Chapter

Integrative Systems Biology Resources and Approaches in Disease Analytics

Marco Fernandes and Holger Husi

Abstract

Currently, our analytical competences are struggling to keep-up the pace of in-deep analysis of all generated large-scale data resultant of high-throughput omics platforms. While, a substantial effort was spent on methods enhancement regarding technical aspects across many detection omics platforms, the development of integrative downstream approaches is still challenging. Systems biology has an immense applicability in the biomedical and pharmacological areas since the main goal of those focuses in the translation of measured outputs into potential markers of a Human ailment and/or to provide new compound leads for drug discovery. This approach would become more straightforward and realistic to use in standard analysis workflows if the collation of all available information of every component of a biological system was ensured into a single database framework, instead of search and fetch a single component at time across a scatter of databases resources. Here, we will describe several database resources, standalone and web-based tools applied in disease analytics workflows based in data-driven integration of outputs of multi-omic detection platforms.

Keywords: systems medicine, bioinformatics, omics, data integration, pathway analysis

1. Introduction

Over the last decade the emergence of high-throughput screening platforms and the increase in availability of large-scale-omics data, as well as clinical data from electronic health records comprising phenotypic, therapeutic and environmental factors information opened the possibility to mechanistically understand diseases and diseases stages at the molecular level. Thereby, a great number of wealth data in many kidney and cardiovascular conditions was generated, however these findings were neither translated nor reached the clinical setting and are still enclosed in peer-reviewed literature and across general scope expression profiling databases. Simultaneously it has become apparent that the existing systems to integrate and correlate this data are either inadequate or non-existent. Due to the multi-factorial molecular phenotype of disease, it is evident that development of novel therapeutic and disease detection approaches should be based upon the study of the entire "System" simultaneously. **Figure 1** gives a general overview in the fundamental difference between conventional and systems approaches, whereby in the context of conventional approaches a hypothesis is put forward that is assumed to be of importance in the disease or biological condition. This hypothesis is then tested

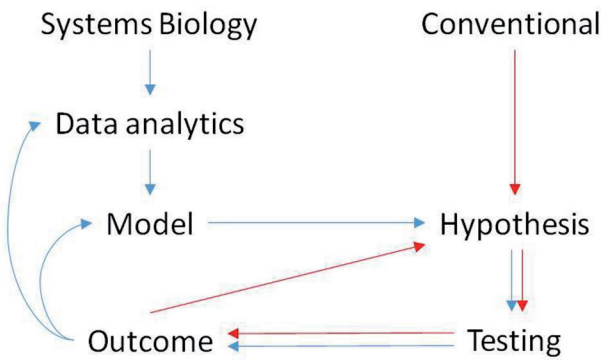

Figure 1.
Overview of general differences between conventional and Systems Biology approaches in biological and disease analysis research. Red arrows show the path of the conventional hypothesis-driven methodology including testing of a hypothesis, usually employing lab-based investigations, and re-adjusting the hypothesis dependent on outcomes. Blue arrows denote a systems approach, where data are integrated and analysed, producing a model system and a hypothesis that can be verified using conventional methods. Outputs of such an approach are usually fed back into the model or the data analysis stream to refine models, adjust hypothesis or confirm the established model.

and either validated or refuted based on the outcome of this hypothesis-driven methodology. Yet, it is obvious that it is easy to investigate any hypothesis and then choose the one that appears most correct, in the real world constraints such as time and financial resources do not allow for such an approach, and hypotheses are usually generated on a best-guess basis which can lead to a substantial amount of bias, resulting in skewed or partial insights and can often be misleading. In order to avoid such scenarios, research driven by the data itself rather than a hypothesis has been proposed a long time ago, but could not be properly implemented due to the lack of unbiased large-scale data or the ability to integrate disparate data in the first place. Additionally, a successful systems approach requires underlying prior knowledge, such as physicochemical parameters in how molecules interact with each other, what reactions they are involved in and other unconnected information. This knowledge has only slowly been accumulated through conventional research and has only over the last 10–15 years been available to such an extent where a systems approach became feasible. Data-driven systems biology-based diagnostic and prognostic models consisting of relevant panels of molecules—key branches of the cellular network, appear to more accurately reflect pathophysiology than traditional hypothesis-driven approaches, consequently, may have a much higher chance of success and implementation in the clinical setting. Of the most pronounced effects is the crossing between research borders and the urge for multidisciplinary integration of biology, chemistry, computing sciences, mathematics, and medicine to tackle the complexity of such system. To get a holistic view of a system's biology, multiple and different types of observations must be combined, such as clinical which includes pathological, demographical, epidemiological, and as well as molecular, which includes large-scale genotyping, gene expression, proteomics, metabolomics, and lipidomics data. The downside of such an approach in disease analytics or data integration is the rise in complexity both in output as well as in methods needed to generate those, and the skills required to interpret and contextualise outcome parameters. However, biological and disease models generated this way allow for a higher confidence in generating testable hypotheses, disease classifications on a molecular level and identification of overlapping and divergent pathways of malignant conditions. Ultimately, the removal of bias and integration of all available data, both clinical and biological, leads to a far better understanding of disease and enables the identification of intervention points with higher confidence and accuracy.

2. Disease classification boundaries

The standard resource for disease taxonomy relies primarily on the International Classification of Diseases (ICD) which displays information on diseases and health conditions, and a continuous monitoring of the associated epidemiological statistical trends World Health Organisation [1]. The foundations of the ICD disease classification relies mainly in a type of evidence-based medicine with distinction of clinical features, including patient symptoms, histological assessment, and evaluation of risk factors [2]. While widely used in the clinical setting, in the era of "big-data" and precision medicine, its rigid hierarchical structure lacks the flexibility needed to accommodate the fast and expanding molecular-insights of disease-phenotypes captured across many -omics platforms [3]. Moreover, to support this notion of undefined disease boundaries across current disease classification, we can observe the existence of co-occurring conditions that if seen as a unified biological network, could provide information about common multi-functional genes, cellular pathways, as well the impact of lifestyle [4]. Additionally, analysis of disease progression with the presence of overlapping conditions through evaluation of temporal correlation and disease progression patterns condensed from a population can become useful in the prediction and prevention at the patient's individual level in future disease-associate events [5].

Further disease taxonomy refinement can be achieved by applying network analysis [3] of combined disease phenotypes sourced from ICD-9 with protein-protein interactions (PPI's) data from STRING [6] and additional curation efforts of gene-disease associations (GDA) from several data sources. The network analysis allowed for reclassified of pancreatic cancer into 11 subclasses, which is consistent with the number of molecular subtypes observed in the Bailey et al. [7] study. They also proposed the use of such approach in drug repurposing, for instance therapy with metformin, a well-known agent used to treat type 2 diabetes mellitus (T2DM), that could regulate the imbalanced status of the microbiota community in the gut mucosae, a known cause of pathological chronic bowel inflammation as occurs in Crohn's disease and ulcerative colitis [8], and also act as preventable agent to reduce the risk of colorectal cancer. Moreover, molecular profiling associated with histologic assessment seems to yield enhanced probabilistic scores in graft survival predictions. For instance, joint integration of multi-center histology features in renal biopsies and gene-array data yielded a new molecular score system able to predict renal graft survival [9] and improving the diagnosis of antibody-mediated rejection of transplanted in hearts [10]. Such approaches can also be implemented to assess disease trajectory, treatment selection and monitoring in many neoplasms, and could be specially tailored for cases where the tumour primary site is of unknown origin [11].

3. Systems biology towards systems medicine

Over the last 15 years, the rise of systems biology as a research field has changed how we look at human normal physiological function and has helped to uncover disease complexity. Now scientists use systems biology approaches to understand the big picture of how all the pieces interact in an organism. The inference of genotype-phenotype relationships boosted by the assembly of a high-quality human genome opened the avenue for the development of reference maps of interactome networks, [12] consisting of binary association pairs, for instance PPI's, protein-DNA/RNA, or protein-metabolite interactions. **Figure 2** shows the essential biological molecular interactions governing cell behaviour in an over-simplified

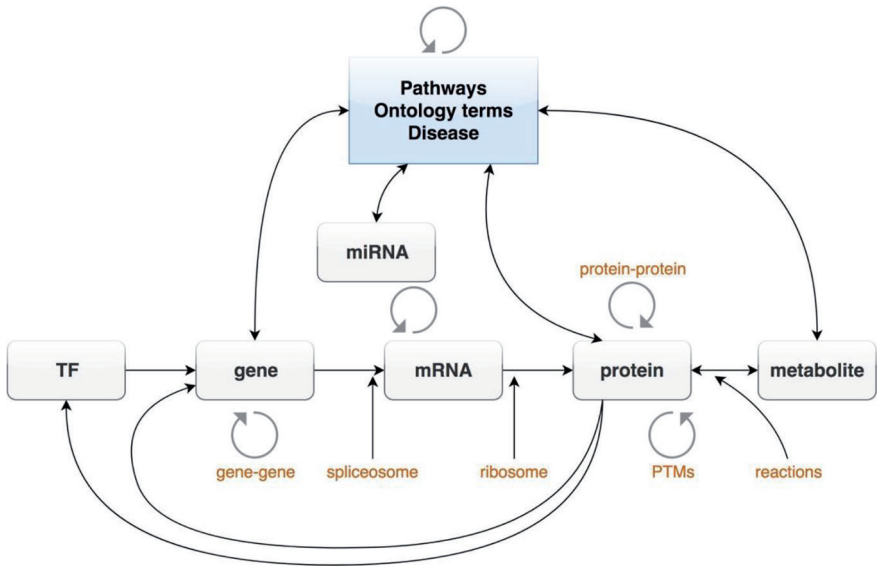

Figure 2.
Description of the essential known relationships/interactions in an over-simplified biological system. Transcription factor (TF), microRNA (miRNA), post-translational modifications (PTMs). The illustration does not account for epigenetic modifications, for instance DNA methylation and histone modifications known to occur and regulate gene expression. Dark coloured arrows denote entity associations, while self-circular arrows describe self-pair interactions or modifications.

biological system. A curated compilation of high-quality sources of binary interactions is considered a prime resource in the Systems Biology field and thereby enabling a deeper understanding of the larger picture—be it at the level of the organism, organ, tissue, or cell—by putting its components together. It's in stark contrast to decades of reductionist biology, which merely focuses on the properties of its individual components [13]. Most disease conditions exhibit expression of complex disease phenotypes [13], such as obesity, metabolic syndrome, autoimmune diseases and renal diseases.

Using the words of Ronald Germain to provide a definition of Systems Biology, he advocates that: "There are an endless number of definitions, it's even worse than the elephant," that infamous elephant that stymies the attempts of blind men to describe it because each feels just one part, "Some people think of it as bioinformatics, taking an enormous amount of information and processing it." "The other school of thought thinks of it as computational biology, computing on how the systems work. You need both parts." Ironically, to best understand this novel approach, we should take a reductionist approach to defining its parts. The system, it seems, is more than the sum of its parts [14]. Systems Biology requires comprehensive data at all molecular levels, a profound understanding of biological systems, data-criteria based assessment and in-deep understanding of the limitations of the techniques used in the experimental setup. Moreover, systems biology requires prior knowledge either published or sourced from biological databases and newly predicted and frequent molecular events requires further *in vivo*/*vitro* validation [15]. Systems Biology is cross-disciplinary: "[...] a scientific approach that combines the principles of engineering, mathematics, physics, and computer science with extensive experimental data to develop a quantitative as well as a deep conceptual understanding of biological phenomena, permitting prediction and accurate simulation of complex (emergent) biological behaviours" (Ronald Germain in [14]). Furthermore, systems biology promotes understanding of the functional roles and interplays of all molecules in cells in health and disease. Also provides a framework for large-scale

data-driven analysis and predictions based on prior knowledge of experimentally identified interactions and pathways [16]. Thus, more relevant that the underlying high-throughput screening methods, including genomics, proteomics, metabolomics, and also bioinformatics approaches is the use of such methods in a integrative manner to holistically understand how nonlinear processes and their outcomes are regulated in a biological system [17].

3.1 Bridging the gap between fields

Over the last 10 years, major efforts to reclassify diseases based on molecular insights from advances in molecular biology, bioinformatics and high-throughput screening yielded novel disease subtypes among many disease conditions. The use of multiple data types, including clinical endpoints—omics and ontology-based data have been used to reconstitute disease phenotypes, classify and to refine disease-relationships [18]. Nevertheless, the development of a molecular-based disease taxonomy that links global molecular networks with pathological phenotype landscapes remains elusive. Systems medicine can be perceived as a multi-disciplinary collaborative effort driven by the application of systems biology approaches, which includes methodological workflows from high-throughput-omics technologies to generate data, warehousing management systems for data flow and handling and methods for data analytics and interpretation in the context of biomedical research [19]. Ultimately, with further adoption of a systems-based approach patients will benefit of a measurable improvement of their health status since processes of disease onset and progression will be mechanistically identified, leading to new insights regarding disease-disease boundaries, and disease subtyping which facilitates ideal pharmacological interventions as drug repurposing [20]. For instance, the identification of digoxin, a drug used as therapy for atrial fibrillation and congestive heart failure [21] as potential drug candidate for pharmacological intervention in medulloblastoma subtypes 3 and 4 [22]. The authors of the study implemented an integrative systems biology approach using genomic data and collating existing drug-drug, drug-targets interactions information into a tridimensional functional-drug network. This approach involved handling omic data sets such as DNA-seq—mutated genes, copy-number variation (CNV)—repeated sections of the genome, RNA-seq and methylation profiles, combined with clinical measurements of patient outcomes (survival data) and fused using network-based and probabilistic methods that yielded a network composite with disrupted driver signalling networks and potential drug candidates [22].

4. Large-scale data: omics platforms

The advent of new high-throughput technologies (sequencing, array-based and mass spectrometry) led to an explosion of available data, not only by the number of experiments performed, but also by the data density obtained per experiment. Here, we will provide description of detection platforms handling molecular datasets; for medical imaging data types and analysis strategies please see the following review [23].

4.1 DNA microarrays and next-generation sequencing (NGS)

Microarray technologies have been widely used in research for primary screening, including gene expression profiling and providing genotype-phenotype relationship. Moreover, if properly designed, microarrays will not only provide information on gene expression and expressed single nucleotide polymorphisms (SNPs), but

also detect exon junctions and fusion genes [24]. However, identical to PCR-based techniques, the design of probes requires prior knowledge. Therefore, microarrays are mostly applied in the quantification of known sequences and not for the discovery of new variants, transcripts or other unknown features [25]. Microarrays have numerous limitations. For instance, they render an indirect measurement of the relative concentration of a particular nucleic acid sequence [26]. Another limitation is based that a DNA-array can only detect sequences that the array was designed for. In addition, non-coding RNA's that are not yet recognised as expressed are typically not represented on an array [26]. Microarrays are still considered a reliable technique for routine and/or initial screening that allows multiplex quantitation of microRNAs and gene probes expression in a fast, simple and affordable way. Nevertheless, the continuous drop in the cost of NGS at a level that virtually matches the cost of DNA microarray-based platforms, thus is foreseen that DNA-arrays will be fully replaced by sequencing methods within the next decade [26].

4.2 Proteomics

The use of omics technologies, including quantitative proteomics methods aims to identify and quantify the dynamics of protein abundance, in order to gain a deeper understanding of the associated biological functions. Thereby, the quantification of the expression level and state of all proteins at a given time can characterise physiological-states at the cellular-level [27]. Mass spectrometry (MS) technology, particularly tandem mass spectrometry (MS/MS), has been utilised as a discovery engine in proteomics [28]. This technology allows for identification and simultaneously quantification of hundreds or even thousands of proteins in an experimental setup, which enables real-time comparisons for instance between two or more physiological states [29]. Furthermore, peptide sequence composition will directly impact on ionisation efficiency, and their intensities observed in a spectrum often do not reflect their abundances, [30] thereby many label-free or label-based quantitation methods have arisen to allow comparative proteomic analysis. For instance, label-free proteomic approaches such as ion intensity, spectral counting have a simplified workflow when compared to labelling techniques; have no theoretical limit concerning multiplexing capability providing an improved proteome coverage, but lower quantification accuracy when compared with labelling methods (e.g. iTRAQ: isobaric tags for relative and absolute quantitation, SILAC: stable isotope labelling by/with amino acids in cell culture) [30]. In proteomics, several algorithms have been developed to query and cross compare MS data. The most popular used to identify proteins from raw MS data are for instance, MASCOT, SeQuest, OMSSA, X!Tandem [31], Andromeda [32], MS-GF [33], Paragon [34] and more recently, Morpheus [35] and an improved SEQUEST-like algorithm—ProLuCID [36]. The rise in the number of algorithms and specialised computational tools for analysis of MS-based proteomics data sets led to the development of workflows/pipelines such as PEAKS [37], MaxQuant [38], OpenMS Proteomics Pipeline (TOPP) [39], Trans-Proteomic Pipeline (TPP) [40] and others for further downstream data analysis—Perseus [41].

4.3 Metabolomics

In many metabolomics studies the identification and quantification of metabolites mainly rely on the application of analytical methods based on mass spectrometry (MS) (either coupled with a liquid or gas-chromatograph) and nuclear magnetic resonance (NMR) spectroscopy [42]. Metabolites are defined as small molecules, usually less than 1000 Da, which suffer several changes during cellular metabolism [43]. The selection of a particular platform depends upon the aims of

the experimental study and is typically driven by establishing a compromise among sensitivity, specificity, and scanning speed [44]. Metabolomics approaches can be globally split either by the full range measurement/analysis of all compounds in a given sample—untargeted metabolomics, or targeted metabolomics, in which a set of predefined and biochemically well-characterised compounds are measured in a sample [44]. MS has become an essential method for non-targeted profiling of metabolites in complex bio-samples, particularly low-abundance metabolites, due to its high sensitivity and selectivity capabilities when using liquid chromatography (LC) coupled to tandem MS/MS [45]. Metabolomics data from NMR and MS platforms are complex because they usually contain thousands distinct peaks therefore, multivariate statistical analysis plays an important role in metabolomics for reducing data dimensions, differentiating similar spectra, and in the development of predictive models [46]. Metabolomics is used as a screening tool in current healthcare settings, and could be greatly utilised to monitor therapy efficacy, and assess potential drug side-effects [47].

5. Data-driven approaches and multi-omics data integration

In the field of biomedical research adopting an unbiased approach or "hypothesis-free" (depending of the author and field of study, also defined as hypothesis-generating approach, data-driven research, or discovery research) to research can bring several benefits when compared with the widely used scientific approach—hypothesis-driven research (traditional approach). In which, the latter, in some cases encourages poor scientific practices by forcing/imposing qualitative and weak hypotheses that Are not prepared for strong statistical inference or quantitative analysis (QA) modelling, thereby in such cases an explicitly exploratory approach should be set as default [48]. In order to overcome this problem, large-scale approaches such as expression profiling started to become very popular in the mid '90s, and beginning of 2000, with the advent, rapid development and availability of high-throughput mass spectrometry, other methods followed [49]. Computational methods to analyse this flood of data were developed accordingly, however the majority only focused on one specific technology or experimental setup and up to this day are very often not interchangeable in other technological platforms. Large-scale approaches employed in omics research need a different analysis methodology, which is especially true if integrative analysis techniques are employed. True integrative (as opposed to integrating linear relationship data such as gene-protein data) approaches go beyond simple data fusion and gave rise to the field of Systems Biology. On the other hand, hypothesis-generating research (systems biology-derived hypotheses) and hypothesis-driven research are complementary, thus combining both approaches will certainly sustain more chances of a complete understanding of complex biological systems, than either approach on its own [48]. With the advent of high-throughput technologies their application in the biomedical field was a foreseen logical step. However, until recently integration of multi-omic data was not a common approach in former analysis workflows. The literature and publicly available databases are awash with data, yet the main approach of integrating all this information in a disease-specific context is traditionally based on meta-analysis at best or cannot be accomplished using standard computational methods. This molecular information can then be integrated in a further stage by means of meta-analysis or by cross-normalisation of data from different acquisition platforms [50]. A combinatorial stepwise data integration (**Figure 3**) approach can be used in order to incorporate data from different biological layers of information to predict phenotypic outcomes [51]. On the other side, by recreating the cell

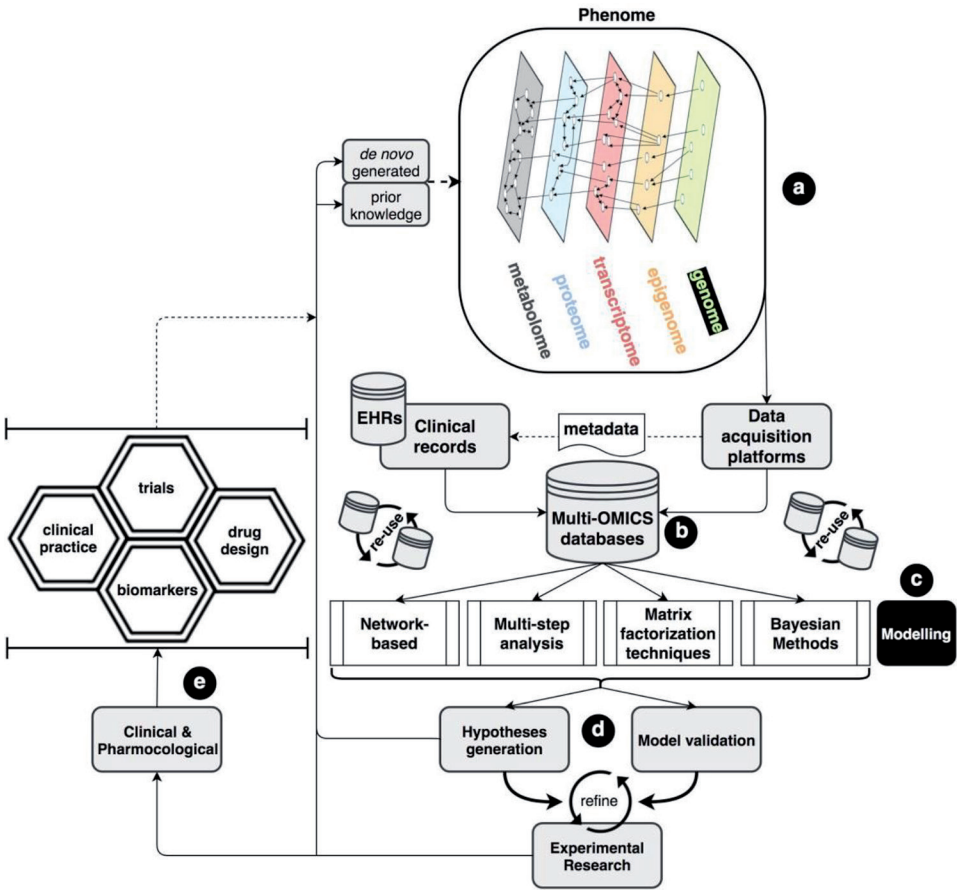

Figure 3.
Purposed workflow for a data-driven approach. Data generation from omics platforms plus existing biological information (a), development of a multi-omics database (b), selection of suitable modelling methods (c), model validation and use for hypothesis-generating research (d), lead optimization and candidate selection (e).

environment and dynamics by describing their interactions on a qualitative and quantitative manner and relying on underlying data (prior biological knowledge) for connectivity, e.g. PPI's, molecular co-occurrence, ontologies and enzymatic reactions [52]. Large-scale data sets for instance derived from multi-omics platforms may also be used to infer novel relationships by network learning approaches using Bayesian inference models [51] and extracting molecular information from multi-layered networks. This approach (as in many others) is challenging since it requires enough statistical power, higher number of samples to deduce all the possible interactions. Another challenge is due to the lack of uniformisation regarding the 'gold 'standards (criterions for evaluation) for accepting or rejecting relationships of the inferred model; however the ability to recreate a well-accepted interaction can at least be used for benchmarking methods in biological systems [53].

6. Biological databases and database systems

Databases form the basis for most applications in bioinformatics. The number of biological databases available now is enormous, the journal of Nucleic Acids Research (NAR) catalogues a total of 1737 molecular biology databases (2018 edition) [54]. The 2018 edition contains an enormous set of 181 papers that describe the adding of 82

new biological databases, 84 updates and as well 15 databases published elsewhere. However, a prominent issue concerns that many databases are not maintained over time and abandoned, yet they persist in database listings. There are many different types of databases, ranging from primary databases containing sequence data such as nucleic acid or protein; secondary databases or also known as pattern databases hosts, that results from the analysis of the sequences held in primary databases.

6.1 General scope expression databases

The Gene Expression Omnibus (GEO) [55] is a public repository that functions as both warehouse of raw microarray and other gene-based high-throughput data, and additionally serves as a platform for gene differential expression (DE) analysis using the GEO2R tool across a multitude of experimental conditions of user-submitted pre-processed data sets. In the same way, the European counterpart for storing of high-throughput genomics exists such as the European Bioinformatics Institute (EMBL-EBI) throughout the ArrayExpress database [56]. These data resources are both in compliance with community guidelines for description of an experimental setup for microarray and high-throughput NGS experiment. Comparatively, there is currently much less support for sharing of proteomics and metabolomics data sets despite the increasing demand. Public efforts for proteomic data sharing yielded the Proteomics Identification Database (PRIDE) that contains over 10,100 user-submitted MS-based raw proteomic data sets (September 2018) [57]. PeptideAtlas [58] handles re-analysed data sets via the TPP pipeline to provide end-users a consistently view over their data. MetaboLights [59] hosts user-submitted metabolomics experiments, which currently houses 439 experiments (November 2018). The standards for reporting proteomics and metabolomics experiments are coordinated by the Human Proteome Organisation's Proteomics Standards Initiative (HUPO-PSI), and Metabolomics Standards Initiative (MSI) respectively.

6.2 Disease profiling databases

Our group developed more specialised databases resources in several disease conditions handling pre-selected data sets containing DE molecules. In nephrology, we developed the Chronic Kidney Disease database (CKDdb) [60] storing microRNA, genomics, peptidomics, proteomics and metabolomics information relevant to CKD, collected from over 300 studies in the literature and integrated into the Pan-omics Analysis DataBase (PADB). The PADB framework (www.padb.org) uses gene and protein clusters (CluSO) and mapping of orthologous genes (OMAP) between species therefore facilitating data harmonisation from a diverse range of omics platforms and across several species, which makes it an invaluable resource for systems biology data-driven approaches. Also, many conditions associated with the cardiovascular system are covered in the Cardio/Vascular Disease (C/VD) database [61], which gives special emphasis on coronary artery disease (CAD). In neurological associated conditions such as Multiple Sclerosis we also developed the MuScle database [62] that stores and integrates curated data sets mined from large-scale studies with focus on genomics and miRNA. Likewise, we built a cancer-related differential expression database: the Multi-Omics Cancer database (MoCadb) that integrates clustered molecular information covering multi-omics studies in many gastro-intestinal cancers. In the same framework we also cover an assorted disease profiling database valuable for subtractive disease analysis studies, the Large-Scale Screening Resource (LSSR) that contains 81,980 entries, referring to 13,589 molecules. Moreover, a peak profiling database for biomarker patterns research, the Urinary Peptidomics and Peak-maps (UPdb) [63] database that

comprises Human urinary fingerprints from 200 subjects analysed mainly through surface enhanced laser desorption ionisation-time of flight mass spectrometry (SELDI-TOF-MS).

7. Software tools and solutions

Many modern high-throughput technologies lead to the generation of exceptionally large-scale and complex datasets, which includes PPI's, protein-DNA interactions, kinase-substrate interactions, qualitative and quantitative genetic-interactions gene co-expression [64]. The "Big Data" challenge can be fulfilled by the development of Bioinformatics tools to handle these large-datasets to reduce their complexity to a level that enables rationale interpretation and in this way is more likely to provide new biological insights to the Life Sciences. The compilation (not an exhaustive list) of many web-based, standalone tools and R-based packages are described in **Table 1**. They allow the accomplishment of different-omics tasks

Name	Description	Webpage	Ref.
iClusterPlus	Integrative clustering	bioconductor.org/packages/iClusterPlus	[84]
mixomics	Data integration (CCA,PLS,PCA)	mixomics.org	[85]
omicade4	MCIA and CIA	bioconductor.org/packages/omicade4	[86]
pwOmics	Pathway-based integration of omics	bioconductor.org/packages/pwOmics	[87]
PRESTO	Dimensionality reduction of multivariate data	github.com/saramcardle/PRESTO	[88]
caret	Classification and regression training	cran.r-project.org/web/packages/caret	—
GEO2R	Identify DE genes using GEOquery & limma R packages	ncbi.nlm.nih.gov/geo/geo2r	[55]
Metabo Analyst	Metabolomics analysis	metaboanalyst.ca	[89]
Networkanalyst/INMEX	Integration of gene DE via network approaches	networkanalyst.ca	[90]
ExAtlas	Meta-analysis & visualisation of gene DE	lgsun.irp.nia.nih.gov/exatlas	[91]
Elastic net	Gene DE with fitted GLM	https://zenodo.org/record/16006	[92]
ATHENA	Integration of genomics with clinical data	ritchielab.org/software/athena-downloads	[93]
Network propagation	Gene DE, mutations, PPI's	http://apps.cytoscape.org/apps/Diffusion	[94]

PMA, Penalised Multivariate Analysis; RGCCA, Regularised and Sparse Generalised Canonical Correlation Analysis for Multiblock Data; caret, Classification and REgression Training; ATHENA, Analysis Tool for Heritable and Environmental Network Associations; CCA, Canonical-Correlation Analysis; PLS, Partial Least Squares; PCA, Principal Component Analysis; CIA, Co-Inertia Analysis; MCIA, Multiple Co-Inertia Analysis; GO, gene ontology; DE, differential expression; GLM, generalised linear models.

Table 1.
Web-based, standalone tools and R packages dedicated to different-omics tasks such as feature selection, sample classification, multivariate approaches in data integration and meta-analysis.

such as feature selection, sample classification, multivariate methods. Cytoscape [65] is a tool primarily designed for network visualisation and analysis and has useful plugins available through the hosting website. Cytoscape makes use of a wide wealth variety of plugins to extend its functionality which are designed by the scientific community. The platform counts with several freely available apps/plugins (over 300 apps available on November 2018) for a diverse array of uses and analysis types.

7.1 Gene ontology (GO) and pathway-term-enrichment

The Gene Ontology (GO) consortium [66] aims to capture the increasing knowledge on gene function in a controlled vocabulary applicable to a wide range of organisms. GO represents genes and gene products attributes on matters of their associated biological processes (BP), cellular components (CC) and molecular functions (MF). GO is considered roughly hierarchical, with 'child' elements (terms) being more specific than their 'parent' elements (terms), nevertheless, a 'child' element (term) might have more than one parent element. The ClueGO app [67] is used for the integration and visualisation of GO and pathway terms sourced from KEGG [68], WikiPathways [69] and Reactome [70]. The resultant ClueGO network is established based in kappa statistics which shows the agreement on how any given gene and/or gene products pairs share similar terms. The ClueGO analysis output is conditioned by thresholding of the kappa coefficient, in which a higher coefficient conducts only to the visualisation of close-related terms with very identical gene products. While, lower kappa coefficients will let visualisation of less associated terms.

7.2 Gene-disease associations (GDA)

The conclusion of the Human Genome Project led to the massification of research related with uncovering genotype—disease phenotype associations [71]. This event translated in a disparate growth in the number of publications and on the other side a limited and slow paced biocuration of these newly discovered evidences. Currently, DisGeNET [72] unifies biomedical literature evidence based on GDA collated from a multitude of databases. This database makes use of the Medical Subjects Headings (MeSH) tree structure for disease classification by a Unified Medical Language System. The potential of the database is extended by disgenet2r package and optional programmatic access.

7.3 Protein-protein interactions (PPIs)

STRING database [6] collates molecular information to cover both known and predicted PPI's. All molecular interaction data is originally from primary interaction databases such as IntAct [73], BioGRID [74] and additional text-mining, coexpression and high-throughput experiments and computationally predicted PPIs. The up-to-date database version 10.5 comprises nearly 26 million PPI with a confidence score greater than 0.9 of more than 9 million proteins across 2031 organisms. GeneMANIA is another source for PPIs analysis and is accessible via web interface [75], and also as a Cytoscape app that can be used to detect related genes of a input query by means of a "guilt-by-association" strategy, which explores the realisation that a protein function can be obtained from another by seeing whether it interacts with another of known function. The app uses a large database of functional interaction networks, indexing 2152 association networks containing more than 500 million interactions mapped to 166,084 genes from nine organisms.

7.4 Combining metabolomic and gene expression data

Multi-omics datasets might not only contain protein and gene data, but also expression profiles of chemical compounds. While it is easy and straightforward to combine protein/DNA/RNA expression data using common identifiers, this is not the case for metabolism end-products—metabolites. This requires a guilt-by-association, which explores the rationale that metabolites are frequently produced by enzymes and a shift in metabolite expression can reflect an up-stream shift in protein or gene expression. This involves semantic searches in enzyme repositories—BRENDA to identify potential proteins and has some inherent pitfalls such as uncertainty which enzyme/isoform is responsible for the metabolic change. Additionally, the same compound could also be generated by several proteins, which adds to the uncertainty. Therefore, metabolic datasets are often treated as separate entities in multi-omics studies and analysed independently and then converged only at the level of final outcomes [76]. The MetScape 3 app [77] for the Cytoscape can perform joint analysis of both metabolomic and gene expression data and allows visualisation of the entire fused network, or by selecting custom views based on metabolic pathways When dealing with large-scale datasets, there is the option to use a concept file based on pre-computed gene set enrichment analysis (GSEA), along with statistical and fold-change thresholds.

7.5 Transcription factor (TF)-driven modules and microRNA-target regulation

Transcription factors (TF) are critical for the regulation of gene expression since they control if gene's DNA is transcribed into RNA [78]. A compendium on non-redundant TF and TF binding sites can be found at JASPAR [79]. The number of human TF ranges from 1500 to 2600, depending on source and stringency [78]. Direct analysis of modulated events due to TFs is not only valuable but might shed light on hidden elements that conventional pathway analysis cannot reveal. However, many TF binding sites and modulated genes are very hypothetical and often a random guess. Therefore, network-based analysis and interpretation involving TF elements should be taken with caution. CyTargetLinker [80] for extends existing biological networks by adding interactions associated with regulatory elements such as TF-target, miRNA-target or drug-targets. The application requires a loaded network with network attributes preferentially mapped to Ensembl, NCBI gene, UniProt, miRBase or DrugBank. Similarly, in CluePedia [81] users can perform miRNA analysis, by matching it to target-genes via selection of different database resources custom versions. Users can upload a list of genes and query the app to perform gene/miRNA enrichments. Then it will generate a miRNA-target interaction network that can be reused for inline integration with GO and pathway term clustering [81] within ClueGO.

7.6 Pathway mapping and visualisation

7.6.1 PathVisio pathway mapping and edition

PathVisio [82] allows drawing, edition, and visualisation of pathways handling gene, protein and metabolite data that can be further cross-mapped via the BridgeDb [83]. Inference of relevant pathways is based on an archive of pre-existent pathway maps from WikiPathways [69] and Reactome [70], establishing pathway over-representation based on a Z-score statistical procedure under the hypergeometric distribution and a *P*-value ranking based on a permutation procedure (randomisation test) that compares actual and permuted Z-scores. Pathways with a permuted $P < 0.05$ are considered significant by default.

7.6.2 KEGG pathway mapping

KEGG is an integrated database resource of biological systems integrating genomic, compound and functional information. KEGG allows analysis of datasets from high-throughput omics technologies by uploading a list of genes/proteins or metabolites along with optional statistical scores and fold-change values. After converting to KEGG internal identifiers, the molecular data is matched (KEGG mapper) into a collection of curated pathways, covering metabolism, signalling transduction pathways, specific pathways for several disease conditions and drug development.

8. Conclusions and future perspectives

The availability of large-scale multi-omics data has opened the avenue to gain an unrivalled insight in disease-associated molecular pathophysiological changes. Simultaneously it has become apparent that systems to integrate and correlate this data are either inadequate or non-existent. The literature and publicly available databases are awash with data, yet the main approach of integrating all this information in a disease-specific context is traditionally based on meta-analysis at best or cannot be accomplished using standard computational methods. In order to better model complex organisms, samples from multiple tissues of the same individuals should be studied simultaneously using omics data, which will require the development of novel analysis methods. Acquiring the relevant tissues and/or body fluid sources from Human study cohorts can of course be difficult, thereby comparative systems biology may help identify which organisms may be similar enough in each aspect to be used as models. It is sometimes suggested that omics technologies and systems biology have failed to deliver many breakthrough enhancements to the treatment of complex diseases. In some cases, it may be that in fact such diseases are not truly one disease from a system or reductionist point-of-view, but several with the same or similar phenotypic end-points—i.e., with the current terminology they are unknown subtypes of disease. If this is the case, then the overlap between the systems is poor and statistical methods which the approach relies on require very large cohorts for identification of these subtypes and subsequent description of each system. Other possibilities are that longitudinal data or samples from different tissues are required. Other relevant concerns arise from biomarker validation studies, such as correlated observations (i.e. multiple observations per patient), multiplicity (testing multiple biomarkers or endpoints), multiple clinical endpoints (interest in more than one relevant endpoint) and selection bias (from retrospective data or observational study). Data-driven investigations using systems biology approaches, although offer complete views over the function of biological systems in health and disease its limited by the state of completeness of prior biological information.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007–2013 under grant agreement FP7-PEOPLE-2013-ITN-608332. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Author details

Marco Fernandes[1] and Holger Husi[1,2*]

1 Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, United Kingdom

2 Division of Biomedical Sciences, Centre for Health Science, University of the Highlands and Islands, Inverness, United Kingdom

*Address all correspondence to: holger.husi@uhi.ac.uk

IntechOpen

References

[1] World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems: tenth revision. Geneva: World Health Organization. 2004

[2] National Research Council Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine. Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC: National Academies Press; 2011

[3] Zhou X, Lei L, Liu J, Halu A, Zhang Y, Li B, et al. A Systems Approach to Refine Disease Taxonomy by Integrating Phenotypic and Molecular Networks. eBioMedicine. 2018;**31**:79-91

[4] Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. Nature reviews Genetics. 2016;**17**(10):615-629

[5] Jensen AB, Moseley PL, Oprea TI, Ellesoe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Communications. 2014;**5**:4022

[6] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;**43**(Database issue):D447-D452

[7] Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;**531**(7592):47-52

[8] Lee SY, Lee SH, Yang EJ, Kim EK, Kim JK, Shin DY, et al. Metformin ameliorates inflammatory bowel disease by suppression of the STAT3 signaling pathway and regulation of the between Th17/Treg balance. PLoS One. 2015;**10**(9):e0135858

[9] Reeve J, Bohmig GA, Eskandary F, Einecke G, Lefaucheur C, Loupy A, et al. Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes. JCI Insight. 2017;**2**(12):pii:94197

[10] Parkes M, Reeve J, Kim D, Macdonald P, Aliabadi A, Goekler J, et al. Molecular assessment of heart transplant biopsies: Emergence of the injury dimension. Transplantation. 2018;**102**:S62-SS3

[11] Birner P, Prager G, Streubel B. Molecular pathology of cancer: How to communicate with disease. ESMO Open. 2016;**1**(5):e000085

[12] Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014;**159**(5):1212-1226

[13] Chuang HY, Hofree M, Ideker T. A decade of systems biology. Annual Review of Cell and Developmental Biology. 2010;**26**:721-744

[14] Germain RN. Systems-biology-as-defined-by-nih. Available from: https://irp.nih.gov/catalyst/v19i6/systems-biology-as-defined-by-nih

[15] Ashburn TT, Thor KB. Drug repositioning: Identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery. 2004;**3**(8):673-683

[16] Robinson SW, Fernandes M, Husi H. Current advances in systems and integrative biology. Computational and Structural Biotechnology Journal. 2014;**11**(18):35-46

[17] Gottschalk RA, Martins AJ, Sjoelund V, Angermann BR, Lin B, Germain RN. Recent progress using systems biology approaches to better understand molecular mechanisms of immunity. Seminars in Immunology. 2013;**25**(3):201-208

[18] Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science (New York, N.Y.). 2015;**347**(6224):1257601

[19] Apweiler R, Beissbarth T, Berthold MR, Bluthgen N, Burmeister Y, Dammann O, et al. Whither systems medicine? Experimental & Molecular Medicine. 2018;**50**(3):e453

[20] Ayers D, Day PJ. Systems medicine: The application of systems biology approaches for modern medical research and drug development. Molecular Biology International. 2015;**2015**:698169

[21] Virgadamo S, Charnigo R, Darrat Y, Morales G, Elayi CS. Digoxin: A systematic review in atrial fibrillation, congestive heart failure and post myocardial infarction. World Journal of Cardiology. 2015;**7**(11):808-816

[22] Huang L, Garrett Injac S, Cui K, Braun F, Lin Q, Du Y, et al. Systems biology-based drug repositioning identifies digoxin as a potential therapy for groups 3 and 4 medulloblastoma. Science Translational Medicine. 2018;**10**(464):pii:eaat0150

[23] Papp L, Spielvogel CP, Rausch I, Hacker M, Beyer T. Personalizing medicine through hybrid imaging and medical big data analysis. Frontiers in Physics. 2018;**6**:51

[24] Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Medical Science Monitor Basic Research. 2014;**20**:138-142

[25] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008;**5**(7):621-628

[26] Bumgarner R. Overview of DNA microarrays: Types, applications, and their future. Current Protocols in Molecular Biology. 2013; Chapter 22, Unit 22.1:1-11

[27] Cox J, Mann M. Is proteomics the new genomics? Cell. 2007;**130**(3):395-398

[28] Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annual Review of Biomedical Engineering. 2009;**11**:49-79

[29] Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: Status in 2013. Nucleic Acids Research. 2013;**41**(Database issue):D1063-D1069

[30] Lindemann C, Thomanek N, Hundt F, Lerari T, Meyer HE, Wolters D, et al. Strategies in relative and absolute quantitative mass spectrometry based proteomics. Biological Chemistry. 2017;**398**(5-6):687-699

[31] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database—2009 update. Nucleic Acids Research. 2009;**37**(Database issue):D767-D772

[32] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: A peptide search engine integrated into the MaxQuant environment. Journal of Proteome Research. 2011;**10**(4):1794-1805

[33] Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. Journal of Proteome Research. 2008;**7**(8):3354-3363

[34] Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, et al. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Molecular & Cellular Proteomics. 2007;**6**(9):1638-1655

[35] Wenger CD, Coon JJ. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. Journal of Proteome Research. 2013;**12**(3):1377-1386

[36] Xu T, Park SK, Venable JD, Wohlschlegel JA, Diedrich JK, Cociorva D, et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. Journal of Proteomics. 2015;**129**:16-24

[37] Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, et al. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry. 2003;**17**(20):2337-2342

[38] Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, et al. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nature Protocols. 2009;**4**(5):698-705

[39] Kohlbacher O, Reinert K, Gropl C, Lange E, Pfeifer N, Schulz-Trieglaff O, et al. TOPP--the OpenMS proteomics pipeline. Bioinformatics. 2007;**23**(2):e191-e197

[40] Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, et al. A guided tour of the trans-proteomic pipeline. Proteomics. 2010;**10**(6):1150-1159

[41] Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods. 2016;**13**(9):731-740

[42] Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chemical Society Reviews. 2011;**40**(1):387-426

[43] Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: The apogee of the omics trilogy. Nature Reviews Molecular Cell Biology. 2012;**13**(4):263-269

[44] Lei Z, Huhman DV, Sumner LW. Mass spectrometry strategies in metabolomics. The Journal of Biological Chemistry. 2011;**286**(29):25435-25442

[45] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrometry Reviews. 2007;**26**(1):51-78

[46] Blekherman G, Laubenbacher R, Cortes DF, Mendes P, Torti FM, Akman S, et al. Bioinformatics tools for cancer metabolomics. Metabolomics: Official journal of the Metabolomic Society. 2011;**7**(3):329-343

[47] Trivedi DK, Hollywood KA, Goodacre R. Metabolomics for the masses: The future of metabolomics in a personalized world. New Horizons in Translational Medicine. 2017;**3**(6):294-305

[48] Carroll S. Defining the scientific method: Editorial. Nature Methods. 2009;**6**(4):237

[49] Zhang Z, Wu S, Stenoien DL, Pasa-Tolic L. High-throughput proteomics.

Annual Review of Analytical Chemistry (Palo Alto, CA). 2014;**7**:427-454

[50] Walsh CJ, Hu P, Batt J, Santos CC. Microarray meta-analysis and cross-platform normalization: Integrative genomics for robust biomarker discovery. Microarrays (Basel). 2015;**4**(3):389-406

[51] Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics. 2017;**8**:84

[52] Gligorijevic V, Przulj N. Methods for biological data integration: Perspectives and challenges. Journal of The Royal Society Interface. 2015;**12**(112):pii:20150571

[53] Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nature Methods. 2012;**9**(8):796-804

[54] Rigden DJ, Fernandez XM. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Research. 2018;**46**(D1):D1-d7

[55] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets--update. Nucleic Acids Research. 2013;**41**(Database issue):D991-D995

[56] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—Simplifying data submissions. Nucleic Acids Research. 2015;**43**(Database issue):D1113-D1116

[57] Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Research. 2018;**47**:D442-D450

[58] Deutsch EW. The PeptideAtlas project. Methods in Molecular Biology. 2010;**604**:285-296

[59] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. Nucleic Acids Research. 2013;**41**(Database issue):D456-D463

[60] Fernandes M, Husi H. Establishment of a integrative multi-omics expression database CKDdb in the context of chronic kidney disease (CKD). Scientific Reports. 2017;**7**:40367

[61] Fernandes M, Patel A, Husi HC. VDdb: A multi-omics expression profiling database for a knowledge-driven approach in cardiovascular disease (CVD). PLoS One. 2018;**13**(11):e0207371

[62] Cervantes-Gracia K, Husi H. Integrative analysis of multiple sclerosis using a systems biology approach. Scientific Reports. 2018;**8**(1):5633

[63] Husi H, Barr JB, Skipworth RJ, Stephens NA, Greig CA, Wackerhage H, et al. The human urinary proteome fingerprint database UPdb. International Journal of Proteomics. 2013;**2013**:760208

[64] Mohr S, Bakal C, Perrimon N. Genomic screening with RNAi: Results and challenges. Annual Review of Biochemistry. 2010;**79**:37-64

[65] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research. 2003;**13**(11):2498-2504

[66] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The Gene

Ontology Consortium. Nature Genetics. 2000;**25**(1):25-29

[67] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;**25**(8):1091-1093

[68] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000;**28**(1):27-30

[69] Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, et al. WikiPathways: Capturing the full diversity of pathway knowledge. Nucleic Acids Research. 2016;**44**(D1):D488-D494

[70] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2016;**44**(D1):D481-D487

[71] Ozgur A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;**24**(13):i277-i285

[72] Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database: The Journal of Biological Databases and Curation. 2015;**2015**:bav028

[73] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research. 2014;**42**(Database issue):D358-D363

[74] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: A general repository for interaction datasets. Nucleic Acids Research. 2006;**34**(Database issue):D535-D539

[75] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research. 2010;**38**(Web Server issue):W214-W220

[76] van der Knaap JA, Verrijzer CP. Undercover: Gene control by metabolites and metabolic enzymes. Genes & Development. 2016;**30**(21):2345-2369

[77] Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics. 2012;**28**(3):373-380

[78] Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;**527**(7578):384-388

[79] Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2016;**44**(D1):D110-D115

[80] Kutmon M, Kelder T, Mandaviya P, Evelo CT, Coort SL. CyTargetLinker: A cytoscape app to integrate regulatory interactions in network analysis. PLoS One. 2013;**8**(12):e82160

[81] Bindea G, Galon J, Mlecnik B. CluePedia cytoscape plugin: Pathway insights using integrated experimental and in silico data. Bioinformatics. 2013;**29**(5):661-663

[82] van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, et al. Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics. 2008;**9**:399

[83] van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, et al. The BridgeDb framework: Standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics. 2010;**11**:5

[84] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;**25**(22): 2906-2912

[85] Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology. 2017;**13**(11):e1005752

[86] Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014;**15**:162

[87] Wachter A, Beissbarth T. pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge. Bioinformatics. 2015;**31**(18):3072-3074

[88] McArdle S, Buscher K, Ehinger E, Pramod AB, Riley N, Ley K. PRESTO a new tool for integrating large-scale-omics data and discovering disease-specific signatures. bioRxiv. 2018:302604

[89] Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Research. 2018;**46**(W1):W486-W494

[90] Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock RE. INMEX—A web-based tool for integrative meta-analysis of expression data. Nucleic Acids Research. 2013;**41**(Web Server issue):W63-W70

[91] Sharov AA, Schlessinger D, Ko MS. ExAtlas: An interactive online tool for meta-analysis of gene expression data. Journal of Bioinformatics and Computational Biology. 2015;**13**(6):1550019

[92] Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Research. 2015;**43**(12):e79

[93] Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData Mining. 2013;**6**(1):23

[94] Carlin DE, Demchak B, Pratt D, Sage E, Ideker T. Network propagation in the cytoscape cyberinfrastructure. PLoS Computational Biology. 2017;**13**(10):e1005598