systematic functional group searches in KEGG and HMDB. Furthermore, the demonstrated polynomial behavior of our algorithm (**Figure 9E**) is the best expected performance, given the debate on whether the common subgraph isomorphism problem has polynomial or NP-complete behavior (de Melo et al., 2013). Also, **Figure 10** further highlights the relative differences between the Ullmann algorithm and CASS on a log scale. This difference in performance increases substantially with respect to *m*.

However, the improvement in our new algorithm with short-circuiting is sporadic (**Figure 7B** and Table S3) and is dependent on the order of the search of **M** and the number of valid isomorphic mappings in **M** (i.e., number of isomorphic **M'**). But an excellent case for utilizing the short-circuiting variant of our algorithm is when searching for stereoisomeric compounds within databases. Two large stereoisomeric compounds, A and B, will have a very large number of possible mappings as they contain an identical number and type of atoms. A single valid mapping of all atoms in A to all atoms in B is sufficient to determine that A and B are stereoisomeric. Additional valid mappings beyond the first convey no additional information regarding the relationship of compounds A and B and do not need to be determined. For a number of possible stereoisomers from KEGG Ligand, both the short-circuiting and non-short circuiting algorithms were compared, providing sporadic results where the short-circuiting either performed better or comparably to the non-short circuiting algorithm (Table S4 in Supplementary Material).

SYSTEMATIC ISOMER ANALYSIS

The increased performance of CASS compared to the Ullmann algorithm allows for the rapid detection of functional groups and stereoisomers, which we used to create functional group-resolved SQLite versions of metabolite databases. From our functional group-resolved SQLite versions of the HMDBv3.5 and KEGG Ligand (as of March 2014), several additional analyses were performed. First, the number of distinct molecular formulae in both databases was determined as well as the number of molecular formulae the two databases have in common (**Figure 11A**). The 3557 molecular formulae were then compared against both databases to determine if the molecular formula was isomeric in neither, both, or either database. 39% were isomers in neither database, 32% were isomers in both, while 17 and 12% were isomers only in the HMDB and KEGG, respectively (**Figure 11B**).

In addition to determining the shared isomers between the databases, a historic trend analysis of isomerism was performed

Table 1 | List of representative database compounds and functional groups.

Atoms and bonds in database compounds and functional groups			
	Compound name	**Atoms**	**Bonds**
Database compounds	Deoxycytidine	29	30
	R-3-Hydroxybutyric acid	15	14
	2-Hydroxybutyric acid	15	14
	Deoxyuridine	28	29
	1-Methylhistidine	23	23
	Cortexolone	55	58
	2-Methoxyestrone	46	49
	Deoxycorticosterone	43	69
	1,3-Diaminopropane	15	14
	2-Ketobutyric acid	13	12
Functional groups	Carboxylic acid	5	4
	Epoxide	3	3
	Alkene	6	5
	Alcohol	3	2

The number of atoms and bonds in the database compound and in the functional group being searched for have an indirect impact on the performance of each algorithm.

Table 2 | Possible node mapping counts for paired functional group/database compound searches.

Possible node mapping count (*m*)				
Database compounds	**Functional group**			
	CA	**Epoxide**	**Alkene**	**Alcohol**
Deoxycytidine	46	22	18	26
R-3-Hydroxybutyric acid	25	11	8	15
2-Hydroxybutyric acid	25	11	8	15
Deoxyuridine	47	23	18	26
1-Methylhistidine	34	16	14	20
Cortexolone	84	46	42	155
2-Methoxyestrone	71	41	38	46
Deoxycorticosterone	70	45	42	43
1,3-Diaminopropane	18	6	6	13
2-Ketobutyric acid	23	11	8	13

The possible node mapping counts (m) between functional group atoms and database compounds have a direct impact on algorithm performance. In most cases, larger functional groups have larger m values than smaller functional groups and therefore should take longer to search for. These values of m are based on an element node coloring scheme.

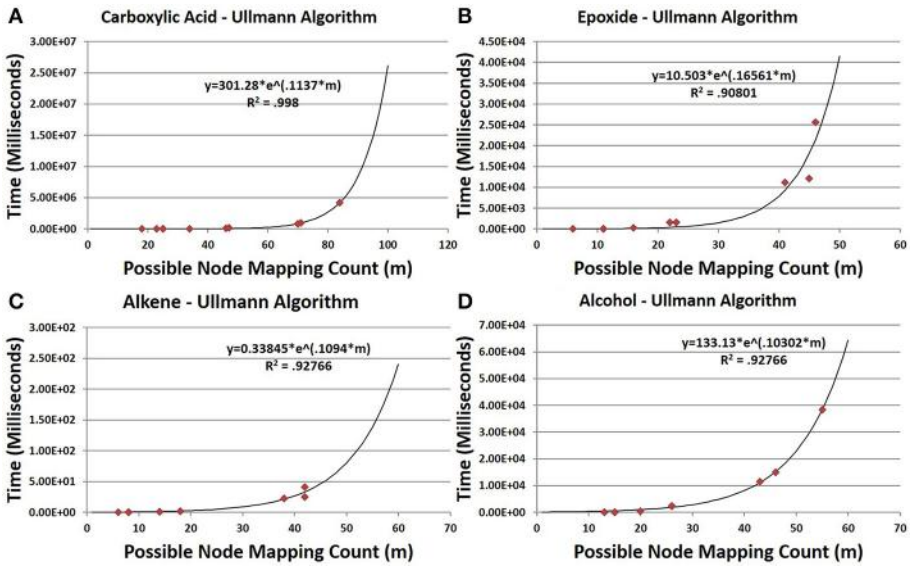


FIGURE 8 | Time trials for the Ullmann algorithm. In all cases, the time needed for the Ullmann algorithm to find all instances increases essentially exponentially with increasing m. Data were analyzed by non-linear regression to $t = a\exp(bt)$. **(A,B,D)** The time to search for carboxylic acids, epoxides and alcohols shows exponential growth with respect to m. **(C)** The time needed to search for alkenes, while also exponential, is much smaller than that needed for all other functional groups, as the alkene group is the smallest of the four groups in both number of bonds and atoms. This indicates that the time needed to find a specific group varies with respect to its size in a strong nonlinear exponential manner (alcohol with only one more bond and atom takes much longer for $m > 25$).

(**Figures 12A–C**). The percentage of isomeric molecular formulae in the HMDV3.5 and a combined database appear to have plateaued at 43 and 46%, respectively (**Figures 12A,C**). KEGG has reached a 28% isomeric content based on molecular formulae. This lower percentage of isomers in KEGG is likely due to the inclusion of pharmaceuticals and synthetic compounds that have unique molecular formulae that are not found in nature, which is probably why the isomeric content has not plateaued. What is also interesting is that the percent isomeric entries in all three databases (**Figure 12B**) is appreciably higher than the percent isomeric molecular formulae, indicating that a moderate number of isomeric molecular formulae are represented by more than 2 isomeric entries.

SYSTEMATIC STEREOISOMER ANALYSIS

Additionally, the higher performance of CASS allows for the comparison of two database structures in order to determine if they are stereoisomers of one another. For each database, all compound pairs in which the two compounds have identical formulae and the same number of bonds are tested for potential stereoisomerism. Since a single isomorphic instance of one database entry in another is sufficient to identify stereoisomeric compounds, our short-circuiting can be used to greatly accelerate these comparisons. As database structures can be very large, the potential number of mappings must be kept small for efficient analysis; this is achieved using 2-bond and 3-bond node coloring.

In addition to searching for stereoisomerism within each database, stereoisomerism was checked for compounds with the same formula and number of bonds between the two databases. Entries with duplicate names were excluded from this analysis to reduce the likelihood of comparing identical entries. The percentage of stereoisomeric compounds in the HMDBv3.5 and KEGG is 1.14 and 9.43%, respectively. The combined database has a percent stereoisomerism of 8.3%. Additionally, a historical trend of stereoisomers in HMDBv3.5, KEGG, and the combined database show early instability, followed by a downward trend that is plateauing. The large difference in stereoisomerism between KEGG and HMDB likely reflects the different portions of metabolism best represented by either database. The HMDB contains a large number of lipids and large aliphatic structures that typically have numerous structural isomers but few stereoisomers while KEGG has numerous sugars and other structures with a high number of potential stereoisomers.

CS-TAGGING STRATEGY ANALYSIS

All instances of each functional group were identified in a combined KEGG and HMDB database with duplicate entries removed. Using the functional group-resolved SQLite version of the combined KEGG and HMDB database with duplicate entries removed, we systematically tested different experimental CS-tagging strategies to determine, optimal strategies with 3, 5, 10, or 15 functional group adducts (Tables S5A–C in Supplementary Materials).

In all tests, 15 iterations were performed, with the top 50 strategies kept in the first iteration and the top 15 strategies kept in subsequent iterations, and, with a performance cutoff of 0.1%. The analysis was repeated under stoichiometric, non-stoichiometric and pseudostoichiometric quantification expectations and different degrees of allowed overlap between functional groups. Selected strategy types were compared against

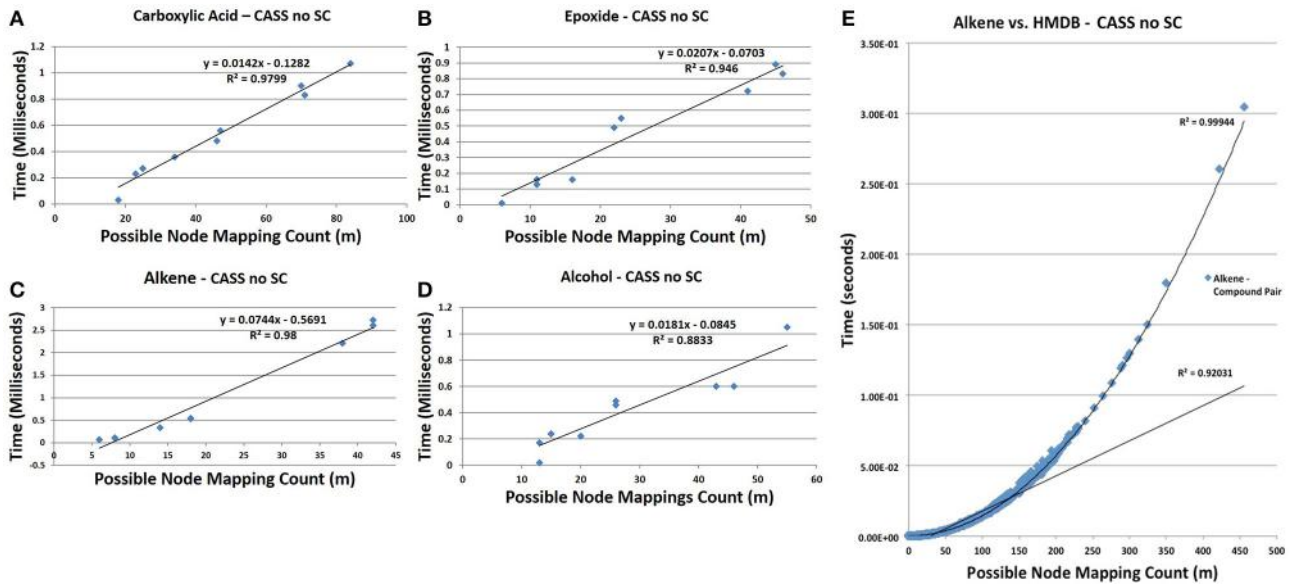


FIGURE 9 | Time trials for CASS with no short circuiting. (A) The time needed to find carboxylic acids is pseudo-linear at $m < 100$ ($R^2 = 0.9799$). **(B)** Similar to carboxylic acids, the time needed to find epoxides is pseudo-linear for observed m, ($R^2 = 0.946$). **(C)** Similarly, the time needed to search for alkenes grows pseudo-linearly with m ($R^2 = 0.98$). **(D)** The time needed to find alcohols remains low at all observed values of m but is less strongly linear than with other functional groups ($R^2 = 0.8833$). It is likely that with respect to high values of m, all functional groups would show polynomial growth; however, for most values compounds, m will be sufficiently small to allow our algorithm to show pseudo-linear performance. **(E)** The time needed to find all alkenes in the HMDB demonstrates the non-linear performance of CASS for values of $m > 150$, as the overall trend matches a second-order polynomial with an R^2 of 0.9994. Although non-linear, the time needed grows slowly enough to allow all functional group searches to complete in a relatively short amount of time. To estimate an upper bound on the values of m likely to be observed during functional group searching within metabolic databases, the value of m for each pair of functional group with database compound within the HMDB was determined. The values of m for each functional group were recorded and the largest value of m for each functional group was selected to create a set of the largest observed values of m. This set of largest values of m represents the most strenuous calculations that must be performed by our algorithm. The average largest value of m is 1007 with a standard deviation of 508. The largest value of m observed for all pairs was 4104. Although, these largest values of m are within the non-linear performance region of CASS, these values of m are still small enough to allow for efficient functional group searching within any given database structure. The time needed to find the other three functional groups in all HMDB entries was also determined and shown in Figure S2.

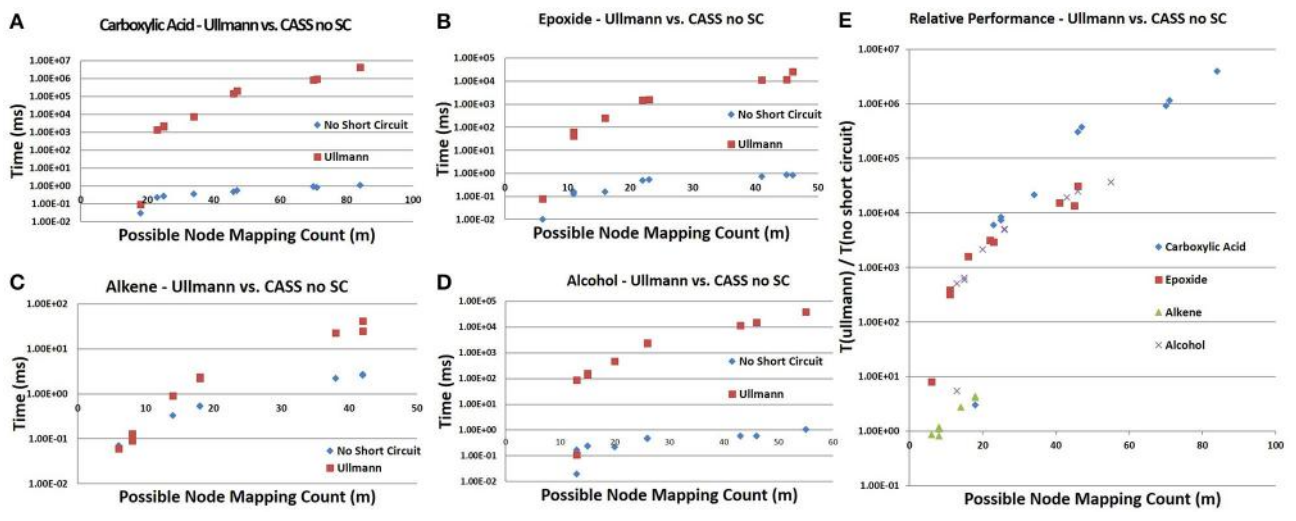


FIGURE 10 | Direct comparison of the Ullman algorithm to CASS with no short circuiting. (A–D) For all four functional groups, our algorithm shows linear performance at these values of m while the Ullmann algorithm does not. **(E)** The ratio of the time needed by Ullmann vs. our algorithm with respect to m demonstrates that our algorithm is faster than the Ullmann algorithm in all cases. This ratio increases with m and varies between functional groups.

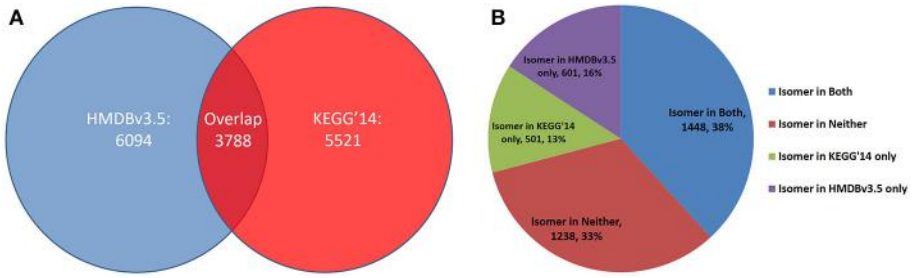


FIGURE 11 | Distinct molecular Formulae distributions across HMDBv3.5 and KEGG Ligand. (A) In each database, the number of distinct formulae was determined. Each formula was tested to see if it existed in the other database. Of a total 15,403 formulae, only 3788 (24.6%) exist in both. **(B)** Each shared formula was compared against the HMDBv3.5 and KEGG'14 to test for isomerism in either database. The majority are isomers in either both or neither database, but some are isomers in one database exclusively.

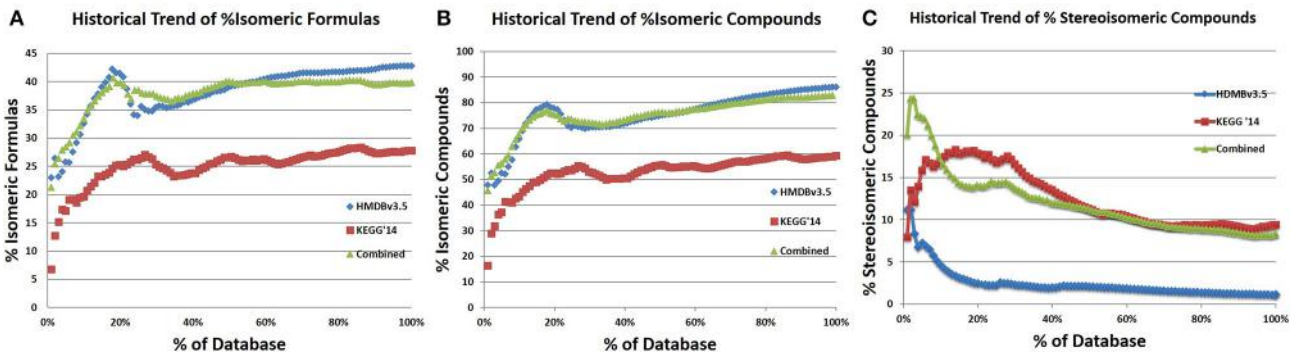


FIGURE 12 | Historic trends of % isomeric molecular formulae and compound entries. (A) The historical trend in % isomeric formulae was determined by analyzing the isomeric content ofa growing percentage of the database, following the sorted order of IDs. This was done by iteratively adding 1% of each database to a temporary database, calculating how many distinct formulae there are in the database and the number of formulae for which there is only one compound, which is the number of isomeric formulae. For both HMDBv3.5 and KEGG'14, the % isomeric formulae increase with the number of compounds, but both seem to plateau. **(C)** The combined analysis was performed identically to the individual analysis, but 1% of both databases sorted by ID was added after each trial. The combined analysis reveals that the % isomeric formulae seems to plateau at around 40%, remaining at that level from 55% onwards. **(B)** In contrast, the % of isomeric compounds in the databases seems to grow at a slow but constant rate for the HMDB, (fits linear with R^2 of 0.9883 after 30%) and a fluctuating but slowly increasing rate for KEGG.

no adduct formation to visualize the improvements in compound disambiguation (**Figure 13**). Unfortunately the distribution of isomers within the HMDB makes it a poor representation of the effectiveness of CS-tagging strategies. Over 53% of the isomeric compounds in the HMDB are isomers of 9 or more other compounds (Figure S1). This level of high isomerism within the HMDB is due to the inclusion of a very large number of lipids and triglycerides, many of which are structural isomers of one another (different positions of double bonds in the acyl chains and positions of acyl chains on the backbone) and cannot be easily disambiguated by CS-tagging and MS alone. Additional information from other methods such as LC and tandem MS will be needed to resolve lipid structural isomerism, especially for triglycerides. KEGG on the other hand has a much more manageable isomer distribution. Strategy analysis was performed using both the combined HMDB and KEGG database as well as KEGG separately.

Stoichiometric adduct formation consistently generates the best increases in percent unambiguous compounds for both databases and the ideal strategy of 3 functional groups varies very little with varying the amount of overlap or with which database was analyzed. The optimal three adducts with distinct functional groups only increases the percent of unambiguous compounds in the combined database and the KEGG database from 17.13 to 30.35% and from 40.98 to 61.63%, respectively. Strategies with 15 functional groups perform slightly better with performances of 36.67% for the combined database and 69.13% for KEGG alone. Allowing for detection of overlapping, subgraph or super functional groups offers only minimal improvement; less than 1% for 3 functional groups and less than 2.5% for strategies of 15 functional groups.

In contrast, non-stoichiometric strategies provide the worst increases in percent unambiguous compounds. For the combined database, the ideal 3 functional group strategy only allows for 23.18% of compounds to be uniquely identified. The performance is better in KEGG alone, with the ideal strategy of 3 allowing 49% of compounds to be uniquely identified. As with stoichiometric analysis, ideal strategies are similar between the

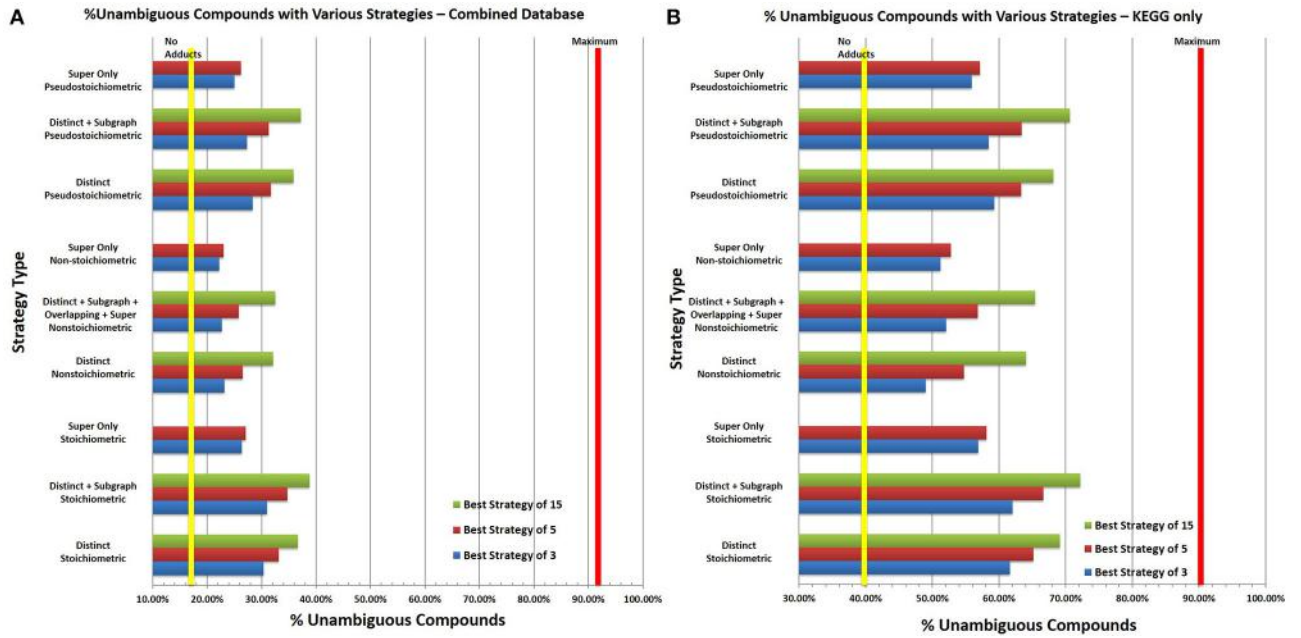


FIGURE 13 | Percent of Unambigiuous Compounds with Selected CS-Tagging Strategies. (A) In general, using more adducts improves the disambiguation of compounds for all strategies. For pseudostoichiometric and stoichiometric strategies, the difference seen in utilizing 15 functional groups compared to 5 is roughly equal to the difference in performance seen with 5 and 3 functional groups indicating that multiplexing multiple groups offers increased performance but with diminishing returns. The yellow vertical line represents the percent of unambiguous formulae without adduct formation, the red line represents the percent of formulae without stereoisomers, which cannot be disambiguated with CS-Tagging alone due to stereoisomerism. In all three stoichiometric environments, very good performance is achieved by detecting distinct instances of the functional groups alone. The ability to detect subgraph or overlapping instances of each functional group offers only marginal improvements in performance. Super functional group only strategies consistently provided the worst performance. Stoichiometric tagging offers the maximum percent of disambiguation; however, it will be difficult to ensure consistent results experimentally. Pseudostoichiometric adduct formation provides performance close to stoichiometric tagging and much greater performance than non-stoichiometric tagging. However, the performance of all the strategies falls short of the theoretical maximum expected from stereoisomerism alone. This is likely the result of the distribution of isomers in the HMDB. Many structural isomers in the HMDB differ only slightly and map to a relatively small number of formulae, making them difficult to disambiguate. The functional groups comprising each strategy differ slightly; please refer to Tables S5A–C to find the set of functional groups comprising each strategy shown as well as the performance of additional strategies. **(B)** Similarly the same analysis was performed using the KEGG database only. As with 13A, the yellow line represents the percent of unambiguous compounds without adduct formation and the red line the theoretical maximum based on stereoisomerism within the database. As with the combined database, stoichiometric strategies outperform non-stoichiometric and pseudostoichiomeric strategies but pseudostoichiometric strategies have only slightly lower performance. The highest performing strategies allow for coverage of over 70% of all database compounds but stills fall short of the theoretical 90.6% maximum (see Tables S6A–C for the functional groups comprising each strategy). The better performance of the various strategies in KEGG compared to the combined database is the result of the different distribution of isomeric compounds in KEGG. Unlike the HMDB, the isomeric compounds in KEGG are more evenly distributed among KEGG molecular formulae, allowing easier disambiguation of isomeric compounds (see Figure S1 in Supplementary Material).

combined and KEGG database, however, in non-stoichiometric analysis, allowing for detection of overlapping, subgraph or super groups does allow for noticeable improvements for smaller strategies. Detection of overlapping, subgraph or super groups has an unpredictable effect on the performance of each strategy depending on what database is considered and the number of functional groups. In the combined database, detection of overlapping, subgraph or super groups decreases performance of three functional group strategies by a marginal amount, while for KEGG, marginal improvements are observed. However, their detection improves performance of all strategies with 10 or more functional groups in both databases marginally.

In reality due to the complexity and differing reactivity of metabolites, stoichiometric adduct formation is unlikely to occur for all compounds. However, pure non-stoichiometric adduct formation is unlikely to occur as well; adduct formation will likely occur in a pseudostoichiometric manner, wherein only one to three instances of a functional group can be reliably identified in a stoichiometric manner. Pseudostoichiometric strategies perform significantly better in both databases than non-stoichiometric strategies but only marginally worse than stoichiometric ones. For the combined and KEGG databases, the best pseudostoichiometric strategy of 3 allows for unique identification of 28.37 and 59.32% of compounds. The performance of these strategies increases steadily up to 15 functional groups for both databases up to 35.83 and 68.13% for the combined and KEGG databases, respectively. Detection of overlapping, subgraph, and super functional groups has a mixed effect for strategies with less than three functional groups, but is marginally helpful for all strategies with greater than 5 functional groups.

Additionally, strategies were generated using only the super functional groups under stoichiometric, pseudostoichiometric, and non-stoichiometric conditions. In all cases, the super only strategies delivered the worst performance by a significant margin and the algorithm terminated early due to the performance cutoff in all cases.

Collectively the optimal strategies determined by this analysis can be generalized to help aid in CS-tagging reagent development and use. The most common functional groups in strategies with five or fewer functional groups are alkene, methyl, ketone, carboxylic acid, dialkyl ether, and enol; therefore, adducts for these functional groups will allow for the greatest disambiguation of metabolites. Although reagents already exist for forming adducts with most of these groups, no CS-tagging agent exists for methyl groups nor can one be easily developed due to the group's lack of chemical reactivity. However, supplementary techniques such as NMR could be used in lieu of a CS-tagging agent to determine the number of methyl groups pseudostoichiometrically. Additionally, the marginal performance increases achieved by allowing the detection of overlapping, subgraph and super functional groups in addition to distinct instances of each functional group, indicates that reagents that can detect instances of functional groups within other chemical moieties will not be necessary for effective CS-tagging strategies. Instead, multiple reagents capable of forming adducts pseudostoichiometrically or stoichiometrically against specific moieties should be multiplexed. The poor performance of the super only strategies demonstrate that optimally, reagents should form adducts with functional groups that are neither exceedingly rare within the database nor ubiquitous.

DISCUSSION

Our new algorithm, CASS, significantly outperforms the Ullmann algorithm in finding complete isomorphisms in chemical structures. Although the prototypical solution to the MCSI problem and by extension the common subgraph isomorphism problem that we have solved, the modernization of the Ullmann algorithm shows that it not suitable for identifying identical regions between compounds. Additionally, the modernization of the Ullmann algorithm revealed a typographical mistake in the original publication.

CASS allows for the creation of functional group-resolved databases necessary for assigning functional group resolved molecular formulae derived from FT-MS analysis of CS-tagged metabolites to specific chemical structures. Additionally, the short-circuiting and advanced node coloring abilities of CASS allows the detection of all stereoisomers in the KEGG and HMDB metabolite databases within a few hours on a single midrange workstation (less $5K). We use CASS to determine the theoretical number of compounds (∼9%) that cannot be distinguished using the combined functional group (from CS-tagging) and molecular formula (from FT-MS) information.

Furthermore, conversion of the molfile flat file databases into SQLite provides a number of advantages such as portability, ease of query with CS-tagging and molecular formula data as well as improvements in database access speed. Also, our variant of the molfile file format expands on the traditional file format, enabling the designation of more complex substructures within specific chemical contexts. This is achieved by allowing dynamic element typing for given atoms and support for contextual atoms to delineate functional groups with common features (e.g., aldehydes and ketones). Additionally, unlike many previous functional group search programs, CASS does not require hard coding in order to search for a given structure; therefore, the end user can easily add, remove, or modify functional groups to his or her choice without introducing errors into the program.

Our analysis of the HMDBv3.5 and KEGG Compound'13 shows only a low amount of overlap as only 24% of the distinct formulae from each databases exist in both. Thus, current database searches for metabolites based on molecular formulae could be biased, depending on the choice of the database. In addition, the significant presence of isomeric molecular formulae in these databases (i.e., 43% in HMDBv3.5, 28% in KEGG Compound 13', and 46% in a combined database) indicates that additional structural features such as functional groups determined by CS-tagging will need to be included in molecular formula-based database searches to facilitate unambiguous metabolite assignment of a large number of detected mass peaks. Moreover, a unique assignment of a molecular formula in one database could map to multiple compounds in another. Therefore, unique assignments should be checked in multiple databases to prevent potential misidentification of MS-detected compounds.

As an aside, the apparent plateauing at roughly 46% percent isomeric compounds in a combined database (from HMDBv3.5 and KEGG'13) may indicate a biologically relevant percent isomeric content of metabolomes in the biosphere. This would naturally be due to the significant number of stereospecific enzyme-catalyzed chemical reactions in cellular metabolism that appears to maintain an approximately 50% stereospecific chemical environment in living systems. The specific biological significance of this phenomenon is not completely apparent, but we suspect it may be due to some fundamental principle in information theory that living systems take advantage of at the stereochemical level.

Also, our analysis of CS-tagging strategies indicate that by multiplexing several functional group derivatizations in a single sample, using the unique isotope labeling distributions inherent in the design of the reagents, it is possible to determine: (i) the numbers of distinguishable metabolites having each functional group, (ii) the exact mass of the desired radical with high resolution MS, and (iii) chemical shift and molecular connectivity information with NMR. Together these can distinguish between many isomeric species with the same molecular formula but different functional groups, and therefore greatly reduce the ambiguity of structural assignment, especially for non-lipid metabolites. However, isomeric disambiguation of lipids will require additional methods that identify specific substructure.

In conclusion, by coupling molecular formula determination from ultra-high resolution FT-MS with additional chemical substructure information like functional group identification from CS-tagging or substructure determination from tandem MS-MS or NMR, our chemically aware substructure search algorithm CASS can provide robust assignment of FT-MS raw data to various metabolites and their isotopic enrichment profiles (e.g., ^{13}C

isotopologs of UDP N-acetylglucosamine or UDP-GlcNAc) in SIRM studies. The identity and fractional enrichment of labeled metabolites thus obtained are valuable parameters for modeling the contribution of various pathways to the synthesis of given labeled metabolites from tracer precursors such as done for UDP-GlcNAc synthesis from $^{13}C_6$-glucose (Moseley et al., 2011). Thus, the combined molecular formula and chemical substructure-based computational tools described here are key components of our computational pipeline to facilitate systems biochemical understanding of human metabolome and its perturbations by disease development.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation NSF 1252893 (Hunter N. B. Moseley), NIH R01ES022191-01 (Teresa W.-M. Fan, Richard M. Higashi, and Hunter N. B. Moseley), NIH P01CA163223-01A1 (Andrew N. Lane and Teresa W.-M. Fan), NIH 1U24DK097215-01A1 (Richard M. Higashi, Teresa W.-M. Fan, Andrew N. Lane, and Hunter N. B. Moseley) and NIH R25-CA134283 (David W. Hein). We thank Harrison Simrall and Andrew McCollam for advice and technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00237/abstract

REFERENCES

Armitage, E. G., and Barbas, C. (2014). Metabolomics in cancer biomarker discovery: current trends and future perspectives. *J. Pharm. Biomed. Anal.* 87, 1–11. doi: 10.1016/j.jpba.2013.08.041

Benecke, C., Grund, R., Hohberger, R., Kerber, A., Laue, R., and Wieland, T. (1995). MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Anal. Chim. Acta* 314, 141–147. doi: 10.1016/0003-2670(95)00291-7

Csizmadia, F. (2000). JChem: java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Model.* 40, 323–324. doi: 10.1021/ci9902696

Dalby, A., Nourse, J., Hounshell, W. D., Gushrust, A. K. I., Grier, D. L., Leland, B. A., et al. (1992). Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Model.* 32, 244–255. doi: 10.1021/ci00007a012

Daylight Chemical Information Systems, I. (2008). *4. SMARTS—A Language for Describing Molecular Patterns* [Online]. Available online at: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (Accessed May 13, 2014).

de Melo, V. A., Boaventura-Netto, P. O., and Bahiense, L. (2013). QAPV: a polynomial invariant for graph isomorphism testing. *Pesqui. Oper.* 33, 163–184. doi: 10.1590/S0101-74382013000200002

Eustis, S. E. (2011). *ETH Zurich Chemical Database*. Available online at: https://www.mcneillab.ethz.ch/?db=9

Fan, T., Lane, A., Higashi, R., and Yan, J. (2011). Stable isotope resolved metabolomics of lung cancer in a SCID mouse model. *Metabolomics* 7, 257–269. doi: 10.1007/s11306-010-0249-0

Fan, T. W., and Lane, A. N. (2008). Structure-based profiling of metabolites and isotopomers by NMR. *Prog. Nucl. Magn. Reson. Spectros.* 52, 69–117. doi: 10.1016/j.pnmrs.2007.03.002

Fan, T. W., Lane, A. N., Higashi, R. M., Farag, M. A., Gao, H., Bousamra, M., et al. (2009). Altered regulation of metabolic pathways in human lung cancer discerned by (13)C stable isotope-resolved metabolomics (SIRM). *Mol. Cancer* 8:41. doi: 10.1186/1476-4598-8-41

Fan, T. W.-M. (2012). Considerations of sample preparation for metabolomics investigation. *Handb. Metabol.* 17, 7–27. doi: 10.1007/978-1-61779-618-0_2

Fan, T. W.-M., Tan, J. L., Mckinney, M. M., and Lane, A. N. (2012). Stable isotope resolved metabolomics analysis of ribonucleotide and rna metabolism in human lung cancer cells. *Metabolomics* 8, 517–527. doi: 10.1007/s11306-011-0337-9

Fan, T. W.-M., Yuan, P., Lane, A. N., Higashi, R. M., Wang, Y., Hamidi, A., et al. (2010). Stable isotope-resolved metabolomic analysis of lithium effects on glial-neuronal metabolism and interactions. *Metabolomics* 6, 165–179. doi: 10.1007/s11306-010-0208-9

Feldman, H. J., Dumontier, M., Ling, S., Haider, N., and Hogue, C. W. (2005). CO: a chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Lett.* 579, 4685–4691. doi: 10.1016/j.febslet.2005.07.039

Fu, X., Li, M., Biswas, S., Nantz, M. H., and Higashi, R. M. (2011). A novel microreactor approach for analysis of ketones and aldehydes in breath *Analyst* 136, 4662–4666. doi: 10.1039/c1an15618g

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22, 245–252. doi: 10.1016/j.tibtech.2004.03.007

Gori, S. S., Lorkiewicz, P., Ehringer, D. S., Belshoff, A., Higashi, R. M., Fan, T. W.-M., et al. (2014). Profiling thiol metabolites and quantification of cellular glutathione using FT-ICR-MS spectrometry. *Anal. Bioanal. Chem.* 406, 4371–4379. doi: 10.1007/s00216-014-7810-z

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002). LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* 30, 402–404. doi: 10.1093/nar/30.1.402

Guo, K., and Li, L. (2009). Differential 12C-/13C-isotope dansylation labeling and fast liquid chromatography/mass spectrometry for absolute and relative quantification of the metabolome. *Anal. Chem.* 81, 3919–3932. doi: 10.1021/ac900166a

Haider, N. (2010a). *Creating a Web-Based, Searchable Molecular Structure Database Using Free Software* [Online]. Available online at: http://merian.pch.univie.ac.at/∼nhaider/cheminf/moldb.html

Haider, N. (2010b). Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules* 15, 5079–5092. doi: 10.3390/molecules15085079

Hattori, M., Tanaka, N., Kanehisa, M., and Goto, S. (2010). SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.* 38, W652–W656. doi: 10.1093/nar/gkq367

Hummel, J., Strehmel, N., Selbig, J., Walther, D., and Kopka, J. (2010). Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics* 6, 322–333. doi: 10.1007/s11306-010-0198-7

Kaddurah-Daouk, R., Kristal, B. S., and Weinshilboum, R. M. (2008). Metabolomics: a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.* 48, 653–683. doi: 10.1146/annurev.pharmtox.48.113006.094715

Kind, T., and Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinform.* 7:234. doi: 10.1186/1471-2105-7-234

Kotera, M., Mcdonald, A. G., Boyce, S., and Tipton, K. F. (2008). Functional group and substructure searching as a tool in metabolomics. *PLoS ONE* 3:e1537. doi: 10.1371/journal.pone.0001537

Lane, A. N., Fan, T. W., Bousamra, M. I. I., Higashi, R. M., Yan, J., and Miller, D. M. (2011). Stable Isotope-Resolved Metabolomics (SIRM) in cancer research with clinical application to nonsmall cell lung cancer. *OMICS* 15, 173–182. doi: 10.1089/omi.2010.0088

Lane, A. N., Fan, T. W., and Higashi, R. M. (2008). Isotopomer-based metabolomic analysis by NMR and mass spectrometry. *Methods Cell Biol.* 84, 541–588. doi: 10.1016/S0091-679X(07)84018-0

Lane, A. N., Fan, T. W.-M., Xie, X., Moseley, H. N., and Higashi, R. M. (2009). Stable isotope analysis of lipid biosynthesis by high resolution mass spectrometry and NMR. *Anal. Chim. Acta* 651, 201–208. doi: 10.1016/j.aca.2009.08.032

Le, A., Lane, A. N., Hamaker, M., Bose, S., Gouw, A., Barbi, J., et al. (2012). Glucose-independent glutamine metabolism via TCA cycling for proliferation and survival in B cells. *Cell Metab.* 15, 110–121. doi: 10.1016/j.cmet.2011.12.009

Lorkiewicz, P., Higashi, R. M., Lane, A. N., and Fan, T. W. (2012). High information throughput analysis of nucleotides and their isotopically enriched isotopologues by direct-infusion FTICR-MS. *Metabolomics* 8, 930–939. doi: 10.1007/s11306-011-0388-y

Mattingly, S. J., Xu, T., Nantz, M. H., Higashi, R. M., and Fan, T. W. M. (2012). A carbonyl capture approach for profiling oxidized metabolites in cell extracts. *Metabolomics* 8, 989–996. doi: 10.1007/s11306-011-0395-z

Moseley, H., Lane, A., Belshoff, A., Higashi, R., and Fan, T. (2011). A novel deconvolution method for modeling UDP-GlcNAc biosynthetic pathways based on 13C mass isotopologue profiles under non steady-state conditions. *BMC Biol.* 9:37. doi: 10.1186/1741-7007-9-37

Nesvizskii, A. I., Vitek, O., and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 4, 787–797. doi: 10.1038/nmeth1088

Owens, M. (2006). *The Definitive Guide to SQLite.* New York, NY: Apress.

Pan, Z., and Raftery, D. (2007). Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Anal. Bioanal. Chem.* 387, 525–527. doi: 10.1007/s00216-006-0687-8

Ramautar, R., Berger, R., Van Der Greef, J., and Hankemeier, T. (2013). Human metabolomics: strategies to understand biology. *Curr. Opin. Chem. Biol.* 17, 841–846. doi: 10.1016/j.cbpa.2013.06.015

Raymond, J. W., and Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.* 16, 521–533. doi: 10.1023/A:1021271615909

Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *J. ACM* 23, 31–42. doi: 10.1145/321921.321925

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36. doi: 10.1021/ci00057a005

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0-The human metabolome database in 2013. *Nucleic Acids Res.* 41, D801–D807. doi: 10.1093/nar/gks1065

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., et al. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* 37, D603–D610. doi: 10.1093/nar/gkn810

Wood, P. L. (2014). Mass spectrometry strategies for clinical metabolomics and lipidomics in psychiatry, neurology, and neuro-oncology. *Neuropsychopharmacology* 39, 24–33. doi: 10.1038/npp.2013.167

Ye, T., Mo, H., Shanaiah, N., Gowda, G. A., Zhang, S., and Raftery, D. (2009). Chemoselective N-15 Tag for sensitive and high-resolution nuclear magnetic resonance profiling of the carboxyl-containing metabolome. *Anal. Chem.* 81, 4882–4888. doi: 10.1021/ac900539y

Zhang, Y., Qiu, L., Wang, Y., Qin, X., and Li, Z. (2014). High-throughput and high-sensitivity quantitative analysis of serum unsaturated fatty acids by chip-based nanoelectrospray ionization-Fourier transform ion cyclotron resonance mass spectrometry: early stage diagnostic biomarkers of pancreatic cancer. *Analyst* 139, 1697–1706. doi: 10.1039/c3an02130k

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: 25 March 2014; accepted: 03 July 2014; published online: 28 July 2014.
*Citation: Mitchell JM, Fan TW-M, Lane AN and Moseley HNB (2014) Development and in silico evaluation of large-scale metabolite identification methods using functional group detection for metabolomics. Front. Genet. **5**:237. doi: 10.3389/fgene.2014.00237*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.
Copyright © 2014 Mitchell, Fan, Lane and Moseley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 26 August 2014
doi: 10.3389/fgene.2014.00253

Systems biology and brain activity in neuronal pathways by smart device and advanced signal processing

***Gastone Castellani*[1]*, *Nathan Intrator*[2] *and *Daniel Remondini*[1]**

[1] *Department of Physics and Astronomy, L. Galvani Center for Biocomplexity, Biophysics and Systems Biology, University of Bologna, Bologna, Italy*
[2] *Department of Computer Science, Exact Sciences Faculty, Tel Aviv University, Tel Aviv, Israel*

Edited by:
Pietro Lio, University of Cambridge, UK

Reviewed by:
Nicola Neretti, Brown University, USA
Armando Bazzani, University of Bologna, Italy

***Correspondence:**
Gastone Castellani, Department of Physics and Astronomy, L. Galvani Center for Biocomplexity, Biophysics and Systems Biology, University of Bologna, Viale Berti Pichat 6/2, Bologna, Italy
e-mail: gastone.castellani@unibo.it

Contemporary biomedicine is producing large amount of data, especially within the fields of "omic" sciences. Nevertheless, other fields, such as neuroscience, are producing similar amount of data by using non-invasive techniques such as imaging, functional magnetic resonance and electroencephalography. Nowadays a big challenge and a new research horizon for Systems Biology is to develop methods to integrate and model this data in an unifying framework capable to disentangle this amazing complexity. In this paper we show how methods from genomic data analysis can be applied to brain data. In particular the concept of pathways, networks and multiplex are discussed. These methods can lead to a clear distinction of various regimes of brain activity. Moreover, this method could be the basis for a Systems Biology analysis of brain data and for the integration of these data in a multivariate and multidimensional framework. The feasibility of this integration is strongly dependent from the feature extraction method used. In our case we used an "alphabet" derived from a multi-resolution analysis that is capable to capture the most relevant information from these complex signals.

Keywords: multivariate analysis, multiple networks, electroencephalography, genomics, metagenomics

INTRODUCTION

Brain activity is without doubt the most complex process in nature. While the body of research is exponentially growing, it is quite amazing that fundamental building blocks or atoms of this process are still quite unknown. Two of them indicate how far we are in understanding brain processes; the first is the fundamental synaptic modification rule in a single neuron, and the second is internal brain representations of the physical world (and sensory input).

For a long time, it was assumed that it would be possible to describe the synaptic modification rule by deducing from observations, and analyzing them mathematically (Lynch et al., 1990; Cooper et al., 2004) in a similar way as other physical rules have been discovered. As the process turns out to be extremely complex in terms of the different neuro-transmitters, neuro-receptors and the chemical interactions which lead to the changes, it is now assumed that further deductions and a potential breakthrough in understanding synaptic modification may be obtained by massive computer simulations (Kandel et al., 2013). This is motivated by the immense progress computers have made in the last two decades, and the believe that computational power and memory which resembles the brain will be reached in a decade (Kurzweil and Grossman, 2005).

The quest for understanding the internal brain representation is somewhat independent of the quest for understanding synaptic plasticity. To illustrate how little we know about internal representations, we can take an object such as a desk, and point out that we do not know what it is that makes the simple combination of a surface and legs be represented (or recognized) as a desk. Specifically, what is the difference in representation for two (similar desks), is it mainly temporal, namely a different form of oscillation of the same neurons, or spatial, mainly activity of different neurons (Biederman, 1987; Edelman, 1999).

This somewhat frustrating description of the current state of the art suggests that a certain change in the way we collect data about the brain may be necessary so as to drive us to more meaningful conclusions.

A step in that direction occurred when functional MRI (fMRI) became popular. Then, not only we moved away from determining brain representations, but we also started looking at brain activity in a very crude way. Looking at oxygenated blood to different regions of the brain as a marker for neural activity in those regions, and doing so while integrating data in 3 s time windows. This crude brain activity measure led to great progress in brain activity interpretation and in attributing functional labels to different brain regions. Then came an even more surprising finding; we realized that we do not need to fully understand the role of certain regions in various cognitive and emotional tasks. Instead, it is enough to know the typical (crude) pattern of activity in a group of normal people, and apparently, an attempt to alter the activity in such regions in a group of subjects that suffers from some brain malfunction, may alleviate symptoms of that malfunction.

This paper suggests that another step forward in understanding brain activity and improving brain malfunction may come from developing new methods which like fMRI, provide a view on different functional units of the brain, but, unlike fMRI can be taken outside of the clinical setup and put into continuous mobile use to operate in any environment and thus enrich our ability to observe brain activity under natural settings.

To motivate this, we note, that it is remarkable how much we have learned about brain networks of activity from fMRI given its temporal and clinical limitations (Cabeza and Nyberg, 2000).

The electroencephalography (EEG) is a much older method for sensing non-invasively the functioning brain, with human recordings starting in 1924 (Haas, 2003). The electrical activity mainly results from fluctuations in ionic current flows within (1000 or more) neurons and it provides an indication to the type and degree of activation of different brain regions (Niedermeyer and Lopes da Silva, 2005). Throughout the century of EEG research, EEG energy features were extracted from a small number of frequency bands (e.g., Klimesch, 1999) and other features were extracted from time-locked averaging (ERP and EP) of the response (for review see: Luck, 2005). As the role of EEG in characterizing epilepsy was discovered, it was determined that epilepsy is some form of excessive synchrony between neurons and between brain regions. This has led to the discovery of more advanced signal processing methods which are sensitive to early synchrony changes (Fisher et al., 2005). However, more advanced signal decomposition and feature extraction methods have emerged only very recently in the analysis of EEG data (Duncan et al., 2013; Intrator, 2014).

It is likely that in the near future, there will be several new brain activity representations, all of which will be rich in content and will provide orders of magnitude more data as they will enable continuous mobile monitoring. This paper discusses the usage of such advanced methods, and application of methods which were mainly developed for genomic data analysis, in brain activity interpretation.

There is indeed, a huge overlapping between methods used in genomic data analysis and methods used for brain-activity interpretations. Among the most used we can quote correlation methods, that has been used both for large scale gene-network analysis and for several brain data analysis and modeling (Cooper et al., 2004; Remondini et al., 2005). Other overlapping between these two fields are given by the role of noise in the spontaneous background activity in neural and genomic systems and the subsequent modeling strategies (Milanesi et al., 2009) mutuated from the field of complex systems. In the last 20 years another unifying concept has been developed within the field of statistical mechanics and complex systems: the concept of complex network (Albert and Barabási, 2002). The idea of complex network has been applied to neural systems and to genetic systems by the fundamental tool of connectivity and degree distributions such as the famous power law that is observed in both systems. As a further analogy, at least from the point of view of modeling and data analysis, there is the concept of pathway. The pathways analysis for genomic systems is now a common tool that provide a better interpretation and simplification of this complex data (Francesconi et al., 2008). Nevertheless, the neuronal pathways, or neuronal circuits and areas, have a long history in neuroscience, starting from the classical phrenological idea, about the localization of emotions and neuronal functions. The modern imaging tools and methods are now supporting and confirming the fact that neuronal functions are precisely localized in the brain and that there is a strong relation between the anatomical and the functional localization. This is exactly the same that is observed in cells and tissues by pathways analysis.

In this paper we will take in exam the relations between the genomic and neuronal data analysis and modeling and will illustrate how this can be a powerful method for the analysis of a new generation of data obtained from EEG. We strongly believe that this method will be a further advancement in the field of Systems Biology.

NOVEL BRAIN ACTIVITY INTERPRETATION

Electroencephalography sensing started at the beginning of the 20th century (see Swartz, 1998 for a full review). The first recording of EEG from humans occurred in 1923, with the seminal work of Hans Berger (Haas, 2003), who discovered the Alpha and Beta rhythms of brain-wave oscillations. Later, other typical oscillations were discovered; those below alpha and those above beta. With multi-electrode recording, it became apparent that the EEG signal is not uniform across the skull, and that the signal observed in each electrode is strongly affected by the cortical volume closest to that electrode. This enabled the analysis of correlations of signals between different regions (electrodes), or as is thought now, between different (distributed) cortical networks (Buzsáki and Draguhn, 2004).

While EEG is not considered spatially accurate, the analysis of activity correlations across electrodes gave research a strong boost, in particular, it enabled de-correlating between different sources of brain activity using blind source separation methods such as independent components analysis (ICA; Delorme and Makeig, 2004). The introduction of ICA tools to the EEG community which was mainly done by Delorme and Makeig (2004), led to a large body of work in the analysis of EEG under many brain state conditions. It also enabled an efficient artifact removal (mainly due to muscle activity) from EEG data.

From this short review, one can conclude that separation or decomposition of the EEG signal into different components is a very effective way to study different brain networks in separation. The question becomes, whether an electrode array is essential for such separation.

While the body of work on multi-channel EEG signal decomposition is huge, the amount of work on single-channel EEG decomposition is very small. It was used for example to adapt the features to different subjects for brain computer interface, but from a 32-electrode cap (Yang et al., 2007). In this paper, we concentrate on EEG signal decomposition from a single EEG lead which is given as the difference of two EEG electrodes. The signal difference between two frontal EEG electrodes can provide the simplest measure of Cerebral Asymmetry (Henriques and Davidson, 1990). This asymmetry has long been associated with emotional reaction as well as during cognitive tasks (Davidson, 1988). Thus, if one wants to select a single EEG lead that can cover bot emotional and cognitive brain states, it makes sense to use the difference between Fp1 and Fp2, which are two frontal electrodes.

Luckily, these electrodes reside on the forehead and thus, may be easier to put, and can be dry without the need of a conductive gel.

Using a 3-sensor EEG as in **Figure 1**, Intrator (2014) has discovered features that can be obtained from a single EEG lead and may be useful for emotional and cognitive brain state discovery. These were found using a two stage process: first, a signal processing

FIGURE 1 | The EEG sensor.

and decomposition is applied to propose candidate features, and then, big-data mining and robust statistics methods are used to prune the features and test the robustness and universality of the remaining features across subjects and across conditions. These brain activity features (BAF) provide potential new insights on brain activity and states. They distinguish between three major types of activity: focused, distributed, and chaotic.

Before describing the distinction, we briefly explain what can be seen in **Figures 2** and **3**. Each column of each panel represents the activity of a single BAF (in this case, 121 different features) at a certain consecutive time point of about 1 s. In all panels, the BAFs are the same and are ordered in the same order. Each panel represents about an hour of brain activity. The BAFs which were obtained from different subjects, use the heat color map is used to represent the magnitude of activity, so the more brown/red each pixel is the more active the corresponding feature in the specific time location is. From the activity during the "focused" state, it is apparent that there is a certain correlation and continuity between the features, so that the activity, which can change in time between different features, changes in a continuous way, so that features that are presented close to one another are more likely to become active. The chaotic stage of non-REM sleep is the only exception.

The relation between these features and well-known EEG features or known areas and networks of brain activity is subject to study and will be described elsewhere. Some indications from anecdotal evidence suggest that the activity in the early part of sleep resembles activity during Anesthesia and during some forms of meditation. From studies done on that meditation performed during fMRI scans, we deduce that these specific features correspond to activity in the medial pre-frontal cortex.

Figure 3 depicts the richness of the brain states as is observed by the BAF during sleep and fatigue.

The left panel represents close to 3 h of activity while the right panel represents about an hour and a half of activity. Clear distinction between three known sleep stage are see and they correspond to the early, REM and non-REM stages.

As is well known, sleep monitoring is crucial for the early detection of physical and mental health problems; diagnosis and treatment of insomnia; and diagnosis and monitoring of dementia. Fatigue monitoring is crucial when the brain is engaged in tasks that require fast thinking and response, especially in roles where alertness is essential to performance and safety (e.g., a pilot). The right panel indicates the strength of the BAF for fatigue monitoring: it depicts the brain activity of a subject briefly falling asleep while watching a movie. Temporal regions where stronger and weaker engagement with the movie are clearly visible, as well as the length and depth of sleep.

COMPLEX NETWORK THEORY

In the last decade, physics has been expanding to new research areas. In particular, life-related sciences (ecology, sociology, economics, and last but not least biology) have been showing striking analogies with complex systems arising from various physical areas. Such approaching has happened from both fronts: on the life science side, huge amounts of data have become available for detailed analysis, thanks also to the Internet, through which this data is nowadays easily collectable and queryable (e.g., stock market financial series, social networks, high-throughput biological data). On the other side, many physical and mathematical tools, that had been proven useful in explaining complex phenomena like polymer growth or spin glass, began to spread to other research areas like biological and social sciences in a broad sense.

The common trait of these research fields can be found in the framework of network theory, which focuses on the relationships among elements and allows to draw general conclusions, even though the details of the system are not completely known or easily tractable from a mathematical point of view. Relaxing the attention to the details of the specific interaction or element, network theory aims to provide tools for the characterization of a set of relationships, represented as edges or *links*, occurring among similar elements, referred to as vertices or *nodes*.

One of the most powerful approaches to physical systems is statistical mechanics. Many results (for "ideal" gases or solids) have been obtained by considering random interactions between elements of the system, so that a "mean field theory" could be built from the average behavior of the system. The main drawback of this mean field approach (and the actual challenge at the same time) is that complex systems (to which living and life-related systems belong) are often characterized by a non-trivial set of interactions, and a mean field approach can completely miss the interactions. Moreover, social and biological systems can be considered as constantly far-from-equilibrium systems, since equilibrium for every life-related process equals to death, and a continuous influx and efflux of energy and matter is necessary to maintain life-suitable conditions. It is thus quite hard to fit them into equilibrium-based models that we can say to constitute the "core" of classical statistical mechanics.

An approach that has received renewed attention is based on the so called Master Equation (CME) that describes the temporal evolution of the probability of having a given number of

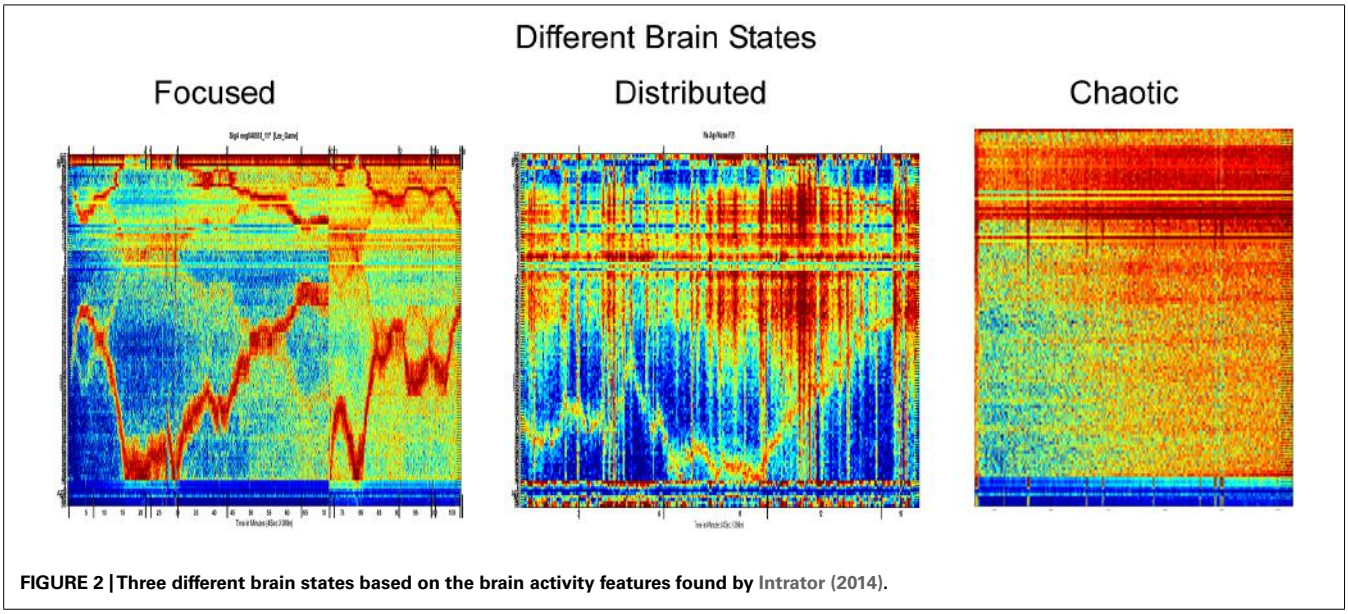


FIGURE 2 | Three different brain states based on the brain activity features found by Intrator (2014).

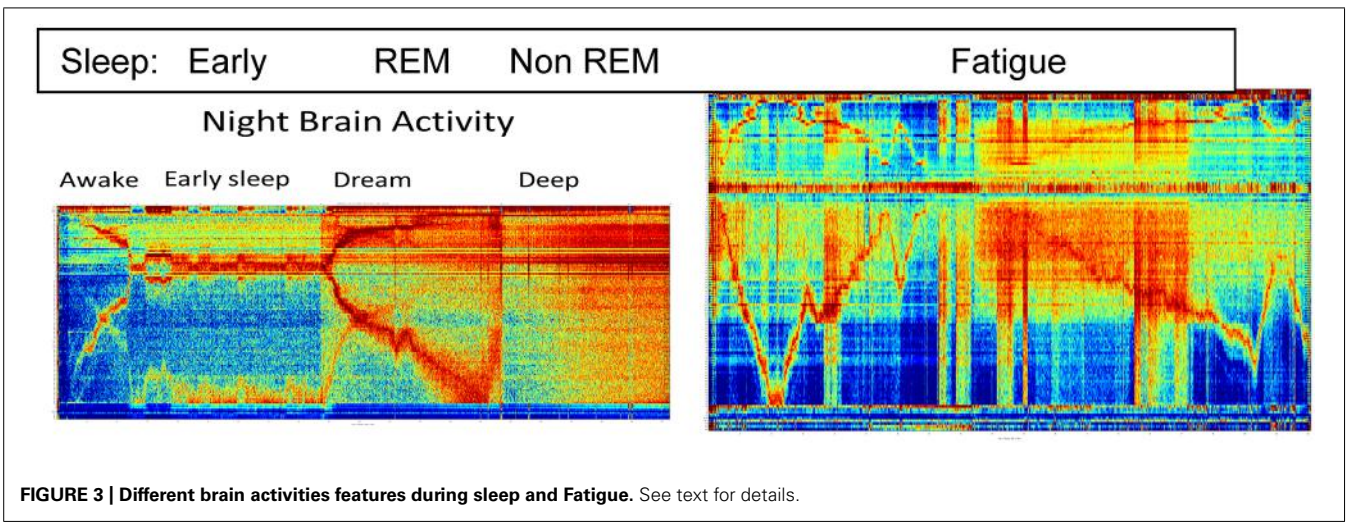


FIGURE 3 | Different brain activities features during sleep and Fatigue. See text for details.

molecules for each chemical species involved. The discrete probabilistic approach, as with CME, is attractive because it ensures the correct physical interpretation of fluctuations in the presence of a small number of reacting elements (as compared to continuum approaches as Langevin and Fokker-Planck formalism; van Kampen, 2007) and because it provides a unitary formulation for many biological processes, from chemical reactions to ion channel kinetics. The CME theory can be related to predictions on the noise levels in selected biological processes, as for example during transcription and translation (Friedman et al., 2006). In particular, the observation that mRNA is produced in bursts varying in size and time has led to the development of new models capable of better explaining the distributions of synthesized products (Cai et al., 2006).

The models based on CME can help to characterize the role of noise in networks reconstruction as well as the role of fluctuation in the enhancement and maintenance of biological functions.

Furthermore, the ME approach, allows to compute all the thermodynamic quantities, including entropy and free energy, with the consequent possibility to characterize the system as a non-equilibrium system if the detailed balance condition is not satisfied.

One of the greatest contributions, which may be given by network theory to the understanding of biological and social systems, is that the network architecture may reflect the dynamical processes that led to it. In a pure statistical-physical fashion, different "universality classes" can be sought for in order to fit the process we are studying, be it the ask-bid mechanism for a stock, the patterns of gene expression or neuronal activation following a stimulus. We remark that the features of a network model are peculiar from a static viewpoint (e.g., the relation between network topology and the evolutionary model that led to it) and from a dynamic viewpoint (e.g., the responses to perturbation, or the noise features of a stochastic dynamics). Recent models of social networks (Holme

and Newman, 2006) show that the situation can be even more complicated, with nodes interactions affecting network topology and network topology affecting node interaction dynamics. This is a common paradigm for biological systems at several levels, for genomic, nervous, and immune (for a recent review, see Gross and Blasius, 2008).

MULTIPLEX NETWORKS

During the last years, a growing interest in the so called multiplex networks has gradually grown within the scientific community. A multiplex network is a topological structure where individual nodes can have links belonging to several layers of networks at the same time. The multiplex, or multivariate network was well known in social sciences at least starting from the seventies (Boorman and Harrison, 1976).

A useful example for pointing out the differences between networks and multiplex is the analogy, from a mathematical-statistical point of view, with univariate and multivariate data.

A univariate variable is identified by single measurements; for example a population survey to estimate the average weight of elderly. Since we are only working with one variable (weight), we would be working with univariate data.

A multivariate variable is identified by multiple measurements for each sampling unit. If for example, in the same population of elderly, we are collecting not only weights, bur also blood pressure, heights, heart rate, etc, we will have 4-uples of values.

In the field of social science and social networks there are many examples of multiplex. In general, each individual node can have different kinds of social ties or relations or transportation systems where each location is connected to another location by different types of transport.

In social sciences a multiplex is defined on the basis of the existence of multiple relations among actors, where actors are defined accordingly to the actor–network theory (ANT; Latour, 1987; Law and Hassard, 1999). At a larger scale relations among nations are characterized by a plethora of cultural, economic, and political exchanges as well as from other form of connections.

Single networks have been studied extensively (Albert and Barabási, 2002; Boccaletti et al., 2006) also from a dynamical point of view (Dorogovtsev et al., 2008) and in social sciences (Wasserman and Faust, 1994). Nevertheless, in nature there exist many systems that cannot be considered as single networks. Noticeable examples are: transportation networks, climatic systems, economic markets, energy-supply networks, ecological networks, human brain and metagenomic systems (Bianconi, 2013).

Multiplexity is thought to play an important role in the organization of large-scale networks. For example, the existence of different link types between agents explains the overlap of community structures observed in ecological, genomic, metagenomic, and social networks (Szell et al., 2010).

The concept of multiplex is taking new space in modern Biology. As a paradigmatic example we will consider metagenomic data and suitable methods for multivariate associations between multiple set of omic data on the same population.

The human metagenome is the set of *Homo sapien*s genes plus the trillions of genes in the genomes of microbes that live in the human body. The microbial genome (microbiome) is in a dynamical relation with the human organism and helps it by crucial functions such as metabolic processes, shaping, control and protective immune (IS) system development, that helped the (co)-evolution of human being and ultimately also the brain development.

With the term Metagenomics, we define the set of omics measurements aimed to quantify the composition and the interactions dynamics between the host and the microbiome. This includes characterization at the level of DNA (metagenome), RNA (meta-transcriptome), protein (meta-proteome), and metabolic network (metabolome), both for the host and the microbiome. Hence, *H. sapiens* is a metaorganism (or super organism) where the different microbiota present in different organs play a major physiological and pathological role.

The interaction between GM and host is personalized, dynamic, bidirectional, history-dependent and is taking place in a multivariate way, by exchange of various molecules: metabolic, genetic, immunitary etc. The dynamic properties of the GM are caused by the fact that GM is a complex ecosystem with a complex dynamics derived by the interactions with components such as the virome (the set of viruses in the human body) the IS and the Neural System. The natural way to characterize the interaction between GM and host is to perform multiple intersection between metagenomic layers an to reconstruct networks and multiplexes.

From this perspective, social systems and biological systems can be seen as a non-linear superposition of complex networks, where nodes represent "actors," "genes" or metabolites and links capture a variety of different social and biological relationships. Human societies and biological systems can be regarded as large numbers of locally interacting agents, connected by a broad range of relationships based on exchange of molecules or social relations. These relational ties are highly diverse in nature and can represent a variety or relations (friendship, love, communication) or ecological interactions (exchange of nutrients, predator/prey relationship, cooperation, amensalism, or neutrality).

The networks in the different slices are not independent, their shapes are interconnected and reciprocally influenced; one network can act as enhancer or inhibitor on the other.

For instance networks in the brain can have excitatory and inhibitory connections, and these can influence the behavior of neurons in other slices. Another example is the transcriptional network where connections intra-slice can modify connections inter-slice (e.g., splicing and transcription factors). Also the case of metagenomic networks is best understood within the framework of multiplex: the cross-talk between host IS and microbiome is influenced by ecological interactions between the Gut Microbiota. Hence we can say that several biological systems, including the brain, can be characterized as a superposition (a linear combination, or also a non-linear combination) of its networks, all defined on the same set of nodes. This superposition is usually called multiplex, multirelational, multimodal, or multivariate network (see **Figure 4**).

NETWORK RECONSTRUCTION FROM GENE-EXPRESSION DATA BY *A PRIORI* BIOLOGICAL KNOWLEDGE

High-throughput gene expression analysis has become one of the methods of choice in the exploratory phase of cellular molecular

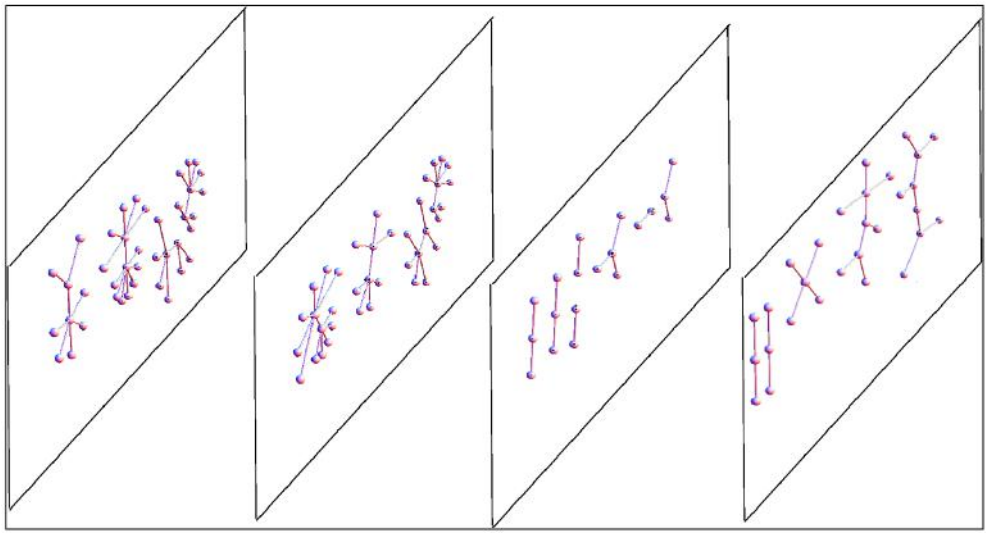


FIGURE 4 | Scheme of a multiplex network with four layers. The same nodes appear in every multiplex layer, but every layer can have different internal connections. In general, in every layer we can have different kinds of networks, both in terms of topology or because of different represented relationships. For example, we could have a multiplex in which in one layer there are genes connected by a transcription network, in the second layer the proteins (produced by the genes) can interact, bind, or be co-expressed, and in the third layer the enzymes encoded by the proteins are embedded in a metabolic network. The typical network observables (e.g., connectivity) that in a single network are scalar values for each node, in a multiplex become a vector (one value for each layer), thus the relationship between nodes based on these vectors can be more complex than in a single network.

biology and medical research studies. Although microarray technology has improved measurement accuracy, and new statistical algorithms for better signal estimation have been developed (Hekstra et al., 2003; Irizarry et al., 2003; Affymetrix Inc.), reproducibility remains an issue (Fortunel et al., 2003). A way to overcome this difficulty is to extend the analysis, in particular the interpretation of the results, from a single-gene level (in which variablity is maximal) to a higher level in which genes are grouped into functional categories. This approach has been shown to be more robust and reproducible (Subramanian et al., 2005; Manoli et al., 2006), since the "integration" of multiple gene expression patterns may "average out" fluctuations (i.e., false positives). Moreover, it mat lead to an easier biological interpretation of the experimental observations, since the single significant genes are embedded into functional categories or processes of clearer biological meaning.

Gene ontology (GO; Ashburner et al., 2000) and biological pathways are the two main gene-grouping schemes in use. GO organizes genes according to a hierarchy of terms, that from a network point of view is defined as a directed acyclic graph (DAG), in simple terms a "tree" in which genes are the "leafs" and the grouping categories are the "branches" (thus following a hierarchy from the external branches to the "root"). This DAG is divided into three categories: "cellular component," "biological process," and "molecular function." Genes appear in more than one level in each of the three categories, but no relation between genes is described (apart from them being in the same group). The biological pathway database cured by the Kyoto University (Kyoto encyclopedia of genes and genomes, KEGG; Kanehisa and Goto, 2000) is probably the most known: it groups genes into pathways of interacting genes and substrates, and contains specific links between genes and substrates that interact directly. Both databases are manually curated but incomplete, also because the knowledge of gene functions and interactions is still evolving. Each gene belonging to the GO database belongs to several categories, nested as in a phylogenetic tree: starting from a gene, we can reach the root through several branches, representing all the categories it belongs to. A limit of GO is the choice of the categories, that might not be so rigorous or univocal. KEGG provides instead a more detailed organization of the genes, since the relations are the exact biochemical interactions occurring inside the cell, but it contains information on fewer genes than GO, since fewer genes are so clearly characterized in terms of their products and interactions.

Different approaches have been proposed to identify significant gene groups based on lists of differentially expressed genes. Several methods have been implemented that can be directly applied to existing gene-grouping schemes. GOstat (Beissbarth and Speed, 2004) compares the occurrences of each GO term in a given list of genes (tested group) with its occurrence in a reference group (typically all the genes on the array) assigning a p value to each term. In the context of pathway analysis, a similar approach is used by Pathway Miner (Pandey et al., 2004) which ranks pathways by p values obtained via a one-sided Fisher exact test. Other methods allow investigators the possibility to define their own gene-grouping schemes. For example, Global Test package (Goeman et al., 2004) applies a generalized linear model to determine if a user-defined group of genes is significantly related to a clinical outcome. With the gene set enrichment analysis (GSEA; Mootha et al., 2003) an investigator can test if the members of a gene set tend to occur toward the top or the bottom of a ranked gene list obtained from the differential expression analysis, and therefore are correlated with the phenotypic class distinction.

In this paper, we extend the significance analysis of gene pathways to higher order structures, i.e., networks of pathways whose intersections contain a significant number of differentially expressed genes. Network structure can reveal the degree of coordination of different biological functions as a consequence of the treatment, as well as the presence of "focal areas" in which groups of genes play central roles. We show examples in which some biological functions (related to specific pathways) are biologically relevant for the studied process, due to their position inside the pathway network. This analysis can be extended to groups of genes at the "interface" between pathways, whose imbalance can affect more than one biological function.

Our approach is aimed at understanding how external perturbations, such as gene activation or tumor induction, can induce in various types of cells, cell lines or derived tissues, behaviors that can generate, integrate, and respond to dynamic informational cues.

The broad question that we are trying to answer is how a cell converts perturbations of its signaling activity into a "binary," or at least discrete, decision, resulting in the appearance of a given phenotype. Thus the signaling activity has to be diffused within the cell between and within pathways. A signaling pathway is not a rigid unit, since it can achieve one ore more functions with different subsets of its elements. The communication with other pathways, due to the fact that many elements are shared between several pathways, may be captured by looking at those elements belonging to the interface between pathways.

NETWORKS AND MULTIPLEX FOR BRAIN MODELING AND DATA ANALYSIS

THE PATHWAY MAPPING

According to the theory of neuronal circuits, a neuronal pathway is formed by a series of interconnected neurons that can be associated with a given response. With this definition, we can use methods for pathway analysis initially designed for gene expression studies and based on network theory (Remondini et al., 2005).

Biological pathways can be identified in two ways:

(1) By *a priori* biological knowledge (supervised method)
(2) By a data driven approach (unsupervised method)

The "*a priori* biological knowledge" approach is based on the idea that we have expert information on pathway structure and interconnections. The classical example is the metabolic and signaling pathways as coded by biochemistry experts (see KEGG, ReconX). In the field of neuroscience this corresponds to relying on the vast literature in brain areas identification based on functional imaging.

The data driven approach, is based on some properties of the collected data. For example, we can define a pathway as a set of neurons (a network) whose activity is associated in time. Correlation with its variants (e.g., parametric and non-parametric) can be used for this purposes. Moreover, it is possible to characterize the causality relationships between data (e.g., brain areas) with several methods. Granger causality (Granger, 1988), is a way to test if a time series X Granger-causes Y, by comparing lagged values of X and Y. It can be used both for searching many-to-one or one-to-one relationships, but for a high-throughput dataset (e.g., fNMR voxel data dynamics) it can be computationally very demanding. Other methods are based on partial correlation (for review Mirowski et al., 2009) and also on the so called Gaussian Graphical Models (Yin and Li, 2012).

Relevance networks (Butte and Kohane, 1999) are a popular method for the analysis of time series of expression levels. The basic idea is to construct a network of similarity of the time patterns. Several similarity measures have been used, such as correlation and mutual information. This technique can represent multiple connections, and capture negative as well as positive correlations. Once the matrix containing the similarity measure for all pairs of genes has been computed, a threshold is used to define the significant links in the network. Network validation can be obtained by permutation testing, i.e., by randomly shuffling the time series or just shifting the phase (Schreiber and Schmitz, 2000). A similar approach has been applied to metabolic networks (Martins et al., 2004; Camacho et al., 2005) using computed metabolite correlations to infer changes in regulation using samples from different physiological states.

An alternative approach is offered by graphical Gaussian models (GGM) that use partial correlation as a measure of independence between two genes. Partial correlations are related to the inverse of the correlation matrix, and in GGMs missing edges indicate conditional independence. One of the biggest problems with GGMs is that the correlation matrix is usually singular and cannot be inverted. Different approaches have been proposed to circumvent this problem: restrict the number of elements analyzed to less than the number of samples (Kishino and Waddell, 2000; Waddell and Kishino, 2000; Toh and Horimoto, 2002) use partial correlation coefficients of limited order (de la Fuente et al., 2004; Magwene and Kim, 2004; Wille et al., 2004); approach the matrix inversion as an ill-posed inverse problem through regularization methods (usually via empirical Bayes, such as variance reduction, see Dobra et al., 2004; Schafer and Strimmer, 2005).

Although co-expression is not a direct indication of co-regulation, and it is neither capable to give informations about causal relationship due to its intrinsic symmetry, it is a very useful tool that can be used to interpret the effect of a perturbation in eliciting different phenotypes when combined with an ontology analysis. Moreover, in a time-series correlation-based approach, the choice of the time window can be critical. Most of the state-of-the art analysis (e.g., for defining functional areas in the brain) are based on whole time-series analysis (one long time window) but recent works seem to show that useful information can be extracted also at shorter time scales (Liu and Duyn, 2013). The key point is to assess if the time resolution available by fMRI is enough for these purposes: some simulation works seem indeed to point in this direction, thus justifying the use of small time windows (Honey et al., 2007). The choice of optimal time window size, besides depending on the time resolution of the experimental setup (fMRI and EEG are very different from this point of view), also depends on the characteristic time scales involved in the brain activity process. This also remains an open issue, even if many experimental observations (Buzsáki and Draguhn, 2004) and theoretical models (Haimovici et al., 2013) show a sort of chaotic, or anyway multiscale on a broad range, spectrum of time scales related to brain activity.

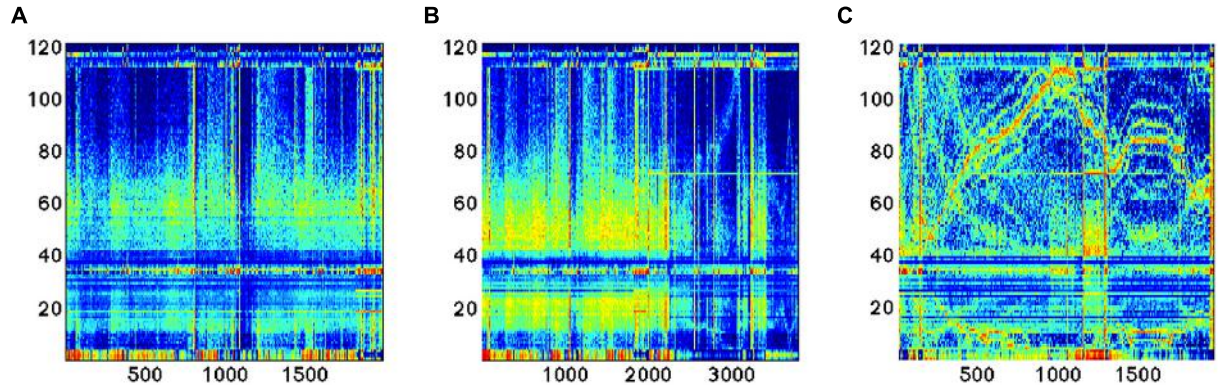


FIGURE 5 | Time series of the 121 features analyzed during EEG recording in three different conditions: (A,B) sleep; (C) dream activity.

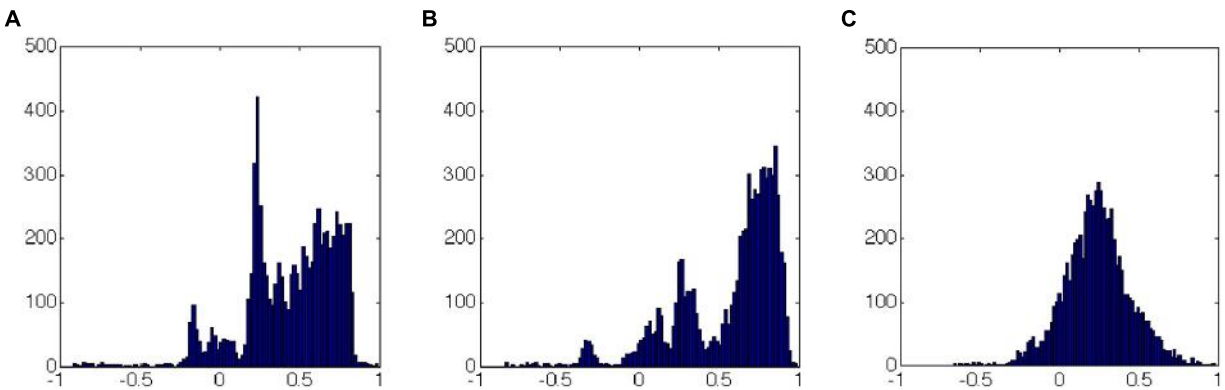


FIGURE 6 | Correlation coefficients distribution (over the whole time series of each experiment) as in Figure 5: (A,B) sleep; (C) dream activity. It can be easily seen that the histograms have similar shapes (in terms of number and range of values) for the two similar rearing states (**A** and **B**, sleep). This picture does not allow to specify if the same links (correlation between features) have similar values.

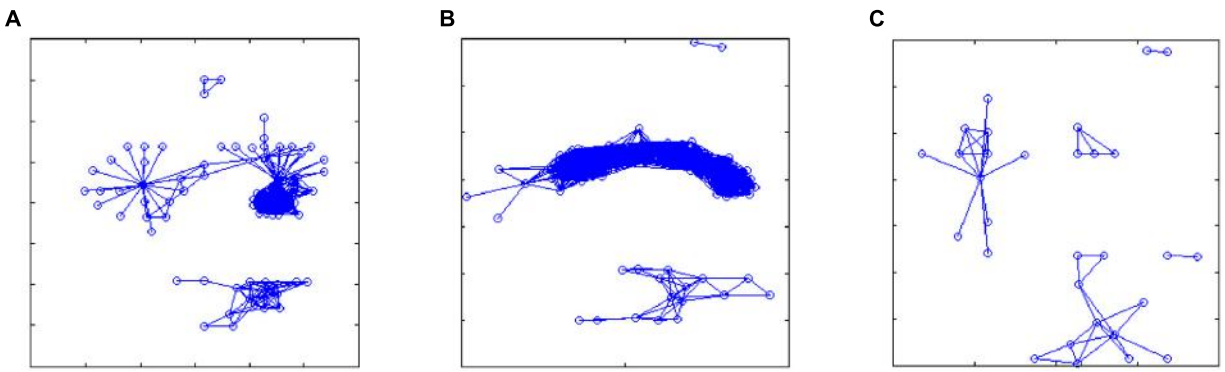


FIGURE 7 | Reconstructed networks in the three cases of Figure 5: (A,B) sleep; (C) dream activity. Starting from the correlation matrices, an arbitrary threshold value was set ($r > 0.8$, but the results were qualitatively similar for a broader range of threshold values, from 0.75 to 0.85) in order to define significant links between features (expressing similarity over time of the linked features). These networks show which features are highly correlated during the different recordings, thus topological observables related to these network may provide a generalized representation of the different rearing states.

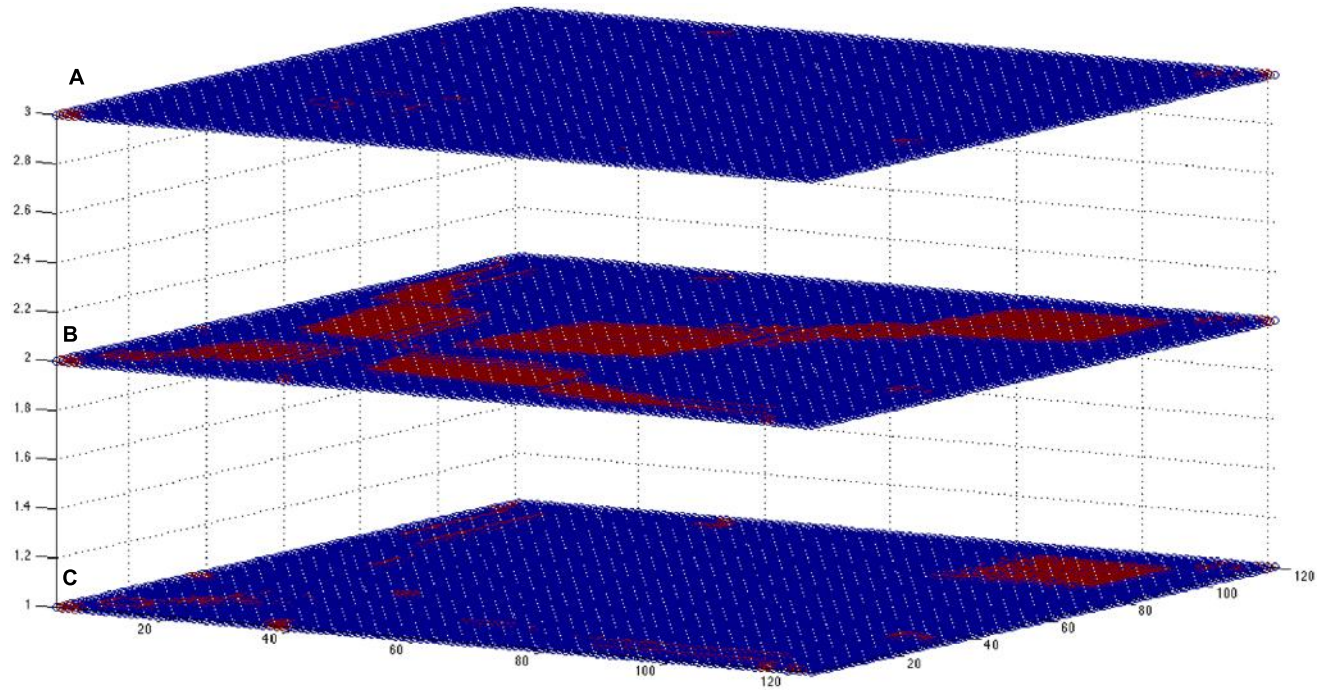


FIGURE 8 | Multiplex-like representation of correlation-based networks. In the picture are shown the three square adjacency matrices (121 by 121, corresponding to the EEG extracted features) obtained for states (**A–C**; from bottom to top, respectively). Blue dots: no link; red dots, existing link. The overlap between the states is higher for cases **(A,B)**, expressing a similar brain state (corresponding to a sleep state): about 5.2% of the possible $N(N-1)$ links are the same for networks **(A)** and **(B)**, whereas for the other intersections the values are about 10 times smaller (0.3–0.5%). Adequate sampling statistics may help to define specific patterns characterizing each rearing state, and similarity measures can be performed to classify the different states.

As an example, here we apply the methods described previously in the cases of reconstruction of the gene expression data to experimental measurements obtained from the EEG device. As it can be seen (**Figure 5**), novel feature extraction methods can emphasize the differences and similarities between brain states. As a second step, a network reconstruction starting from time correlation of the selected features can be performed (**Figures 6** and **7**): the multiplex structure applied on the adjacency matrices in the three states (highlighting the links rather than the node structure of the network, **Figure 8**) allows to find which parts of the network are overlapping for the different states. An increasing number of recordings in different states, applied to different samples (in order to build a "compendium" of observations) will help in building a "library" onto which new experimental observations can be mapped.

CONCLUSION

In our opinion, novel techniques (such as fNMR) and more classical techniques (such as EEG) must be integrated by novel processing and analysis tools, able to extract relevant features of the signal at the single-trace level, but also able to reveal significant interconnections (causal or associative) between traces. Moreover, any possible relevant biological information (e.g., about anatomic regions) must be integrated with the experimental data, in order to enrich the statistical significance of the performed analysis and its biological interpretation.

For these purposes, a great emphasis must be given to feature extraction methods (overcoming the classical Fourier analysis) and to network and multiplex approaches, that may allow to integrate the different informations both in time and space, and to take into account the global complexity of the signal. From this point of view, the panorama of analysis methods for brain data can be enormously enriched by the transfer of knowledge of already existing tools coming from the field of Systems Biology, which is exploiting network approaches and *a priori* biological knowledge since its beginning.

The pathway analysis and its generalization to networks and multiplexes gives the enormous possibility to merge in a unifying framework heterogeneous data as those arising from "omics" measurements and those arising from imaging and EEG. This possibility opens new scenarios for combining microscopic and macroscopic information on single patients that can shed new light in the field of personalized medicine.

ACKNOWLEDGMENTS

Daniel Remondini and Gastone Castellani acknowledge support by the Italian Ministry of Education and Research through the Flagship (PB05) InterOmics and EU projects FibeBiotics (289517) and Mission-T2D (600803), Daniel Remondini, Gastone Castellani and Nathan Intrator acknowledge the European Methods for Integrated analysis of multiple Omics datasets (MIMOmics) (305280) projects.

REFERENCES

Affymetrix Inc.: Technical note: guide to probe logarithmic intensity error (PLIER) estimation. http://www.affymetrix.com/support/technical/technotesmain.affx

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47. doi: 10.1103/RevModPhys.74.47

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Beissbarth, T., and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464–1465. doi: 10.1093/bioinformatics/bth088

Bianconi, G. (2013). Statistical mechanics of multiplex networks: entropy and overlap. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 87, 062806. doi: 10.1103/PhysRevE.87.062806

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115. doi: 10.1037/0033-295X.94.2.115

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: structure and dynamics. *Phys. Rep.* 424, 175–308. doi: 10.1016/j.physrep.2005.10.009

Boorman, S. A., and Harrison, C. (1976). White, social structure from multiple networks. II. Role structures. *Am. J. Sociol.* 81, 1384–1446. doi: 10.1086/226228

Butte, A. J., and Kohane, I. S. (1999). Unsupervised knowledge discovery in medical databases using relevance networks. *Proc. AMIA Symp.* 1999, 711–715.

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929. doi: 10.1126/science.1099745

Cabeza, R., and Nyberg, L. (2000). Imaging cognition II: an empirical review of 275 PET and fMRI studies. *J. Cogn. Neurosci.* 12, 1–47. doi: 10.1162/08989290051137585

Camacho, D., de la Fuente, A., and Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics* 1, 53–63. doi: 10.1007/s11306-005-1107-3

Cai, L., Friedman, N., and Xie, X. S. (2006). Stochastic protein expression in individual cells at the single molecule level. *Nature* 440, 358. doi: 10.1038/nature04599

Cooper, L. N., Intrator, N., Blais, B. S., and Shouva H. Z. (ed.). (2004). *Theory of Cortical Plasticity*. Singapore: World Scientific Publishing.

Davidson, R. J. (1988). EEG measures of cerebral asymmetry: conceptual and methodological issues. *Int. J. Neurosci.* 39, 71–89. doi: 10.3109/00207458808985694

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20, 3565–3574. doi: 10.1093/bioinformatics/bth445

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dobra, A., Hans, C., Jones, B., Nevins, J. R., and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivar. Anal.* 90, 196–212. doi: 10.1016/j.jmva.2004.02.009

Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2008). Critical phenomena in complex networks. *Rev. Mod. Phys.* 80, 1275. doi: 10.1103/RevModPhys.80.1275

Duncan, D., Talmon, R., Zaveri, H. P., and Coifman, R. R. (2013). Identifying preseizure state in intracranial EEG data using diffusion kernels. *Math. Biosci. Eng.* 10, 579–590. doi: 10.3934/mbe.2013.10.579

Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge: MIT Press.

Fisher, R. S., van Emde Boas, W., Blume, W., Elger, C., Genton, P., Lee, P., et al. (2005). Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* 46, 470–472. doi: 10.1111/j.0013-9580.2005.66104.x

Fortunel, N. O., Otu, H. H., Ng, H. H., Chen, J., Mu, X., Chevassut, T., et al. (2003). Comment on 'Stemness': transcriptional profiling of embryonic and adult stem cells and a stem cell molecular signature. *Science* 302, 393. doi: 10.1126/science.1086384

Francesconi, M., Remondini, D., Neretti, N., Sedivy, J. M., Cooper, L. N., Verondini, E., et al. (2008). Reconstructing networks of pathways via significance analysis of their intersections. *BMC Bioinformatics* 9(Suppl. 4):S9. doi: 10.1186/1471-2105-9-S4-S9

Friedman, N., Cai, L., and Xie, X. S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* 97, 168302. doi: 10.1103/PhysRevLett.97.168302

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20, 93–99. doi: 10.1093/bioinformatics/btg382

Grady, D., Thiemann, C., and Brockmann, D. (2012). Robust classification of salient links in complex networks. *Nat. Commun.* 3, 864. doi: 10.1038/ncomms1847

Granger, C. W. J. (1988). Causality, cointegration, and control. *J. Econ. Dyn. Control* 12, 551–559. doi: 10.1016/0165-1889(88)90055-3

Gross, T., and Blasius, B. (2008). Adaptive coevolutionary networks: a review. *J. R. Soc. Interface* 5, 259–271. doi: 10.1098/rsif.2007.1229

Haas, L. F. (2003). Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography. *J. Neurol. Neurosurg. Psychiatry* 74:9. doi: 10.1136/jnnp.74.1.9

Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys. Rev. Lett.* 110, 178101. doi: 10.1103/PhysRevLett.110.178101

Hekstra, D., Taussig, A. R., Magnasco, M., and Naef, F. (2003). Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.* 31, 1962–1968. doi: 10.1093/nar/gkg283

Henriques, J. B., and Davidson, R. J. (1990). Regional brain electrical asymmetries discriminate between previously depressed and healthy control subjects. *J. Abnorm. Psychol.* 99, 22–31. doi: 10.1037/0021-843X.99.1.22

Holme, P., and Newman, M. E. J. (2006). Newman nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 74(Pt 2), 056108. doi: 10.1103/PhysRevE.74.056108

Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10240–10245. doi: 10.1073/pnas.0701519104

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Intrator, N. (2014). *Brain Activity Features: A Continuous Window Into the Mind*. Preprint.

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., and Koch, C. (2013). Neuroscience thinks big (and collaboratively). *Nat. Rev. Neurosci.* 14, 659–664. doi: 10.1038/nrn3578

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

Kishino, H., and Waddell, P. J. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform. Ser. Workshop Genome Inform.* 11, 83–95.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Kurzweil, R., and Grossman, T. (2005). *Fantastic Voyage: Live Long Enough to Live Forever*. London: Rodale.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Milton Keynes: Open University Press.

Law J., and Hassard, J. (eds). (1999). *Actor Network Theory and After*. Oxford and Keele: Blackwell and the Sociological Review.

Liu, X., and Duyn, J. H. (2013). Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4392–4397. doi: 10.1073/pnas.1216856110

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique (Cognitive Neuroscience)*. Cambridge: MIT Press.

Lynch, G., Kessler, M., Arai, A., and Larson, J. (1990). The nature and causes of hippocampal long-term potentiation. *Prog. Brain Res.* 83, 233–250. doi: 10.1016/S0079-6123(08)61253-4

Magwene, P. M., and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.* 5, R100. doi: 10.1186/gb-2004-5-12-r100

Manoli, T., Gretz, N., Gröne, H. J., Kenzelmann, M., Eils, R., and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics* 22, 2500–2506. doi: 10.1093/bioinformatics/btl424

Martins, A. M., Camacho, D., Shuman, J., Sha, W., Mendes, P., and Shulaev, V. (2004). A systems biology study of two distinct growth phases of *Saccharomyces cerevisiae* cultures. *Curr. Genomics* 5, 649–663. doi: 10.2174/1389202043348643

Milanesi, L., Romano, P., Castellani, G., Remondini, D., and Liò, P. (2009). Trends in modeling biomedical complex systems. *BMC Bioinformatics* 10(Suppl. 12):I1. doi: 10.1186/1471-2105-10-S12-I1

Mirowski, P., Madhavan, D., LeCun, Y., and Kuzniecky, R. (2009). Classification of patterns of EEG synchronization for seizure prediction. *Clin. Neurophysiol.* 120, 1927–1940. doi: 10.1016/j.clinph.2009.09.002

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273. doi: 10.1038/ng1180

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* 45, 167–256. doi: 10.1137/S003614450342480

Niedermeyer, E., and Lopes da Silva, F. H. (eds). (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.* Philadelphia, PA: Lippincott Williams & Wilkins.

Pandey, R., Guru, R. K., and Mount, D. W. (2004). Pathway miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* 20, 2156–2158. doi: 10.1093/bioinformatics/bth215

Pauls, S. D., and Remondini, D. (2012). Measures of centrality based on the spectrum of the Laplacian. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 85(Pt 2), 066127. doi: 10.1103/PhysRevE.85.066127

Remondini, D., O'Connell, B., Intrator, N., Sedivy, J. M., Neretti, N., Castellani, G. C., et al. (2005). Targeting c-Myc-activated genes with a correlation method: detection of global changes in large gene expression network dynamics. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6902–6906. doi: 10.1073/pnas.0502081102

Schafer, J., and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21, 754–764. doi: 10.1093/bioinformatics/bti062

Schreiber, T., and Schmitz, A. (2000). Surrogate time series. *Phys. D* 142, 346–382. doi: 10.1016/S0167-2789(00)00043-9

Singleton, A. B. (2014). A unified process for neurological disease. *Science* 343, 497–498. doi: 10.1126/science.1250172

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

Swartz, B. E. (1998). The advantages of digital over analog recording techniques. *Electroencephalogr. Clin. Neurophysiol.* 106, 113–117. doi: 10.1016/S0013-4694(97)00113-2

Szell, M., Lambiotte, R., and Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. U.S.A.* 107, 13636–13641. doi: 10.1073/pnas.1004008107

Toh, H., and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* 18, 287–297. doi: 10.1093/bioinformatics/18.2.287

van Kampen, N. G. (2007). *Stochastic Processes in Physics and Chemistry*, 3rd edn. Amsterdam: North-Holland Personal Library.

Waddell, P. J., and Kishino, H. (2000). Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inform. Ser. Workshop Genome Inform.* 11, 129–140.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511815478

Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana. Genome Biol.* 5, R92. doi: 10.1186/gb-2004-5-11-r92

Yang, H., Long, X. Y., Yang, Y., Yan, H., Zhu, C. Z., Zhou, X. P., et al. (2007). Amplitude of low frequency fluctuation within visual areas revealed by resting-state functional MRI. *Neuroimage* 36, 144–152. doi: 10.1016/j.neuroimage.2007.01.054

Yin, J., and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *J. Multivar. Anal.* 107, 119–140. doi: 10.1016/j.jmva.2012.01.005

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: 03 April 2014; accepted: 10 July 2014; published online: 26 August 2014.
*Citation: Castellani G, Intrator N and Remondini D (2014) Systems biology and brain activity in neuronal pathways by smart device and advanced signal processing. Front. Genet. **5**:253. doi: 10.3389/fgene.2014.00253*
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.
Copyright © 2014 Castellani, Intrator and Remondini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

NETWORK

A network Newman (2003) is the schematical representation of a set of relationships (links) between elements (nodes). Mathematically it can be represented by a NxN square matrix (adjacency matrix, with N the number of nodes) with non-zero elements (equal to one for topological networks and to a real value for weighted networks) where a link exists between two nodes. Other representations are available, eg. a NxL incidence matrix (N number of nodes and L number of links) in which −1 and 1 values are put in each row corresponding to the leaving and the entering node. This formalism represents a sort of "generalized" derivative (or better a finite difference) for a function defined on the nodes, and is the basis for the Laplacian Operator formalism for networks.

CENTRALITY

Measures for nodes, links or network subsets that help ranking these elements based on their topological/structural characteristics. Common centrality measures are connectivity degree (number of incoming/outgoing links), betweenness centrality (ratio of shortest paths passing through a node/link), eigenvalue centrality (like Google PageRank, in which a node is important if it is connected to important nodes, leading to an eigenvalue problem for the adjacency matrix). More recent measures, working in particular for dense and weighted networks, are salient links (Grady et al., 2012) and spectral centrality (Pauls and Remondini, 2012).

MULTIPLEX

A multilayer network (multiplex) represents a set of networks in which the same nodes may appear onto different layers with different relationships. A multiplex can be thought for genes, which proteins appear in Transcription networks (as transcription factors), in Protein–Protein interaction networks (as proteins), and in Metabolic networks (as enzymes controlling metabolic reactions). In neuroscience, we can define a multiplex considering anatomical vs. functional networks, or neuronal networks characterized by different classes of neurotransmitters and receptors.

COMMUNITIES

Networks very often can be dissected into parts, reflecting special relationships between nodes belonging to the same community. These groups can be defined by *a priori* knowledge (like different anatomical or functional regions) or deduced by network topological properties. Clustering methods can be applied to the network as a function of the chosen metrics (e.g., by paths or measures of overlap between node neighborhoods), or communities might arise from dynamical processes applied to the network (e.g., considering transient states of random walks over the network).

NETWORK-BASED STATISTICS

More and more often Systems Biology is integrating common statistical tests (Student's T test, ANOVA and their nonparametric variants) with null models derived from the network structure in which data are embedded. Single-probe statistics (for genes, proteins, neurons) can be scaled up to higher structures like biochemical pathways or brain regions in a recursive manner (Francesconi et al., 2008), and can be enriched by information about significance of their neighbourhood. Moreover, different network structures can be compared and a probability can be assigned to such comparisons in order to assess biological relevance of the observed structure (see a recent comment on Singleton, 2014).

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 30 September 2014
doi: 10.3389/fgene.2014.00342

Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action

Dušica Vidović[1], Amar Koleti[1] and Stephan C. Schürer[1,2]*

[1] *Center for Computational Science, University of Miami, Miami, FL, USA*
[2] *Department of Molecular and Cellular Pharmacology, University of Miami, Miami, FL, USA*

Edited by:
Pietro Lio, University of Cambridge, UK

Reviewed by:
Yaoqi Zhou, Indiana University, USA
Georges Nemer, American University of Beirut, Lebanon

***Correspondence:**
Stephan C. Schürer, Center for Computational Science, University of Miami, Gables One Tower 600, 1320 S. Dixie Highway, Miami, FL 33146, USA
e-mail: sschurer@med.miami.edu

The Library of Integrated Network-based Cellular Signatures (LINCS) project is a large-scale coordinated effort to build a comprehensive systems biology reference resource. The goals of the program include the generation of a very large multidimensional data matrix and informatics and computational tools to integrate, analyze, and make the data readily accessible. LINCS data include genome-wide transcriptional signatures, biochemical protein binding profiles, cellular phenotypic response profiles and various other datasets for a wide range of cell model systems and molecular and genetic perturbations. Here we present a partial survey of this data facilitated by data standards and in particular a robust compound standardization workflow; we integrated several types of LINCS signatures and analyzed the results with a focus on mechanism of action (MoA) and chemical compounds. We illustrate how kinase targets can be related to disease models and relevant drugs. We identified some fundamental trends that appear to link Kinome binding profiles and transcriptional signatures to chemical information and biochemical binding profiles to transcriptional responses independent of chemical similarity. To fill gaps in the datasets we developed and applied predictive models. The results can be interpreted at the systems level as demonstrated based on a large number of signaling pathways. We can identify clear global relationships, suggesting robustness of cellular responses to chemical perturbation. Overall, the results suggest that chemical similarity is a useful measure at the systems level, which would support phenotypic drug optimization efforts. With this study we demonstrate the potential of such integrated analysis approaches and suggest prioritizing further experiments to fill the gaps in the current data.

Keywords: systems-biology, data integration, drug profiling, chemical similarity, kinome profiles, transcriptional signatures

INTRODUCTION

Modern molecular biomedical science relies to a great extent on understanding gene function, and significant progress was made in understanding the roles of numerous individual genes (Silverman and Loscalzo, 2012). However, the most critical unmet medical needs correspond to complex diseases caused by a combination of genetic and environmental factors, such as in cancer.

Many studies have demonstrated that cancer emerges from abnormal protein-protein, regulatory and metabolic interactions caused by concurrent structural and regulatory changes in multiple genes and pathways (Nagaraj and Reverter, 2011; Acencio et al., 2013). Further advances in the prevention, diagnosis and treatment of cancer require a more comprehensive knowledge of the molecular mechanisms that lead to the malignant state. Therefore, understanding cancer pathogenesis requires knowledge of not only the specific contributory genetic mutations but also the cellular framework in which they arise and function (Hong et al., 2008). Cancer cell lines and primary cancer cells have recently been established as powerful model systems to study cancer biology and the pharmacology of drug responses in cancer subtypes. To deconvolute, model, and understand drug sensitivity relies on systems-wide approaches to integrate large-scale biological responses in diseased and healthy cell states, involving various molecular entities such as drugs, proteins, genes, transcripts, cellular, and molecular processes, characteristics (e.g., genetic) of the cell model systems, etc. (Barretina et al., 2012; Heiser et al., 2012; Yang et al., 2013). Of particular interest for the development of novel drugs is their molecular mechanism of action (MoA). MoA describes biochemical interaction through which a drug modulates the corresponding target resulting in a phenotypic response (or pharmacological effect of the drug). Although there are studies linking drug pharmacology to transcriptional

responses (Lamb et al., 2006), the connection to drug targets and the chemical structure of drugs is underexplored, partially because of a lack of large-scale profiling data. Such insights are of particular interest for the rational development of next-generation poly-pharmacology drugs (Hopkins, 2008). Here we present such a study based on data generated at the Library of Integrated Network-based Cellular Signatures (LINCS) project[1].

It is one of the major goals of the LINCS project to generate an extensive reference set of cellular response signatures to representative small molecule and genetic perturbations that can facilitate the development of computational systems-level models of complex diseases and drug action. Common patterns from these data (signatures) include information about gene transcription, protein binding, cell proliferation, cell signaling and other cellular phenotypes with a particular focus on cancer. The LINCS data matrix extends into several dimensions including the model systems (cell lines, primary cells), the perturbations (such as small molecules), and the readout including the genome-wide transcriptional profiles, Kinome-wide binding profiles, and cell-viability and phenotypic profiles against a broad range of cell lines. These biological responses are currently generated, collected, and standardized to facilitate their integration. Data and tools generated in the LINCS consortium are available to the research community via the LINCS website (http://lincsproject.org). The integration of these data and their analysis relies on robust metadata standards developed at LINCS (Vempati et al., 2014). There are also a few recently published approaches that utilize specific LINCS data sets such as transcriptional profiles (Chen et al., 2013a,b) or kinase inhibition profiles (Shao et al., 2013).

Here we apply these standards and report their implementation with a focus on small molecules. We report several case studies involving multi-level integration of such diverse LINCS datasets. Based on large amounts of publically available kinase inhibition and binding data beyond LINCS, we built and applied computational models to fill gaps in the LINCS data matrix to enable much more comprehensive integrative data analyses. We demonstrate some global trends that link chemical features of small molecule perturbations, chemical biology, genomics and cell viability profiles illustrating the complexity and scope of LINCS data and how datasets can be mined. In several examples we show meaningful and biologically interpretable linkages among different signature types in the context of small molecule drugs and known signaling networks.

We hope that our survey and integrative analyses illustrates the wide scope and potential of the LINCS project and will motivate others to use LINCS generated data and knowledge to enhance their research on diverse biological and biomedical problems.

MATERIALS AND METHODS

LINCS ASSAYS AND DATASETS

LINCS datasets cover a range of assays and technologies. Details about LINCS assays, data and tools are available at the LINCS project website (http://lincsproject.org/). For the analyses presented here we used three different types of LINCS data. All data used here can also be obtained via our LINCS Information FramEWork (LIFE) search system[2].

Transcriptional response profiling data (L1000)

For the purposes of this study we selected two L1000 experiments (Peck et al., 2006) with fairly dissimilar cell lines, A549 (non-small cell lung carcinoma) with 1027 compounds tested, and VCAP (prostate carcinoma) with 741 compounds tested, in order to compare expression profiles among the same cell lines as well as between different ones. Although there is no simple measure of cell line similarity (LINCS is one of the first systematic efforts that contribute to the large-scale generation of cellular response signatures), for the purposes of this study we consider these cell lines in the basis of their origin from different organs. In total, here we investigate 1768 "is_ gold" signatures, corresponding to 1,729,104 data points (total number of Z-scores; perturbagens $\times$ transcribed genes measured $\times$ cell lines). All LINCS L1000 data and signatures are available at the Broad LINCS Cloud[3]. For more details on the L1000 data see Supplementary Material.

KINOME-wide binding profiles (KINOMEscan)

LINCS kinase biochemical profiles were generated at Harvard Medical School (HMS) using the DiscoveRx KINOMEscan technology[4], which is a competition binding assay. A panel of 478 purified kinases was profiled against 78 small molecule compounds. However, the majority of LINCS compounds were not profiled in the KINOMEscan assay and we therefore generated predicted KINOME-wide inhibition/binding profiles based on classification models (described below).

Cell growth inhibition profiles

Cell growth inhibition datasets (assay developed at the Center for Molecular Therapeutics at Massachusetts General Hospital) (McDermott et al., 2007; Garnett et al., 2012) were retrieved from the LIFE database and the data were aggregated by averaging replicates. 39 small molecules were tested against 582 previously standardized cell lines at different concentrations (in the range from 0.004 to 15 μM) and one time point (72 h) and number of surviving cells counted. The measured cell viability values center around mean of 81% (corresponding to 19% growth inhibition) with a standard deviation of 31.68 across all concentrations.

SMALL MOLECULE CHEMICAL STRUCTURE STANDARDIZATION, IDENTIFICATION, AND ANNOTATIONS

Compound information for small molecule perturbagens was received from the LINCS Data Production centers, HMS and Broad Institute. To identify unique and common compounds required a rigorous structure standardization pipeline that we implemented for the LINCS program. We used Pipeline Pilot 8.0 (Pipeline Pilot, 2011) components to generate the structures and remove addends and they were then subjected to the PubChem[5] chemical structure standardization procedure using

[1] http://lincsproject.org/

[2] http://life.ccs.miami.edu

[3] http://lincscloud.org/

[4] http://www.discoverx.com/technologies-platforms/competitive-binding-technology/kinomescan-technology-platform

[5] http://pubchem.ncbi.nlm.nih.gov/

the Power User Gateway (PUG) service. In order to further identify PubChem CIDs we used additional service provided by PubChem PUG. The entire process was automated in a custom protocol using Pipeline Pilot. Using this process, a total of 5364 (as of October, 2013) unique LINCS compounds were obtained and LINCS small molecule (LSM) IDs assigned. More details on the procedure can be found in the Supplementary Material.

Additional information and annotations for the standardized structures were retrieved from PubChem but also from numerous external resources including DrugBank[6], the NCBI[7] MLP probe reports, the NCATS pharmaceutical collection (NPC), and the Protein Data Bank (PDB) (Berman et al., 2003). Compounds were annotated as approved drugs, kinase inhibitors, MLP probes, PDB ligands and, if information available, as kinase inhibitor of type I or type II (defined by the kinase ATP-binding site conformation in the ligand-bound form) (Dar and Shokat, 2011). All compound information can be queried, browsed and downloaded via the LIFE search system (http://life.ccs.miami.edu) and the LIFE project website[8].

To characterize the diversity in chemical space of the tested LINCS compounds, we generated a histogram of their pairwise chemical similarities based on the Tanimoto metric using extended-connectivity fingerprints of length 4 (ECFP4) (Rogers and Hahn, 2010).

Based on unique LSM IDs we identified overlap of screened compounds among the different LINCS assays. While many compounds were tested in the L1000 gene-expression assay at the BROAD Institute, only few of those were tested in different assays at HMS.

SMALL MOLECULE KINASE INHIBITOR MODELS

We generated predicted kinase inhibition/binding profiles for all LINCS compounds to fill missing information of those compounds not (yet) tested in the HMS KINOMEscan assay. For that purpose we built Laplacian-corrected naïve Bayesian classification models using the procedure previously described (Schurer and Muskal, 2013); the models used here were rebuilt based on the new kinase inhibition data that doubled in the meantime illustrating rapid growth in published kinase inhibition data. Small molecule kinase activity data was extracted from the Q2 2013 release of the Kinase Knowledge Base (KKB, Eidogen-Sertanty)[9]. After standardization and aggregation based on unique kinases and compounds as previously described, the data amounted to more than 510,000 kinase structure data points with more than 270,000 actives (pIC50 $>$ 6) and more than 590,000 total compounds covering the entire human Kinome. For each model, the number of total data points and actives was considered and only models for kinases with reasonable amount of data were built. For computational kinase profiling, we selected only models with the area under the receiver operating characteristic (ROC) curve greater than 0.9 and if they were based on at least 20 unique activity data points with 10 of them being considered active (pIC50 $>$ 6). This selection resulted in 229 kinase models for which we could make confident predictions (for these 229 kinase models the additional information regarding their characteristics [target, number of data points, number of actives, ROC score, and enrichment factor for 1% for leave-one-out cross validation] can be found in Dataset 1 in the Supplementary Material). The model classifier outcome is a prediction of a compound being active (prediction value is true) or inactive (prediction value is false) for a given kinase. The outcome of performing all models against the LINCS compounds was converted into a 229-bit binary fingerprint for each compound.

KINASE AND SMALL MOLECULE KINASE INHIBITOR ANNOTATIONS

To integrate KINOMEscan results and kinase models, we manually mapped them to Uniprot, standardized descriptions including mutations and posttranslational modification and we added external annotations such as protein name, symbols, IDs and alternate names, and also important details such as gatekeeper amino acid residues. We organized all kinase domains by an extended phylogenetic classification tree that we based largely on the Sugen kinase classification (Manning et al., 2002)[10].

For LINCS standardized compounds a set of additional annotations were derived from the LINCS datasets. We defined active, selective, group selective and promiscuous kinase inhibitors based on the number and the group membership of kinases that are measured in the KINOMEscan assay. Compounds were considered active if they inhibited a kinase more than 90%. If a compound is active toward 5 or more kinases (belonging to different kinase groups) it was considered promiscuous. Compound was defined as selective kinase inhibitor if it is active toward only one kinase, or group selective if it was active only against kinases from the same kinase group. This data is available via the LIFE search system and the LIFE project website.

CELL LINES ANNOTATIONS

Numerous cancer cell lines and non-transformed primary cultures are used as disease model systems in the LINCS project. To facilitate integration and analysis of large-scale cell-based screening profiles generated at LINCS, cell lines were systematically annotated with controlled terms identifying associated organs and diseases (Vempati et al., 2014). Ongoing and future LINCS datasets are also being expanded toward primary tissues, iPS cells and their differentiated derivatives. Here we leverage disease annotations, which are available from the HMS LINCS website[11], and can also be queried in the LIFE search system (http://life.ccs.miami.edu).

BIOPROFILE- AND CHEMICAL STRUCTURE-BASED FINGERPRINTS AND SIMILARITIES

To facilitate comparative analysis of LINCS datasets, we defined several bioprofile fingerprints for tested compound. These bioprofile fingerprints were constructed based on categorical outcomes (active/inactive) in the different LINCS profiling assays. The Tanimoto metric was then used as a

[6]http://www.drugbank.ca/
[7]http://www.ncbi.nlm.nih.gov/
[8]http://lifekb.org/
[9]http://eidogen-sertanty.com/kinasekb.php
[10]http://kinase.com/
[11]http://lincs.hms.harvard.edu/

similarity measure of these profiles (similarities KinomeSim, KinomePredSim, and TranscriptSim for KINOMEscan, predicted kinase inhibition profile, and transcriptional expression profile, respectively). Advantages of this approach include simplicity (binary fingerprints) and computational efficiency (i.e., compute Tanimoto similarities). Chemical similarity of LINCS compounds (ChemSim) was determined based on topological fingerprints derived from the chemical structures also employing the Tanimoto metric. The definition of the fingerprints is provided in the Supplementary Material.

KINASE ENRICHMENT IN CELL GROWTH INHIBITION DATA

We integrated and analyzed the KINOMEscan data and cell growth inhibition assay data, which were retrieved from the LIFE database (http://life.ccs.miami.edu). KINOMEscan data consists of 78 small molecules tested against the panel of 478 kinases (including clinically relevant mutants, lipid, atypical, and pathogen kinases), corresponding to 382 unique kinase UniProt IDs. Cell growth inhibition data represents results of 39 small molecules tested in 582 cell lines (standardized as described above) at different concentrations (in the range from 0.004 μM to 15 μM) and one time point (72 h). Twenty one compounds were tested across the two described datasets and were used to integrate the data. For each kinase we calculated an enrichment score to reflect how much more likely it is to find activity in the cell growth inhibition assay among compounds that inhibit that particular kinase over the background probability of a compound inhibiting cell growth (the further details are provided in the Supplementary Material). Kinase enrichment scores were further used in the hierarchical clustering analysis performed by TIBCO Spotfire software (TIBCO Spotfire, 2013). Clustering was based on the single linkage method and the Euclidian distance was used as a distance measure.

PI3K/AKT/mTOR PATHWAY ANALYSIS

In order to demonstrate systems-level data integration, we considered kinases in the PI3K/AKT/mTOR signaling pathway (Laplante and Sabatini, 2012). We identified and downloaded 213 proteins (including cellular localization variation) from PI3K/AKT/mTOR pathway from Reactome (Joshi-Tope et al., 2005; Vastrik et al., 2007). By matching their genes to the standardized kinase genes symbols in the KKB, we identified 26 unique kinases. We then queried the aggregated KKB (the data that was also used for building the models) for those small molecules with a pIC50 value greater than 6 against any of these kinases and we identified 24,158 unique kinase inhibitors. Their (standardized) structures were compared to the LINCS compounds and we identified an overlap of 35 compounds. Based on the KKB activities, they inhibit 21 out of 26 PI3K/AKT/mTOR pathway kinases. For these 35 compounds that theoretically affect PI3K/AKT/mTOR pathway, we analyze their L1000 responses and the effect on the cell growth inhibition.

SYSTEMATIC PATHWAY ANALYSIS

For the systematic pathway analysis our starting point was the curated pathway database of the National Cancer Institute (NCI)[12]. We retrieved the tab delimited file "NCI-Nature Curated Pathway–UniProt mapping" from their website (http://pid.nci.nih.gov/download.shtml). This file contains a total of 8420 records, which represent a combination of 2688 unique Uniprot IDs and 224 pathways (as of April 3, 2014).

In order to identify kinases, we grouped proteins by the pathways and compared their UniProt IDs to the kinase annotations in the KKB. For each pathway we further identified LINCS compounds that were predicted (by the kinase models, as described above) to be active for the kinases identified in the given pathway, and consequently active in that pathway. For such pathway-active compounds we compared their transcriptional similarities and computed p-values between TranscriptSim of pathway-active and pathway-inactive LINCS compounds in order to demonstrate that (predicted) pathway-active compounds lead to (statistically) significantly more similar transcriptional profiles than the pathway-inactive compounds.

STUDENT *T*-TEST CALCULATIONS

All Student t-test calculations reported here were performed using the R Statistics[13] component "R Two-Variables Tests" implemented in Pipeline Pilot 8.0.

RESULTS

CHARACTERIZATION OF LINCS SMALL MOLECULE PERTURBAGENS

Small molecules tested in different LINCS datasets were compiled, and after removing salts and addends, were submitted to the PubChem web services first for the compound standardization and then for retrieving the PubChem CID identifiers. Unique LSM parent compound IDs were assigned based on the standardized chemical structure representations; a total of 5364 unique compounds were identified across the LINCS assays. Among them, we identified previously known kinase inhibitors, approved drugs, MLP probes, PDB ligands etc. (described in the Materials and Methods). These annotations are illustrated in **Figure 1**; they are available and can be browsed and queried at the LIFE project website (http://lifekb.org/) and the LIFE search engine (http://life.ccs.miami.edu).

We explored the diversity of compounds in the LINCS chemical space by pairwise Tanimoto similarities based on extended-connectivity fingerprints (**Figure 2**).

As shown in the similarity histogram (**Figure 2**), the distribution is skewed toward low similarity suggesting LINCS compounds are fairly diverse (Tanimoto coefficient below 0.4). LINCS compounds were selected by the centers to cover a broad biological space including known drugs, kinase inhibitors and probes from the Molecular Libraries program.

OVERLAP OF LINCS COMPOUNDS AND CELL LINES ACROSS ASSAYS

Cell lines were previously standardized by a joint effort of several LINCS centers (Vempati et al., 2014).

Standardized compounds and cell lines were compared across the LINCS Data Generation Centers and selected assays. One hundred and fifty compounds and thirty one cell lines were tested

[12]http://pid.nci.nih.gov/index.shtml
[13]http://www.r-project.org/

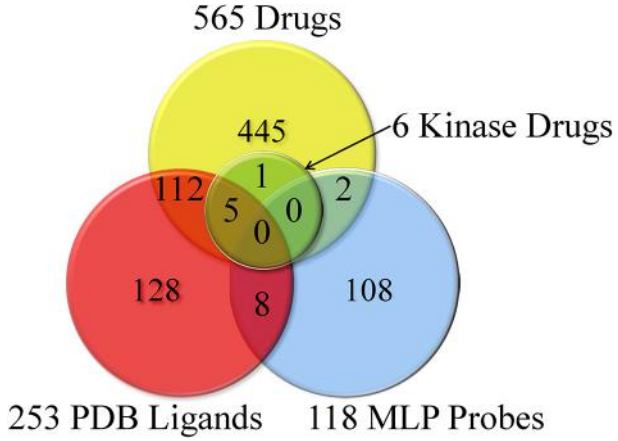


FIGURE 1 | Identified known PDB ligands, MLP probes, approved drugs, and kinase drugs among LINCS compounds.

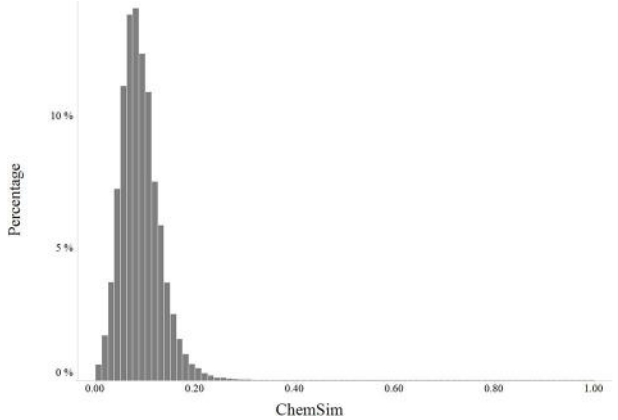


FIGURE 2 | Distribution of pairwise chemical similarity of LINCS compounds.

at both centers (HMS and Broad) across different assays. For the assays considered in this study the overlap between tested compounds and cell lines is shown in **Figure 3**.

From this analysis it becomes obvious that only a small number of compounds were tested in several different assays limiting comprehensive analysis. In order to generate data that would facilitate cross-datasets integration, we built and applied 229 small molecule kinase inhibition models (as described in Materials and Methods) to predict the kinase inhibition profiles for all LINCS compounds. We used these predictions to fill the gaps in the experimental data and to deconvolute the trends between biological responses as described below.

INTEGRATION AND ANALYSIS OF KINASE PROFILING AND CELL GROWTH INHIBITION PROFILING DATASETS

The integration of Kinome-wide small molecule inhibition profiles and phenotypic responses offer a powerful approach to deconvolute likely mechanisms of action of pharmacologically active compounds. Similar, cell line panels, in particular cancer cell lines, are an established approach to characterize small molecule pharmacologically. Using standardized LINCS KINOMEscan and cell growth inhibition signatures generated for the same compounds enables us to map chemical biology binding profiles to cancer cell viability profiles with the potential to contribute to the identification of key kinases and pathways that are relevant for specific cancer subtypes. To investigate this, we generated all combinations of tested kinases and cell lines and for each combination computed a kinase enrichment score that quantifies how much more likely a compound is to be active if it is an inhibitor of a given kinase over the background probability of inhibiting cell growth (see Materials and Methods). Scores of greater than one indicate that inhibitors of that kinase are more likely to inhibit cell growth, suggesting that the pathways to which these kinases belong may be involved in cell death (desirable outcome for the cancer cell lines). Conversely, enrichment scores of less than minus one indicate that such inhibitors would be less likely to kill the cells.

Using the enrichment scores, we performed hierarchical clustering of kinases and cell lines. The resulting heat map is shown in **Figure 4** where red areas represent high kinase enrichment scores, white no enrichment and blue derichment; gray area reflect combinations of kinases and cell lines without overlapping compounds tested in two assays.

KINASE ENRICHMENT AND DERICHMENT IN CANCER

Although there is no clear clustering pattern of kinases vs. diseases in **Figure 4** (which cannot be expected in a relatively limited dataset and cell line model systems), we can still identify individual kinases that are enriched in certain cell lines. For example, kinases ALK, PRKD1, MYLK, CAMKK1, CAMKK2, DAPK3, EGFR, GAK, DCAMKL1 emerge to be more relevant for the lung squamous cell carcinoma (few cell lines originating from this diseased tissue) while kinases MRCKA, MRCKB, DMPK2, HIPK4, CDK2, CDK8, CDK11, PIK3CA, NEK5, ERK3, and CSNK1D appear to be not affected by compounds causing cell death in the same cell lines. Therefore, after identifying kinases that are enriched in one (or several) disease, one could possibly identify novel drug targets or previously known targets that show activity in a new disease and therefore find a case for drug repurposing. In this way, previously unknown side effects of a compound may be discovered and off-targets can be identified among a subset of enriched kinases.

Our analysis approach illustrates how LINCS data can potentially be leveraged to gain important insight into molecular mechanisms that lead to the cell malignant state, especially in the future with the currently expanding LINCS data. The results shown here should be considered as an illustration for data integration and how they can be interpreted. Even with this limited dataset, we were able to identify several examples of known drugs that would confirm potential conclusions derived from this analysis. For example Lapatinib, an approved drug for breast cancer is very potent in the MCF7 breast cancer cell line by killing 83% of cancerous cells (at 2.5 μM). Its known drug target is EGFR. We also found that this drug inhibits EGFR at 100%, as well as majority of its other modifications/mutations in the KINOMEscan panel.

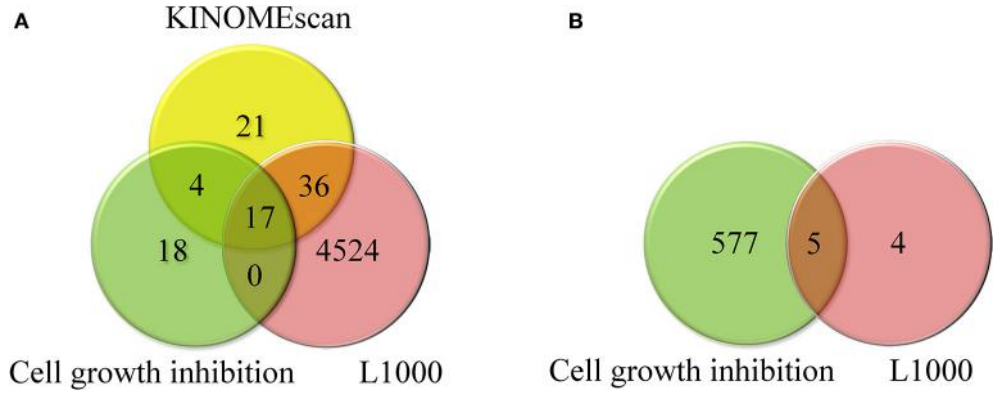


FIGURE 3 | (A) Overlap between compounds tested in KINOMEscan, cell growth inhibition, and L1000 assays and **(B)** overlap between cell lines tested in cell growth inhibition and L1000 assays.

BIOCHEMICAL AND PHENOTYPIC RESPONSE SIGNATURES ARE RELATED AND INTERPRETABLE BASED ON CHEMICAL SIMILARITY

After defining bio-fingerprints to represent cellular signatures generated in the number of LINCS assays (as described in Materials and Methods) we analyzed them to identify correlations and trends between different biological and cellular phenotypic response profiles.

Kinome-wide binding activity (KINOMEscan) profiles

We calculated pairwise Tanimoto similarities (KinomeSim) based on the kinase binding (KINOMEscan) profiles for 78 compounds that were tested in that assay (see Materials and Methods). For the same compounds we computed the corresponding pairwise molecular similarities (ChemSim). KinomeSim thus represents the similarity of a compound pair based on their biochemical (kinase) binding profile while ChemSim quantifies the similarity of two compounds based on features of their chemical structures. Chemical structure similarity is an important concept in cheminformatics where it is generally assumed that more structurally similar compounds are more likely to have similar biological activity (similarity property principle) (Martin et al., 2002). Here we apply this concept to a biological profile. **Figure 5** illustrates the global relationship between pairwise biological profile and chemical similarities; specifically ChemSim is binned and within each bin the average KinomeSim is calculated and shown as the corresponding bar height. As **Figure 5** illustrated, there is a general trend that highly similar compounds also have very similar kinases panel activity (KINOMEscan) profiles. A two-sided Student *t*-test confirmed the statistical significance of this trend. For example using a ChemSim cutoff of 0.8, which can be considered reasonable similar for the fingerprints applied here (see Materials and Methods), the average biological profile similarities of the corresponding KinomeSim distributions are (statistically) significantly different with a *p*-value of 1.9·10–61.

Predicted small molecule kinase inhibition profiles

Using predicted kinase inhibition profiles rather than the experimental binding profiles allowed us to investigate a much larger number of compounds. Whereas KINOMEscan profiles were available for 78 compounds, we generated predicted kinase inhibition profiles for all 5364 LINCS standardized compounds as described in Material and Methods. Although we don't expect perfect predictions, we have shown that the predictors are highly accurate (Schurer and Muskal, 2013); we only applied models with sufficient data and very good cross-validation performance. An important characteristic of the kinase classification models is that they are derived from a large corpus of published and patented results comprising many different assay technologies and assay conditions aggregated by unique chemical structures and kinase protein target. It may therefore be the case that such results are in fact more robust in terms of reproducibility as oppose to comparing just two different assay methods or assay conditions, which can sometimes give considerably different outcomes (Haibe-Kains et al., 2013). It was therefore of much interest how the predicted profiles would perform statistically.

In the same manner as described above, we compared pairwise similarities based on (predicted) kinase activity profiles (KinomePredSim) and chemical structural features (ChemSim). **Figure 6** illustrates the global trend.

As before, structurally similar compounds exhibit similar (in this case predicted) biological response profiles. We corroborated this trend by a *t*-test comparing two distributions of KinomePredSim corresponding to a ChemSim split of 0.8 (reflecting similar and dissimilar compound pairs) and obtained a *p*-value of 1.62•10–79. As expected no such trend is observed when the kinase predictions are randomized.

Gene expression (L1000) profiles

After demonstrating a robust, perhaps expected trend that the similarity of compounds based on their biochemical activity profiles (KINOMEscan as well as predicted) increases significantly with their chemical similarity, it was of interest to compare chemical similarity to gene expression similarity. To evaluate transcriptional similarity we considered not just one response (active vs. inactive) for each feature (e.g., kinase target), but two responses, overexpressed and underexpressed for each feature (i.e., gene); this was implemented in a binary fingerprint simply by doubling the features as described in Materials and Methods. With that

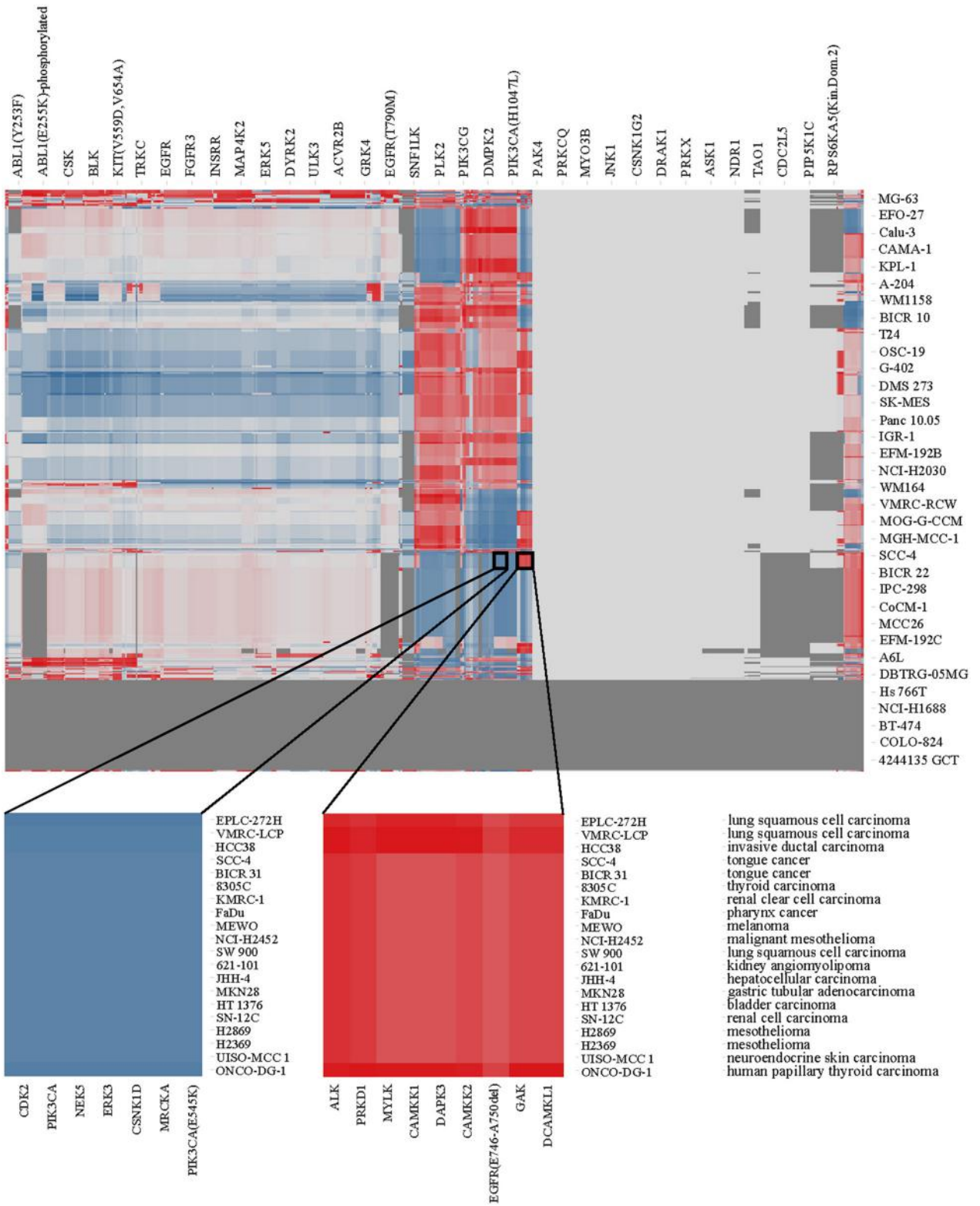


FIGURE 4 | Heat map of kinase enrichment across cell lines as described in the text. Two circles focused on the area of high enrichment (red) and derichment (blue), respectively (the gradient color mode range shown from red, for the maximum enrichment score of 3.13, via white for average score of −0.32, to blue color corresponding to minimum derichment score of −5.50). Kinases, cell lines and diseases are annotated.

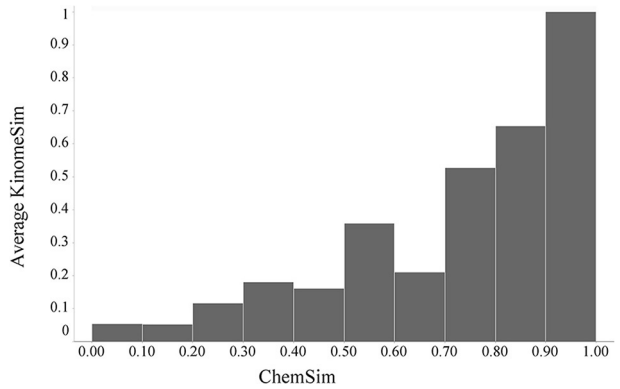


FIGURE 5 | Global trend of kinase binding profile similarities (KinomeSim) and chemical structure similarities (ChemSim) for 78 compounds, illustrated as average KinomePredSim values by ChemSim ranges.

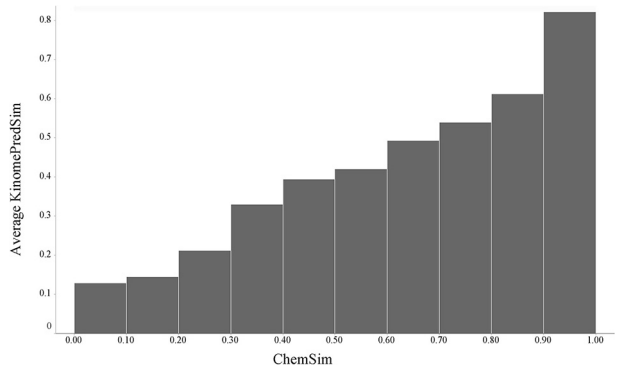


FIGURE 6 | Global trend of pairwise predicted kinase inhibition profile similarities (KinomePredSim) and chemical structure similarities (ChemSim) for 5364 compounds, illustrated as average KinomePredSim values by ChemSim ranges.

we can again compare pairwise similarities, this time based on the gene expression profiles (TranscriptSim) vs. chemical similarity (ChemSim). We found that a similar global trend holds even in this case, when there are no direct interactions between small molecule perturbagens and the molecular entity underlying the biological profiles, i.e., transcribed gene in this case. **Figure 7** illustrates this trend for two dissimilar cell lines, A549 (non-small cell lung carcinoma) and VCAP (prostate carcinoma).

As before we quantified the statistical significance of this trend by the two-tailed *t*-test using a ChemSim cutoff of 0.8 to differentiate similar vs. dissimilar compounds. The *p*-values of the corresponding TranscriptSim distributions are 2.06•10–14 and 9.64•10–14, for the A549 and VCAP cell lines, respectively.

In the same way we also compared compound L1000 response profiles across both cell lines. Although there is the general trend of increasing transcriptional similarity with molecular similarity holds, the effect is much smaller (about half the average similarity) compared to the trend on one cell line alone (shown in **Figure 8**). This is expected, because the cell lines can be expected to have a very different response to the same compounds; in particular that is the case for kinase inhibitors that was evaluated above. The response of kinase inhibitors tested (for example) in A549 and VCAP growth inhibition assays can be explored in our LIFE software (http://life.ccs.miami.edu). A global effect across two very different cell lines is noteworthy and probably related to conserved pathways.

Relating small molecule predicted kinase inhibition profiles and gene expression profiles

After establishing a general global trend of biochemical and transcriptional similarity with compound similarity, it was of interest to compare gene expression (L1000) signatures and kinase inhibition profiles. Because of the limited number of experimental KINOMEscan profiles and encouraged by our results, we compared compound pairwise similarities based on transcriptional response profiles to the predicted kinase inhibition profiles. As shown in **Figure 9**, compounds that are more similar based on their biochemical kinase profile are also more similar with respect to changes in gene expression. We estimated statistical significance of this trend for the KinomePredSim cutoff of 0.8 (above the cutoff considered similar biochemical kinase profile) with the *p*-values of 1.28·10–21 and 6.70·10–30 for A549 and VCAP, respectively. While it is known that kinases are mechanistically related to downstream gene expression via various signaling pathways and networks, these results suggest some level of global systems-wide stability of gene transcription with respect to modulating the entire human Kinome. We did not incorporate any systems-level information to group kinases (this is described in more detail below), but look only at the global profiles.

Earlier observed trend of increasing transcriptional similarity for more similar chemical purturbagens reasonably could be rationalized based on the assumption that more similar compounds are more likely to bind to similar targets. The kinase profile similarity analyses above confirm that assumption, even at large scale of more than 5000 compounds using predicted kinase profiles. To investigate further the dependencies of chemical similarity, biochemical similarity and transcriptional similarity we analyzed TranscriptSim vs. KinomePredSim for different cutoffs of ChemSim as shown for the two cell lines, A549 and VCAP in **Figures 10A,B**, respectively. Specifically **Figure 10** compares three ChemSim cutoff values, namely 1 (keep all compounds, green), 0.8 (remove compound pairs with similarity higher than that, blue), and 0.5 (leave practically only non-similar compounds, red).

As **Figure 10** illustrates, as chemically similar compounds are removed from the analysis, the observed trend between transcriptional similarity and biochemical similarity of compound pairs decreases, but still holds even for only dissimilar compounds (ChemSim cutoff 0.5). This is the case again for two very different cell lines.

To evaluate these trends statistically, we performed Student *t*-tests for the different datasets corresponding to a ChemSim cutoffs of 0.8 (426,331 and 219,163 compound pairs for A549 and VCAP, respectively) and 0.5 (425,452 and 218,648 of compound pairs for A549 and VCAP, respectively). In both cases the

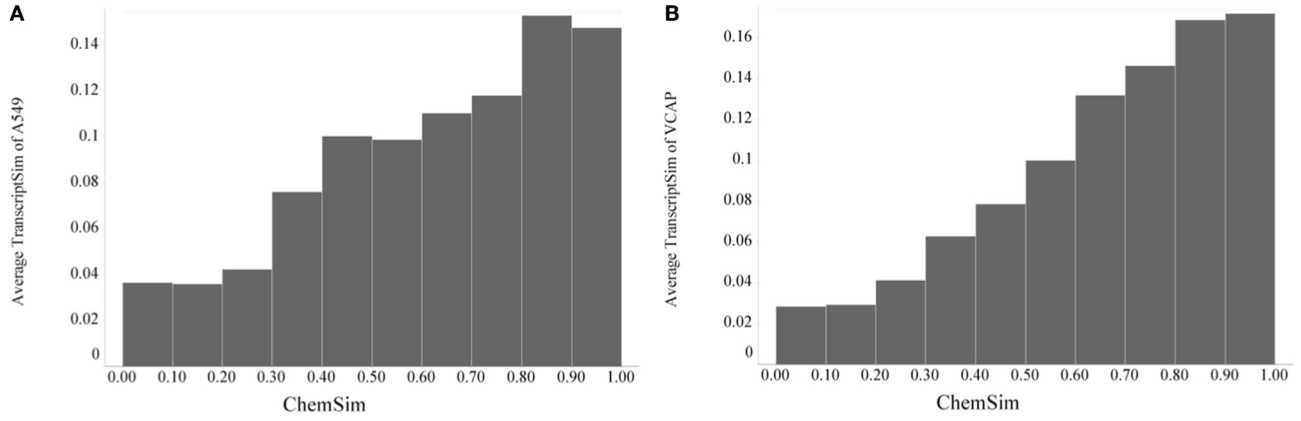


FIGURE 7 | Global trend of pairwise transcriptional similarity (TranscriptSim) in (A) A549 cells and (B) VCAP cell and chemical structure similarities (ChemSim) for 1027 and 741 compounds per cell line, respectively, illustrated as average TranscriptSim values by ChemSim ranges.

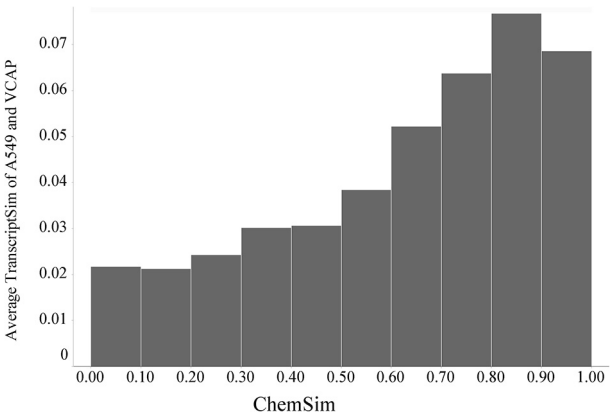


FIGURE 8 | Global trend of pairwise transcriptional similarity (TranscriptSim) across A549 and VCAP cells and chemical structure similarities (ChemSim) for 1027 and 741 compounds per cell line, respectively, illustrated as average TranscriptSim values by ChemSim ranges.

dataset was split by a KinomePredSim cutoff of 0.8 (similar and dissimilar based on their predicted kinase inhibition profile) and *p*-values characterizing the difference in mean for the corresponding distributions of transcriptional similarity were calculated. The *p*-values for the ChemSim cutoff of 0.8 are 2.15•10–19 and 1.15•10–28 for A549 and VCAP cell lines, respectively, while for the ChemSim cutoff of 0.5 the *p*-values are 0.004 and 0.014 for A549 and VCAP cells, respectively. These results confirm that the observed trend between the biochemical kinase profile and transcriptional profile similarities is statistically significant even for structurally dissimilar compound pairs. This is noteworthy as a global trend suggesting that transcriptional response signatures may be modeled based on biochemical response profiles alone. With this, it is of course not surprising that this trend is more pronounced with increasing chemical similarity, because—as shown above—chemical similarity would results in higher biochemical similarity. For example, **Figure 11** illustrates two highly similar compounds (ChemSim = 0.88) with high KinomePredSim (of 0.70) and TranscriptSim (of 0.56).

An example of high biochemical similarity and high gene expression similarity for two structurally dissimilar compounds is illustrated in **Figure 12**; specifically ChemSim = 0.25, KinomePredSim = 0.83, and TranscriptSim = 0.47. Identifying pharmacologically similar, but structurally diverse compounds as demonstrated here using LINCS signatures, is an important approach in drug lead development; for example to overcome undesired physicochemical properties, such as solubility or brain penetration, or for patent reasons.

SYSTEMS-LEVEL INTEGRATION AND ANALYSIS OF LINCS SIGNATURES

The above analyses suggested that the transcriptional profiles are correlated (to some extent) to the MoAs of kinase inhibitors as characterized by their kinase inhibition profiles. We therefore anticipated that small molecule perturbagens that affect same pathway would also exhibit similar transcription. To demonstrate that in a specific example, we selected and analyzed the PI3K/AKT/mTOR pathway, which is in the regulation of cell apoptosis and a target of many cancer drug discovery studies. For this example we extracted experimental kinase inhibitor activities from the KKB to identify those compounds that would interact physically with a protein target in the pathway.

In addition we pursued a systematic approach analyzing transcriptional response for all currently available pathways from the NCI database. Here we used the kinase models (described above) to predict LINCS compounds that could affect kinases in the considered pathways.

PI3K/AKT/mTOR pathway analysis

For 21 kinases previously identified in the mTOR pathway we identified (using the KKB) 35 active kinase inhibitors among LINCS compounds (see Materials and Methods; see Supplementary Material Dataset 2 for the list of mTOR pathway proteins, 21 mTOR pathway kinases with the inhibition data, and 35 active compounds). For these, pathway-active, compounds we compared L1000 fingerprint similarities. We found that for the two cell lines, the pairwise mTOR pathway inhibitors' L1000

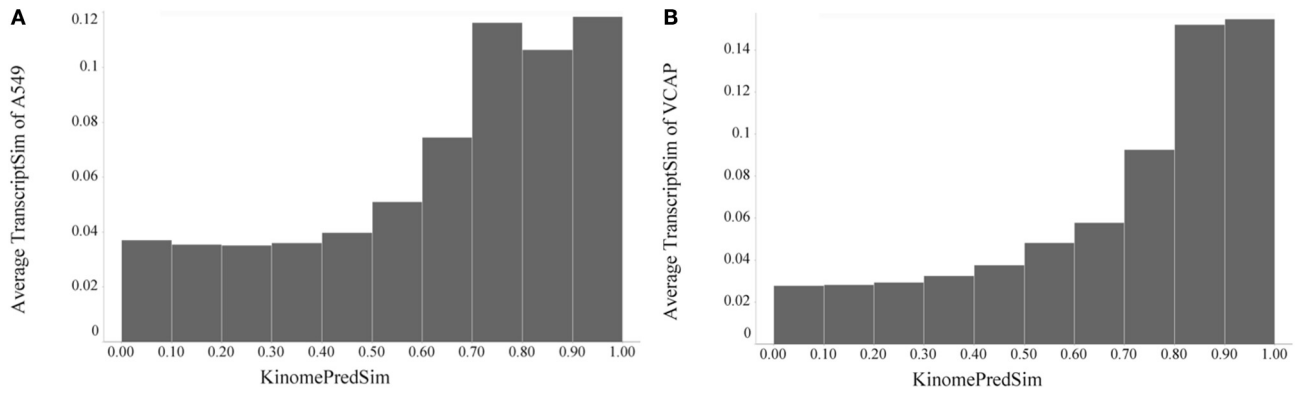


FIGURE 9 | Global trend of pairwise transcriptional similarity (TranscriptSim) in (A) A549 cells and (B) VCAP cells as a function of predicted kinase profile similarity (KinomePredSim) for 1027 and 741 compounds per cell line, respectively, illustrated as average TranscriptSim values by KinomePredSim ranges.

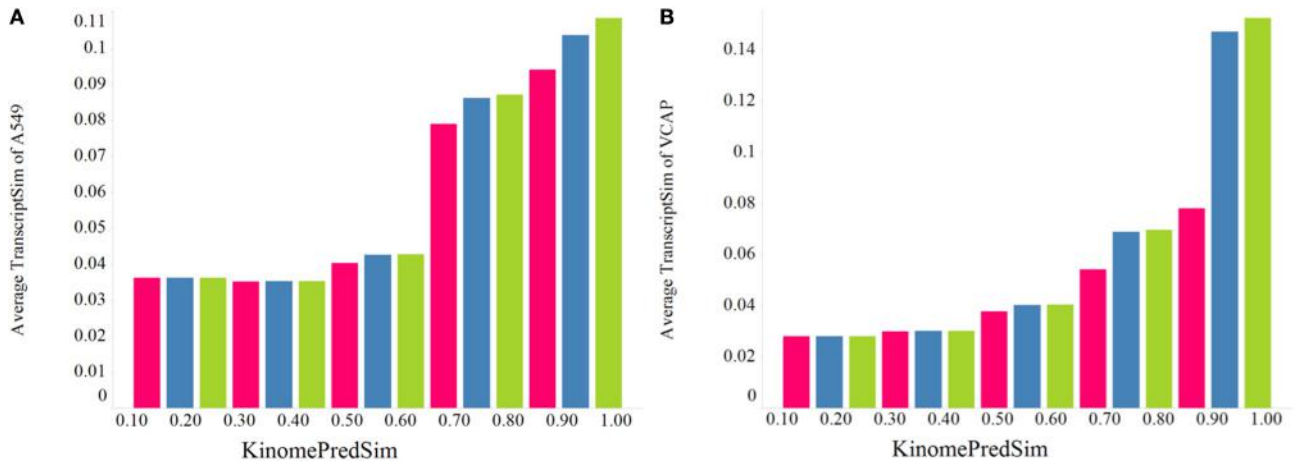


FIGURE 10 | Effect of the chemical similarity (ChemSim) of compound pairs on the trend of the average TranscriptSim as a function of KinomePredSim in (A) A549 cells and (B) VCAP cells. ChemSim cutoff applied are: 1.0 (green) including all compound pairs, 0.8 (blue) removing compound pairs more similar than 0.8, and 0.5 (red) leaving only dissimilar compound pairs (ChemSim < 0.5).

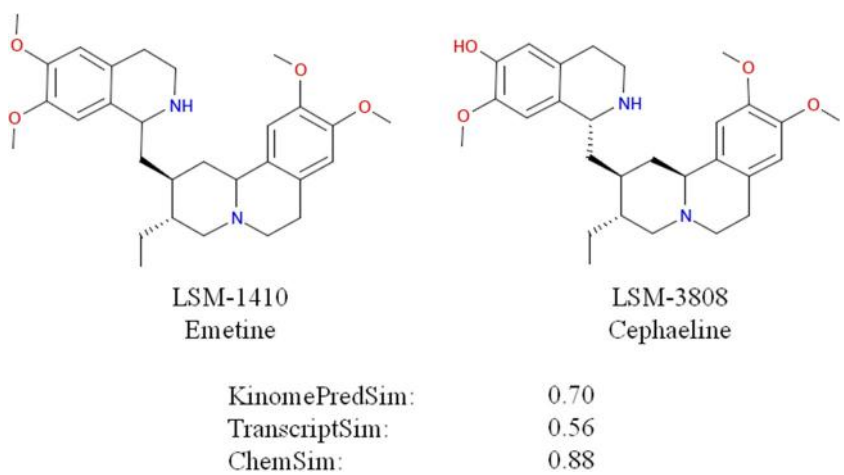


FIGURE 11 | Similar compounds (ChemSim of 0.88) with high KinomePredSim and TranscriptSim (based on L1000 in A549 cells).

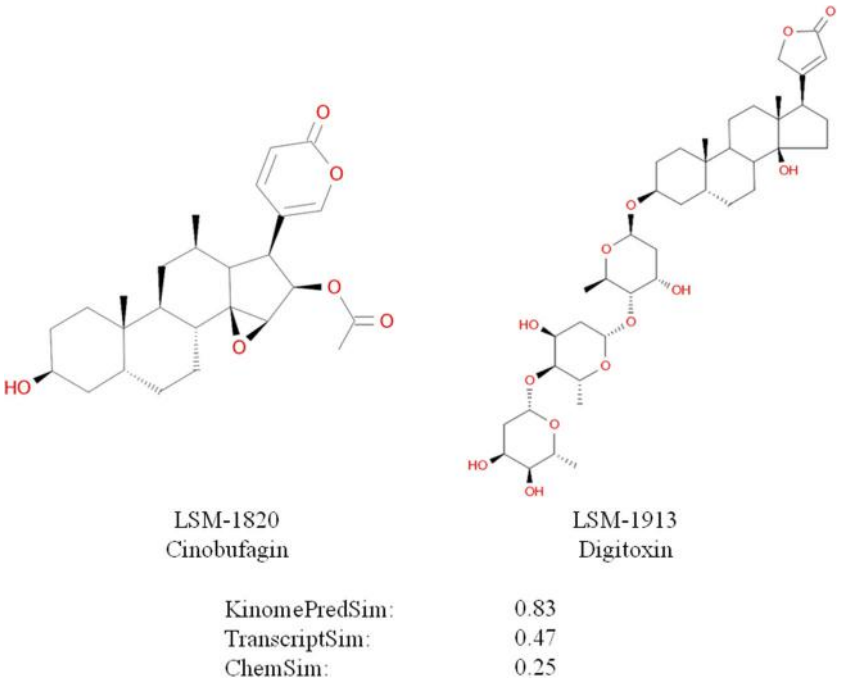


FIGURE 12 | Dissimilar compounds with high KinomePredSim and high TranscriptSim (based on L1000 in A549 cells).

responses are on average more similar then the L1000 responses of all LINCS compounds: for the A549 cell line, the global pairwise L1000 similarity average is 0.035 versus mTOR-pathway compounds' pairwise L1000 similarity average of 0.057; in VCAP cells these number are 0.028 versus 0.043, respectively. The corresponding Student test *p*-values are 1.38•10–28 and 1.1•10–14 for A549 and VCAP, respectively, providing a strong evidence that small molecule perturbagens that interfere with the same pathway (by inhibiting specific kinases in that pathway) result in significantly more similar transcriptional profiles compared to compounds active across different pathways.

Systematic pathway analysis

We also performed a more systematic study by using the NCI pathway database. We utilized the kinase inhibition models to predict the most likely pathway-active LINCS compounds in order to cover as much data as possible. We first annotated all kinase targets covered by our models by pathways (using a total of 224 NCI pathways). Once we had the kinase list for each pathway, we identified LINCS compounds that were predicted to be active for kinases in a given pathway, i.e., pathway-active compounds. Pairwise TranscriptSim values of these pathway-active compounds were compared to the TranscriptSim numbers of the remaining tested compounds and for each pathway the corresponding *p*-values were calculated. The requirement for the *p*-value calculation for each pathway was the presence of at least three pathway-active compounds, i.e., two similarities between them (necessary for the *t*-test mean distribution calculation). This reduced the number of pathways that could be investigated in formal statistics to 191. For the A549 cell line, 156 of 191 pathways have *p*-value below 0.05, suggesting that the greater transcriptional similarity is not random, while for the VCAP cell line we identified 162 of 191 pathways with *p*-value of less than 0.05. Kinases identified per pathway, as well as pathway-active compounds, can be found in the Supplementary Material Datasets 3 and 4 along with the corresponding *p*-values for cell line A549 and VCAP, respectively.

Even though our approach used a simplified assumption of pathway independence (we analyzed each pathway separately and not as a part of the network), it can be seen that transcriptional expression profiles originating from the same pathway (as defined by the participating kinases) are on average significantly more similar compared to result based on compounds that are not related to the same pathway. This is the case for majority of the pathways. For the pathways where this is not the case, we anticipate that additional information of pathway coexistence and dependence may be needed. However, our results provide strong indication that targeting a particular pathway will most likely lead to a certain transcriptional expression profile. And, importantly, it suggests that we can identify pathway-active compounds based on large-scale published data (KKB) or predict their activity via models based on these datasets.

KINASE SIGNATURES SUGGEST DIFFERENT CELL GROWTH INHIBITION PATHWAYS FOR A549 AND VCAP

After illustrating that transcriptional profiles are on average more similar when corresponding to the same cell line then when they are arising from two different cell lines (**Figure 8**), we were interested to contrast the enrichments of kinases for the two cell lines. We used the enrichment scores (as described in Materials and Methods and depicted in **Figure 4**) to identify kinases that are relevant for each cell line. Based on the experimental data we found that, for example, kinases PIK3CG,

NEK5, ERK3, NEK2, PIK3CA, PRKCE, CSNK2A2, PIM1, PKN2, and CAMK2D are enriched in non-small lung carcinoma A549 cell line while kinases DYRK1B, PCTK1, HIPK1, ICK, CDKL5, DYRK1A, MAK, ERK8, CLK1, and CLK2 are enriched in prostate carcinoma VCAP cell line. Mapping these kinases to pathways suggests that cell toxicity may be mediated by different pathways. For example, for VCAP enriched kinase MAK one pathway was identified from the NCI pathway collection: Co-regulation of Androgen receptor activity. In contrast, for A549 multiple pathways were related to the enriched kinases, but 7 pathways had more than one of these kinases as members: PDGFR-beta signaling pathway, CDC42 signaling events, Atypical NF-kappaB pathway, E-cadherin signaling in the nascent adherens junction, IL3-mediated signaling events, IL5-mediated signaling events, GMCSF-mediated signaling events, IL2-mediated signaling events, Role of Calcineurin-dependent NFAT signaling in lymphocytes, RhoA signaling pathway, IL8- and CXCR1-mediated signaling events, CXCR4-mediated signaling events, Class I PI3K signaling events, Thromboxane A2 receptor signaling pathway. These results illustrate the different (systems-wide) characteristics of the two cell lines and likely underlying mechanisms of action related to their growth inhibition. This is valuable for the development of selective and efficacious drugs based on prioritized and cell line-/disease-specific drug targets.

KINASE BINDING AND CELL VIABILITY PROFILES TO GUIDE DRUG REPURPOSING

In contrast to the example above where there appear to be no common kinase targets, repurposing of known drugs is now a common strategy to quickly identify approved drugs that can be applied to a new disease. Here we show an example of Crizotinib (LSM-1027), approved drug for some non-small cell lung carcinomas. Based on the LINCS KINOMEscan data one can identify kinases that are inhibited by this drug (INSR, AURKB, SRC, IGF1R, ROS1, MAP3K1, TYRO3, EPHB4, AXL, TXK, MET, FGR, FLT3, ALK). Furthermore we can identify the related pathways (NCI pathways described in Material and Methods). Although there are several pathways that may be implicated in multiple diseases, we can also identify specific ones, for example Glypican 1 (NCI Pathway ID 200026), which is associated through kinases SRC and FGR. This pathway is implicated in pancreatic cancer (Aikawa et al., 2008). Therefore by using approved non-small lung carcinoma drug Crizotinib, it may be possible to target SRC or FGR and therefore find its new uses in different cancer types.

DISCUSSION AND CONCLUSIONS

The LINCS project is a large-scale coordinated effort to generate a comprehensive systems biology reference resource of cellular and molecular response signatures for a wide range of cell lines, primary cells and stem cells, molecular, genetic, and other perturbations. The goals of the program include the generation of a very large multidimensional data matrix and informatics and computational tools to integrate, analyze, and make readily accessible such diverse data as genome-wide transcriptional profiles, biochemical protein binding, large-scale cellular phenotypic response signatures, and also proteomics and metabolomics data. To produce an integrative view of large and diverse datasets like those in the LINCS project, it is important to systematically standardize and annotate all data. Multiple efforts were carried out within our group and the LINCS consortium to define standards specifications and apply them to annotate a variety of perturbing or detected molecular entities cell model systems and other relevant concepts (Vempati et al., 2014). These efforts continue as the project moves into the next phase. Via tools developed in the program, for example the LIFE search engine (http://life.ccs.miami.edu), LINCS data can already be queried by standardized annotations across different sources.

Here we are particularly interested in small molecule perturbations, because of the potential of small molecules to be developed into therapeutic drugs and a general shift from purely target focused toward a systems poly-pharmacology based approach to drug development that could gain great insights from LINCS. To facilitate the cross-comparison of LINCS signatures, we established a fairly automated process for the standardization of small molecule compounds, which simplifies identification of compounds tested across several assays and also facilitates mapping and annotating of compounds using external sources such as DrugBank, the NCBI MLP probe reports, the NPC collection, and the Protein Data Bank (PDB). Unique compound IDs are also required to better coordinate data generation across centers; as illustrated in **Figure 3**, there are still gaps to be filled in order to achieve a complete data matrix across LINCS assays.

Nevertheless, important insights can be gained by bringing together the current datasets. For example we illustrated the integration of kinase binding profiles (KINOMEscan assay) and cell growth inhibition profiles. We combined these datasets using unique small molecules profiles across and used statistical enrichment to identify kinases that may play a role in the certain cell lines or diseases. The nature of the LINCS data matrix consisting of standardized response profiles enables the prioritization of sets of interesting kinases (signatures) that influence any of the tested cell lines. In that way kinases shared across many cell lines can be identified and such discovery may lead to new target identification or at least novel hypotheses. Also, by discovering common kinases between cell lines related to different diseases may lead to novel starting points for (cancer) drug repurposing.

We demonstrated that the similarity of compounds based on their chemical structure is related to their kinase binding profiles. This could be expected based on the similarity-property dogma, however is still noteworthy at a global level where each profile can represent a characteristic signature, implying that such signatures are related to chemical structures. Looking at the genome-wide transcriptional profiles for a much larger number of tested compounds at the Broad institute (see Materials and Methods), there was a similar trend that relates chemical similarity to global transcriptional similarity. It was more pronounced in the same cell line, but also detectable across cell lines. These chemical similarity trends can be interpreted as a generalization of the classical similarity-property principle, which underlies targeted lead optimization efforts. In particular in the case of transcriptional profiles, which have been related to disease phenotypes and models thereof (Lamb, 2007), these findings appear to support the feasibility of phenotypic lead optimization and utility of phenotypic structure-activity-relationships for drug development.

To link transcriptional responses to the underlying MoA, we compared the transcriptional profiles to the kinase binding profiles. Because of the quite small intersection of compounds for which L1000 and KINOMEscan profiles were available, we developed and applied kinase inhibition classification models based on a very large corpus of published data and applied these to all compounds tested in L1000. In addition to predicting activities for non-tested compounds and extending the current datasets to identify patterns in the data, these computational results can be also used to prioritize compounds for further experimental testing. For example the models could be used to identify a set of diverse compounds that are most likely to efficiently dissect the entire Kinome activity space or to prioritize compounds most likely to interfere in a given biological pathway, or any desirable poly-pharmacology profile to help deconvolute mechanisms of cellular responses.

As expected, the trend we observed for the experimental kinase binding profiles that chemically similar compounds are more likely to have similar kinase inhibition profiles, was also confirmed for the predicted kinase profiles just for all LINCS compounds as the modeling enabled it. We already knew that structurally very similar compounds were also more likely to have similar transcriptional profiles. However, their biochemical kinase similarity appeared related to transcriptional similarity independently from chemical similarity, at least to some extent. This would confirm a mechanistic relationship (by pathways), but more importantly a global response suggests a level of robustness in the cellular responses to chemical perturbation; i.e. small changes in biochemical binding do not have a huge effect on transcriptional response. This may be one reason why most drugs are well tolerated, despite (previously not known) poly-pharmacology and in some cases even alternate indications (drug repurposing). We anticipated that downstream gene expression signatures would be much more closely related by signaling pathways; i.e. compounds inhibiting kinases within a specific pathway should have more similar transcriptional profiles. We tested and confirmed this using actual data for the PI3K/AKT/mTOR pathway and using the kinase inhibition models for a large number of pathways from the NCI database. Although we applied a simplified approach of analyzing individual pathways, we observed that for the majority of pathways the transcriptional expression profiles resulting from small molecules that are active against any kinase in the same pathway are indeed more similar than transcriptional expression profiles of compounds that do not share activity against the same pathway.

Facilitated by common data standards and annotations we were able to integrate diverse biochemical, transcriptional, and phenotypic cell growth inhibition profiles for small molecule drug like molecules. After computing various similarity measures based on the response signatures and chemical information, we illustrated some insightful trends and elucidated the results at the systems-level. Our approach and findings to relate biochemical and transcriptional responses to chemical similarity as well as use of predictive models appear relevant to inform the development of novel poly-pharmacology drugs. We hope that some of the data integration and analysis presented here can inspire others in the research community to leverage LINCS data and the annotations we provided for their own studies and in novel ways.

ACKNOWLEDGMENT

This work was funded by the National Institutes of Health' LINCS project grants U01HL111561, and U01HL111561-02S1, and by the Center for Computational Science of the University of Miami.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00342/abstract

REFERENCES

Acencio, M. L., Bovolenta, L. A., Camilo, E., and Lemke, N. (2013). Prediction of oncogenic interactions and cancer-related signaling networks based on network topology. *PLoS ONE* 8:e77521. doi: 10.1371/journal.pone.0077521

Aikawa, T., Whipple, C. A., Lopez, M. E., Gunn, J., Young, A., Lander, A. D., et al. (2008). Glypican-1 modulates the angiogenic and metastatic potential of human and mouse cancer cells. *J. Clin. Invest.* 118, 89–99. doi: 10.1172/JCI32412

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* 10, 980. doi: 10.1038/nsb1203-980

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

Chen, J., Hu, Z., Phatak, M., Reichard, J., Freudenberg, J. M., Sivaganesan, S., et al. (2013a). Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules. *PLoS Comput. Biol.* 9:e1003198. doi: 10.1371/journal.pcbi.1003198

Dar, A. C., and Shokat, K. M. (2011). The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling. *Annu. Rev. Biochem.* 80, 769–795. doi: 10.1146/annurev-biochem-090308-173656

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 570–575. doi: 10.1038/nature11005

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., et al. (2013). Inconsistency in large pharmacogenomic studies. *Nature* 504, 389–393. doi: 10.1038/nature12831

Heiser, L. M., Sadanandam, A., Kuo, W. L., Benz, S. C., Goldstein, T. C., Ng, S., et al. (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2724–2729. doi: 10.1073/pnas.1018854108

Hong, D., Gupta, R., Ancliff, P., Atzberger, A., Brown, J., Soneji, S., et al. (2008). Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science* 319, 336–339. doi: 10.1126/science.1150648

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690. doi: 10.1038/nchembio.118

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33, D428–D432. doi: 10.1093/nar/gki072

Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* 7, 54–60. doi: 10.1038/nrc2044

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Laplante, M., and Sabatini, D. M. (2012). mTOR signaling in growth control and disease. *Cell* 149, 274–293. doi: 10.1016/j.cell.2012.03.017

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* 298, 1912–1934. doi: 10.1126/science.1075762

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358. doi: 10.1021/jm020155c

McDermott, U., Sharma, S. V., Dowell, L., Greninger, P., Montagut, C., Lamb, J., et al. (2007). Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19936–19941. doi: 10.1073/pnas.0707498104

Nagaraj, S. H., and Reverter, A. (2011). A Boolean-based systems biology approach to predict novel genes associated with cancer: application to colorectal cancer. *BMC Syst. Biol.* 5:35. doi: 10.1186/1752-0509-5-35

Peck, D., Crawford, E. D., Ross, K. N., Stegmaier, K., Golub, T. R., and Lamb, J. (2006). A method for high-throughput gene expression signature analysis. *Genome Biol.* 7, R61. doi: 10.1186/gb-2006-7-7-r61

Pipeline Pilot (2011). *Pipeline Pilot 8.0.* San Diego, CA: Accelrys Software Inc.

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Schurer, S. C., and Muskal, S. M. (2013). Kinome-wide activity modeling from diverse public high-quality data sets. *J. Chem. Inf. Model.* 53, 27–38. doi: 10.1021/ci300403k

Shao, H., Peng, T., Ji, Z., Su, J., and Zhou, X. (2013). Systematically studying kinase inhibitor induced signaling network signatures by integrating both therapeutic and side effects. *PLoS ONE* 8:e80832. doi: 10.1371/journal.pone.0080832

Silverman, E. K., and Loscalzo, J. (2012). Network medicine approaches to the genetics of complex diseases. *Discov. Med.* 14, 143–152.

TIBCO Spotfire. (2013). *TIBCO Spotfire*. Boston, MA: TIBCO.

Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* 8, R39. doi: 10.1186/gb-2007-8-3-r39

Vempati, U. D., Chung, C., Mader, C., Koleti, A., Datar, N., Vidovic, D., et al. (2014). Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS). *J. Biomol. Screen.* 19, 803–816. doi: 10.1177/1087057114522514

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi: 10.1093/nar/gks1111

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

Received: 31 July 2014; accepted: 12 September 2014; published online: 30 September 2014.

*Citation: Vidović D, Koleti A and Schürer SC (2014) Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. Front. Genet. **5**:342. doi: 10.3389/fgene.2014.00342*

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

Copyright © 2014 Vidović, Koleti and Schürer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.