REVIEW

Integrated omics: tools, advances and future approaches

Biswapriya B Misra[1], **Carl Langefeld**[1,2], **Michael Olivier**[1] and **Laura A Cox**[1,3]

[1]Center for Precision Medicine, Section on Molecular Medicine, Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, North California, USA
[2]Department of Biostatistics, Wake Forest School of Medicine, Winston-Salem, North California, USA
[3]Southwest National Primate Research Center, Texas Biomedical Research Institute, San Antonio, Texas, USA

Correspondence should be addressed to L A Cox: laurcox@wakehealth.edu

Abstract

With the rapid adoption of high-throughput omic approaches to analyze biological samples such as genomics, transcriptomics, proteomics and metabolomics, each analysis can generate tera- to peta-byte sized data files on a daily basis. These data file sizes, together with differences in nomenclature among these data types, make the integration of these multi-dimensional omics data into biologically meaningful context challenging. Variously named as integrated omics, multi-omics, poly-omics, trans-omics, pan-omics or shortened to just 'omics', the challenges include differences in data cleaning, normalization, biomolecule identification, data dimensionality reduction, biological contextualization, statistical validation, data storage and handling, sharing and data archiving. The ultimate goal is toward the holistic realization of a 'systems biology' understanding of the biological question. Commonly used approaches are currently limited by the 3 i's – integration, interpretation and insights. Post integration, these very large datasets aim to yield unprecedented views of cellular systems at exquisite resolution for transformative insights into processes, events and diseases through various computational and informatics frameworks. With the continued reduction in costs and processing time for sample analyses, and increasing types of omics datasets generated such as glycomics, lipidomics, microbiomics and phenomics, an increasing number of scientists in this interdisciplinary domain of bioinformatics face these challenges. We discuss recent approaches, existing tools and potential caveats in the integration of omics datasets for development of standardized analytical pipelines that could be adopted by the global omics research community.

Key Words

- integrated
- omics
- genomics
- transcriptomics
- proteomics
- metabolomics
- network
- statistics
- Bayesian
- machine learning
- principal component analysis
- correlation
- clustering

Journal of Molecular Endocrinology
(2019) **62**, R21–R45

Introduction

Access to large-scale omics datasets (genomics, transcriptomics, proteomics, metabolomics, metagenomics, phenomics, etc.) has revolutionized biology and led to the emergence of systems approaches to advance our understanding of biological processes. With decreasing time and cost to generate these datasets, omics data integration has created both exciting opportunities and immense challenges for biologists, computational biologists, biostatisticians and biomathematicians. As an example of a comprehensive analysis approach, Yugi *et al.* (2016) proposed a trans-omics concept of dynamic networks that includes the three most commonly used

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

layers of omics datasets – transcriptomics, proteomics and metabolomics and also included newer datasets such as phosphoproteomics, protein–protein interactions, DNA–protein interactions and allosteric regulation, which can reveal critical components of dynamic biological networks when omics data are successfully integrated. Using three case studies in datasets from bacteria and rats, they showed the interplay of the omics layers, and introduced phenome-wide association, pathway-wide association and trans-ome-wide association (Trans-OWAS) studies to connect phenotypes with omics networks that reflect genetic and environmental factors. These multi-layered, multifactorial approaches are computationally challenging and difficult to display and comprehend visually. Additional data from microRNA/gene, protein/protein, DNA/protein, and protein/RNA interactions further increase the complexity. A recent review enlists genome-based systems biology tools and applications available for network analysis, pathway construction, genome alignments, assemblies, tree viewers and phylogenies, microarray and RNA-Seq viewers, genome browsers, visualization tools for comparative genomics, and tools for building visual prototypes (Pavlopoulos *et al.* 2015). Similarly, tools, resources, databases and software for analysis and visualization of proteomics (Oveland *et al.* 2015) and metabolomics data (Misra & van der Hooft 2016, Misra *et al.* 2017, Misra 2018) are reviewed on a yearly basis. However, none of these recent publications provide a comprehensive overview of approaches for integrating three or more omics datasets.

Although the need for, and the importance of, integration of omics data has been realized for a broad range of research areas, including food and nutrition science (Kato *et al.* 2011), systems microbiology (Fondi & Liò 2015), analysis of microbiomes (Muller *et al.* 2014), genotype–phenotype interactions (Ritchie *et al.* 2015), systems biology (Mochida & Shinozaki 2011, Fukushima & Kusano 2013), natural product discovery (Yang *et al.* 2011) and disease biology (Pathak & Dave 2014), successful implementation of more than two omics datasets is very rare. Since Gehlenborg *et al.* (2010) produced a useful comprehensive compendium for visualization of omics data for systems biology using data from microarrays, RNA deep sequencing, mass spectrometry (MS), nuclear magnetic resonance (NMR) and protein interactions, considerable progress has been made to develop additional tools and approaches for integrated omics analysis. Broad experimental challenges in these integrated omics approaches include, but are not limited to (i) understanding the statistical behavior of readouts from each omics regime independently, (ii) recognizing non-obvious relationships that exist between omics regimes within their original biological context and (iii) capitalizing on time resolution in omics data, such as time course studies, to inform directionality (Buescher & Driggers 2016). A recent review provided data integration strategies for genomics and proteomics datasets (Huang *et al.* 2017), but did not mention and include approaches, which allow integration of metabolomics datasets.

Although all individual omics datasets might not have the four vs associated with integration of 'big data', i.e., volume, variety, velocity and veracity, they pose similar challenges, especially in studies with large sample numbers. In addition, for high-dimensional datasets of more than 1000 variables, popularly known as the 'curse of dimensionality', variances among samples become large and sparse and render cluster analysis uninformative (Ronan *et al.* 2016), further posing challenges interpreting integrated omic datasets. For clarity, we use 'integrated omics' to denote multi-omics approaches integrating three or more omics datasets and include the major omics data types, i.e., genomics, transcriptomics, proteomics and metabolomics.

Strengths and challenges of individual omics

Genomics and transcriptomics

Genomics and transcriptomics have been applied to various aspects of research and clinical applications ranging from the pharmaceutical industry, diagnostics and therapeutics, gene therapy applications, pharmacogenomics and disease prevention, to developmental biology, evolutionary genomics and comparative genomics. Thus, the ability to manage and analyze these types of data has become necessary for a biomedical scientist's skill set. The surge in advancements of next-generation sequencing (NGS) technologies and progress in genomic data analysis have led to high-throughput data generation for genomes (single nucleotide polymorphisms (SNPs), copy number variants (CNVs), loss of heterozygosity variants, genomic rearrangements, and rare variants), epigenomes (DNA methylation, histone modifications, chromatin accessibility, transcription factor (TF) binding) and transcriptomes (gene expression, alternative splicing, long non-coding RNAs and small RNAs such as microRNAs) (Ritchie *et al.* 2015). Generally speaking, the nucleic acid-based omics approaches for data generation rely on five major steps: appropriate sample collection, high-quality nucleic acid extraction, library preparation, clonal

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

amplification, and sequencing (e.g., pyrosequencing, sequencing-by ligation, or sequencing-by synthesis). The specific approach used for each step varies based on the intended downstream application. Following sequencing, the workflow includes data cleaning, filtering, assembly, alignment (*de novo* or reference-based), variant calling, annotation and functional predictions. In addition, pathway and/or network analyses are often used to provide biological context. Heterogeneous datasets pose challenges because quality assurance, quality control, data normalization and data reduction methods differ among the various types of individual datasets. For example, normalization and scale of RNA-Seq data differs from small RNA-Seq data, for example, RNA-Seq datasets typically include tens of thousands of transcripts, while small RNA-Seq datasets typically include less than 2000 small RNAs. With the rapid development of single-cell sequencing technologies, sequencing technologies that produce longer reads, and applications for genomic and transcriptomic analyses, additional challenges are emerging such as appropriate sequence coverage and statistical analysis of single-cell data (Menon 2017). A review of genomics applications and tools is provided in Shendure (2017). Best practices for DNA-seq pipelines are provided by NIH National Cancer Institute. Readers are further directed to Costa-Silva *et al.* (2017) for a comprehensive analysis of current transcriptomic analysis tools. Sequencing-based technologies, which are the most advanced of the omics technologies in terms of availability of laboratory reagents for standardized protocols, analytical tools and public databases for data sharing, provide unique opportunities to obtain high quality from small amounts of tissues or individual cells to address a wide range of biological questions.

Proteomics

Proteomics is used to quantify proteins in multiple sample types using both shotgun and targeted approaches. Recent developments in MS have dramatically increased sensitivity while decreasing the amount of sample required for high-throughput analyses and now allow for the detection of minimal differences in protein abundances, identification of post-translational modifications and other applications from a wide range of samples and tissues (Aebersold & Mann 2016). Whether choosing a chemically labeled or unlabeled quantitative proteomic approach, the six major steps include appropriate sample collection, protein extraction, enzymatic digestion of proteins into peptides, separation/fractionation using liquid chromatography (LC) approaches, followed by MS, peptide and protein identification and quantification and additional bioinformatics analyses such as pathway and network analyses. The field has moved forward from 2D-PAGE-based (dye/fluorescence labeling) protein spot extraction followed by LC-MS or matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS characterization to more system-wide screening approaches with quantitative steps that take advantage of label-based approaches such as Isotope-Coded Affinity Tagging, Stable Isotope Labeling with Amino Acids in Cell Culture (SILAC), ^{18}O Stable Isotope Labeling, Isobaric Tagging for Relative and Absolute Quantitation (iTRAQ) and Tandem Mass Tags (TMT) (Bakalarski & Kirkpatrick 2016) or are label-free (Bantscheff *et al.* 2012, Anand *et al.* 2017). Both label-free (Proffitt *et al.* 2017) and label-based efforts such as TMT proteomics from diverse biological matrices have yielded favorable results. The community has not yet built a consensus in terms of data formatting, cleaning and normalization, for example, the use of ion intensity vs peptide-to-spectrum matches, despite the ongoing efforts through the Proteomics Standards Initiative (Deutsch *et al.* 2017). Nonetheless, proteomics is advancing our understanding in biomedical research, including diagnosis, protein-based biomarker development and therapeutics.

Metabolomics

Metabolites are often end products of complex biochemical cascades that can link the genome, transcriptome and proteome to phenotype, providing an important key tool for discovery of the genetic basis of metabolic variation. Metabolomics can be used to determine relative and absolute amounts of sugars, lipids, amino acids, organic acids, nucleotides, steroids, drugs and environmental constituents from a wide range of sample types including primary cells, cell lines, tissues, biofluids, entire organisms and diverse geo-climatic environments. Depending on the application and instrumentation, metabolomics captures small molecule information in solid (i.e., solid-state NMR), liquid ((liquid chromatography MS (LC-MS), capillary electrophoresis MS (CE-MS)) or gas phase (gas chromatography MS (GC-MS)) using spectroscopy (i.e., NMR) and MS (i.e., LC/GC-MS or tandem MS). Major steps for metabolomics analyses include experimental design, suitable sample collection strategies, quenching of metabolism, optimized metabolite extraction and reconstitution from samples, optional chemical derivatization, MS (with or without a chromatography

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

interface) or NMR and data analysis, including data alignment, filtering, imputation, statistical analysis, annotation and pathway/network analysis. Each of these steps is highly variable depending upon the platform used for sample analysis. In addition, data structure, imputation approaches, identification of unknown metabolites, normalization, scaling and transformation can differ significantly for each data type and instrument. Approaches also differ for targeted or untargeted analyses. Choosing targeted or untargeted analyses is determined by the study question where untargeted analyses are typically used as discovery, hypothesis generating data and targeted analyses are used to test specific hypotheses. For both of these approaches, the combination of LC-MS with complementary GC-MS captures the majority of the chemical space presented by a biofluid or tissue sample.

Unique challenges to specific omics platforms

Unique challenges emanate from each omics platform due to the strengths and limitations of each. These are important to understand when developing methods and approaches for integrating omics data since the complexity and completeness of each data type differs.

Linking genotype to phenotype

High-throughput genomics and transcriptomics datasets critically depend on the ease of nucleic acid amplification from small amounts of biological material, followed by reliable quantification and molecule annotation based on sequence identity. Current sample preparation protocols provide a means to analyze all DNA and RNA in a biological sample, for example, all coding and non-coding RNAs. A major limitation is interpretation of genome and transcriptome data in the context of biological function, i.e., the influence of specific variants on phenotypic variation (Lappalainen *et al.* 2013). Combining data from proteomics and metabolomics with genomics and transcriptomics helps to overcome this limitation by providing molecular information that links genetic and epigenetic variation with phenotypic variation.

Quantification of the proteome

Proteomics data often provide information related to biological function, especially those methods quantifying isoform variation and post-translational modifications. However, proteomics approaches still require significant amounts of sample due to the lack of protein amplification methods, and face difficulties in isolation of membrane proteins, detection of low abundance proteins and insoluble proteins. For example, representation of nuclear proteins in a proteomic dataset typically requires enrichment of nuclei; thus, even untargeted proteomic approaches will not include data for all proteins within a given biological sample. The reliance on separation of complex chemistries (i.e., different charged states and post-translational modifications) using chromatography adds to variability in protein quantification in top-down and bottom-up proteomics. In addition, there is variability in peptide identification due to variation in peptide structure, charge and hydrophobicity, and these biochemical properties of peptides and proteins affect their ability to be detected and identified by NMR or MS. Analysis pipelines for proteomic data must deal with absent data (i.e., is the peptide not detected because it is not ionized efficiently, or is it truly not present in the sample), normalization and absolute vs relative quantification (Bantscheff *et al.* 2012). In recent years, advances in instrument sensitivity, and the development of effective isotopic labeling tools for tissue samples have significantly improved the accuracy and reproducibility of peptide and protein quantification using MS. This now allows the effective quantification (using peptide spectral match counts, peak intensity or peak area quantification or the use of isobaric tags for quantification) of peptides in complex mixtures such as tissue lysates.

Quantification of the metabolome

Metabolomics data can link genetic and proteomic variation to functional variation and provide novel insights into metabolic, regulatory and signaling activities in a given cell or tissue. However, similar to proteins, metabolites are not amplifiable and only 15–30% of the entire mass spectra are identifiable and quantifiable, thus limiting the usefulness of the amount of information generated. In addition, false positives are a challenge due to the use of score-based spectral annotation of molecules. Variability in sample handling, platform used, chemical heterogeneity of small molecules, different quantification methods and lack of standards for data formats and analysis pipelines are major challenges (Spicer *et al.* 2017*a*,*b*). Large-scale efforts in the metabolomics research community are currently ongoing to address these challenges including standardization, annotation of metabolites, interoperability of protocols and methods and statistical considerations.

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Issues shared among the omics platforms

Most omics approaches require knowledge of handling large datasets, annotation of biomolecules within a dataset, sample size vs number of biomolecules quantified, relevance of biomolecules quantified (signal versus noise), quality of output and accessibility of data for sharing due to data volume and complexity. The included Glossary provides definitions of fundamental terms used in this review.

Data handling

Data handling, independent of omics data type, must address issues of data filtering and cleaning (i.e., comparable to data wrangling in data science), imputation, transformation, normalization and scaling. Unfortunately, there are no 'gold standard' unified workflows for any type of omics data (although genomics and sequencing approaches often use widely accepted standards for sequence alignment, QC and/or variant calling), use of one analysis pipeline (or analysis tool, that is, search algorithm for proteomics data or statistical workflows) will yield different results than another, and workflows are constantly evolving as new computational tools are being developed and implemented. For these reasons, it is essential that every analysis pipeline is well documented, including versions of software (i.e., version control) used for each step in the pipeline and rationale for parameters implemented.

Annotation

Annotation of biomolecules for any omics dataset also provides substantial challenges. For example, standard model organisms (fly, nematode, mouse, non-human primate, human) have well-annotated genomes, transcriptomes and proteomes, and the array of tools available for interactive annotation such as miRNA/gene interactions dramatically outnumber those available for non-standard model organisms. Extensive data can be lost when working with non-standard organisms without the use of comparative approaches. That said, non-standard organisms often provide data on molecules that are relevant to human biology, but cannot readily be identified where healthy tissues are required to generate high-quality samples (as these are often challenging to collect invasively in humans). For example, use of an iterative approach to annotate transcripts for non-standard model species, where the species genome is first used for annotation and unannotated transcripts are aligned against multiple other genomes, significantly improves the number of annotated transcripts (Cox *et al.* 2012). In addition, creating peptide reference libraries using species- and individual-specific RNA-Seq transcript sequence data, significantly improves peptide annotation; a study of the baboon liver proteome by Proffitt *et al.* (2017) identified novel unannotated splice variants and 101 unique peptides missed by standard reference databases. In case of metabolomics data, not only the relative metabolite abundance, but also the chemical repertoire of an organism is often unknown, and annotation of molecules is even more challenging without the knowledge of their transcriptomes and proteomes.

Study design and analytic assumptions

To improve inferential robustness and reproducibility, a number of overarching study design and statistical concepts need to be implemented in large, omic studies. Careful study design and subject/sample (experimental unit) recruitment consistent with the study design is necessary for clear, parsimonious testing of *a priori* hypotheses and enables agnostic studies. Convenience samples can be informative but are subject to biases not present, on average, in formal randomized studies. With the exception of ancestry in genetic association studies, often these biases are largely ignored. At best, *ad hoc* methods (e.g., propensity scores analysis) can attempt to reduce the bias but are inferior to randomized designs. Understanding the degree of independence among the experimental units is important to prevent pseudoreplication. Multiple measures on the truly independent experimental unit (e.g., tissue sample, individual) requires analyses using subsampling methods or fixed or random effects repeated-measures modeling to (1) compute the proper variance estimate for tests of hypotheses and interval estimation and (2) reduce bias. Unfortunately, random and mixed-effects models generally require a large number of independent experimental units for proper type 1 error rate control.

Once a study design is selected that best addresses the study question, a set of statistical or machine-learning approaches specific to that study design and question is selected. Each analysis, whether using classical statistical or machine-learning methods, has underlying assumptions that need to be verified. Too often the large number of variables is viewed as making assumption validation impossible or not worth the investment. Further, compounding this issue is the easy access to high-speed computing with programs that use algorithms

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

often not understood by the analyst. Combined with the pressure for rapid results, these perceptions, knowledge and pressures often result in many false inferences, both type 1 and type 2 errors and significantly impact reproducibility, scientific progress and the cost of science. However, large-scale analyses with pretty graphics should not be a permit for poor-quality analyses.

An important step in a proper analysis is to clearly understand from the experimental question whether the omics variable is a predictor or an outcome. Although in some special situations it will not matter, in others it will. For example, consider an experiment with two groups (disease, disease free) and a continuous omic variable meeting the normality assumption (see below). It is well known that the standard equal variance *t*-test is asymptotically equivalent to the score test from a logistic regression model. However, adjusting for a set of covariates (e.g., age, gender, BMI) and computing the analysis of covariance instead of the *t*-test is not equivalent to the logistic model. Thus, aligning the analytic approach to match the outcome is important so that the proper variance is estimated for the test and interval estimates.

As an example, a classical statistical approach to the analysis of omic data from independent subjects (e.g. >1000 metabolites) is a linear model (e.g., analysis of variance, linear regression). Regardless of whether the omic variable is the predictor or the outcome, the methods assume that the residuals from the linear model are independent and approximate a normal distribution with a mean of zero and a constant variance. A transformation of the outcome variable (triglycerides, metabolite) is often required to meet these assumptions. Although it may seem daunting to identify an appropriate transformation for hundreds or thousands of variables, it is easy to implement the Box-Cox power transformation (Box & Cox 1964) (e.g., natural logarithm, square root, inverse) algorithm to quickly identify an appropriate distribution. For variables not easily assigned to these transformations, alternative models may need to be applied (e.g., tweetie or tobit model if the distribution has a large number of zeros and then remainder follows a normal or log-normal distribution). If the omics data are predictors in a model, the concern shifts from conditional normality of the omic variable to assuring that a few outliers do not overly influence inferences from the modeling. Such modest care in analyses can greatly accelerate true discoveries, while reducing false discoveries, thereby increasing reproducibility and lowering the ultimate cost of science.

Statistical power

Major statistical challenges for all omics data include the number of samples in a study versus the number of molecules quantified in each sample (leading to false positives and true negatives), analysis of time series data and treatment of data for targeted and untargeted (unbiased) approaches, i.e. discerning true biological signal from noise. Sample numbers per group may also vary for these different technologies with the ability to highly multiplex samples for genomics and transcriptomics, to moderately multiplex samples for proteomics, with a lack of multiplexing workflows for metabolomics. An additional challenge for integrating different omics datasets is the large variation in the number of observations per sample where a genome typically includes millions of variants, a transcriptome typically includes tens of thousands of quantified molecules, a small transcriptome includes less than 2000 molecules and proteomes and metabolomes include thousands of quantified molecules. Detection of differences in abundance of molecules also varies significantly where a transcriptome may show differences in a range of 10^5 and a metabolome may only show differences in a range of 10^3.

Data archiving and sharing

Additional issues for many omics datasets are the lack of a standardized nomenclature, data formatting and eventual public access to datasets. This has largely been addressed for genomic, transcriptomic, and proteomic data where datasets can be, and are expected to be, deposited in public databases upon manuscript publication. However, standardization of data and development of a central public database for other types of omics data is yet to be implemented and will require the definition of data standards that will allow re-analysis of deposited data, similar to the MIAME (minimum information about a microarray experiment) standards first developed for microarray data (Brazma *et al.* 2001). This was followed by minimum information about a genome sequence (MIGS) (Field *et al.* 2008), minimum information about a proteomics experiment (MIAPE) in proteomics (Taylor *et al.* 2007), metabolomics standards initiative (MSI) in metabolomics (Fiehn *et al.* 2007, Sumner *et al.* 2007), minimum information about a single amplified genome (MISAG) in genomics and minimum information about a metagenome-assembled genome (MIMAG) of bacteria and archaea in metagenomics (Bowers *et al.* 2017). In summary, abiding by these recommended practices of minimum information (i.e., MIXX) will lead to scalable

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

and interoperable protocols for generation of reproducible datasets for comparing standalone omics data sets across multiple biological samples, analytical platforms and research laboratories worldwide.

Tools available for integration of multi-omics data

Analyzing thousands of measurements in each omics experiment is a computationally complex process, where extraction of meaningful correlations and true interactions is not trivial. This is further complicated by the fact that biological systems often yield non-linear interactions and joint effects of multiple factors, making it difficult to discern true biological signals from random noise – noise can come from biological systems, unrelated analytical platforms and diverse data-specific analysis workflows. For instance, cell-type, tissue-type and organ-type specificities of gene, protein or metabolite abundances show inter-individual variability, for which biological levels of organization can pose challenges for extraction of useful data within and among these high-dimensional datasets. Increasing number of studies incorporate a diverse array of relatively newer omics approaches such as fluxomics, ionomics, microbiomics and glycomics with biomedical datasets for identification and prediction of health status or outcomes from interventions. Before omics scale data integration, data normalization is imperative given that data come from different technologies. Figure 1 summarizes a generalized integrated omics workflow. Data integration often requires statistical and even machine-learning tools (Min *et al.* 2016) for a multi-omics view (Libbrecht & Noble 2015). Machine-learning approaches are useful for combined analyses of integrated omics datasets and clinical data to facilitate dimension reduction, clustering, association with clinical measures and prediction of disease (Li *et al.* 2016).

Simplistic, descriptive and exploratory approaches such as multivariate analysis tools like principal component analysis (PCA) can often be used to reduce data dimensionality, while canonical correlation analysis (CCA) can be used to investigate the overall correlation between two sets of variables. Other omics integrative frameworks involve sparse CCA (Parkhomenko *et al.* 2009), multiple factor analysis (De Tayrac *et al.* 2009) and multivariate partial least square regression analysis (Palermo *et al.* 2009). In a recent review, Wanichthanarak *et al.* (2015) identified several available tools and packages for the integration of genomic, proteomic and metabolomic datasets using pathway enrichment, biological network or empirical correlation analysis. Nonetheless, while most of these tools require standard R-statistical programming or Python or Galaxy, implementation has been defined as 'difficult' by the authors. Thus, the need for more user friendly tools remains.

For instance, the integrated omics analyses for understanding different types of cancers at the molecular level pose additional challenges due to very high heterogeneity of samples. Pavel *et al.* (2016) used a fuzzy logic modeling framework (Xu *et al.* 2008) to integrate multiple types of omics data with expert curated biological rules for identification of cancer drivers and to infer patient-specific gene activity. To deal with sample heterogeneity, Wang and Gu (2016) have proposed three clustering categories, direct integrative clustering, clustering of clusters and regulatory integrative clustering. Nibbe *et al.* (2010) demonstrated that integration of complementary data sources (transcriptomic and proteomic data) using a 'proteomics-first' approach can enhance discovery of candidate sub-networks in cancer. This approach, which identifies proteomic targets with significant fold-changes between tumor and control tissues, can be used to 'seed' novel networks that reveal protein–protein interaction (PPI) sub-networks functionally associated with phenotype. This approach has led to the discovery of protein–protein interaction-based changes in human colorectal cancer tissues (Nibbe *et al.* 2010). Ideally, network generation approaches will not rely predominantly on known function(s) of a molecule since many genes and proteins have been shown to have different activities and functions in different biological systems, and the system being investigated may include key molecules with novel functions and/or novel molecules. Even though weighted gene coexpression network analysis (WGCNA) has been heavily used for unbiased integration of genomic and transcriptomic data with quantitative trait data to identify coordinated modules of genes and gene variants associated with variation in phenotypic variation, it remains to be seen whether this algorithm is useful for integration of other omics datasets from diverse analytical platforms (e.g., proteomics, metabolomics, etc.) or more heterogeneous data such as various types of clinical data.

Currently available tools for integration of omics data include web-based tools requiring no computational experience as well as more versatile tools for those with computational experience. User friendly, web-based tools requiring no computational experience include Paintomics, 3Omics and Galaxy (P, M). However, the application of user friendly tools should not be done without an understanding of the underlying methods. Blind application of easy use tools

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

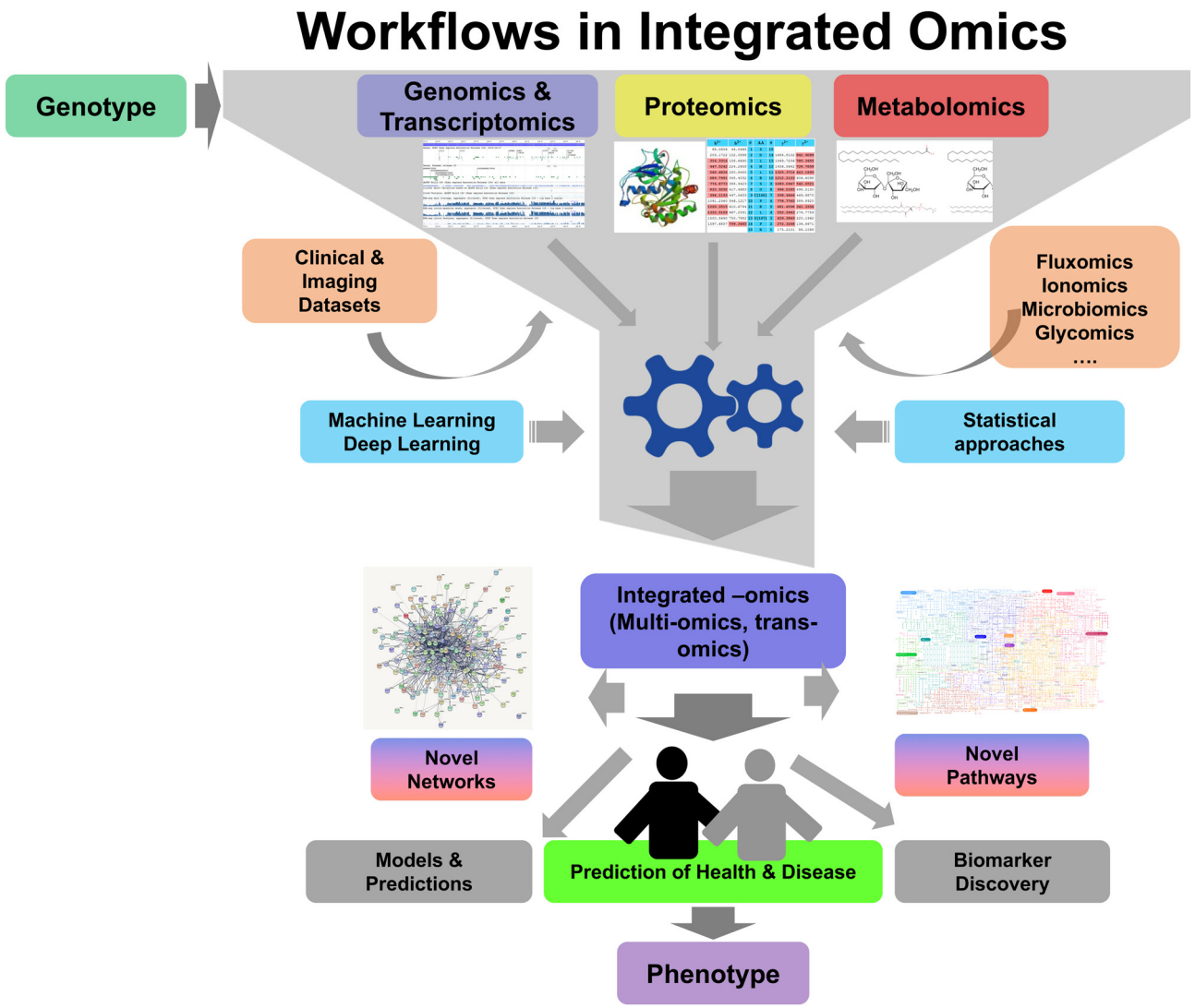

Figure 1
A typical integrated omics workflow showing input datasets, output datasets and results. Using individual omics datasets that are closer to genotype (genomics and transcriptomics), and those closer to phenotype (proteomics and metabolomics), plus a host of other omics platforms, datasets are integrated using statistical or advanced machine learning approaches. Results may be simple pathways or complex networks and may include both known and novel molecules. In addition, results may predict health or disease states, provide insights for effective therapeutic interventions, or reveal spatio-temporal regulation of systems such as cell-, tissue- or organ-type specificity.

often adversely affects progress in the field and ultimately makes science cost more (e.g., unnecessary additional studies to debunk entrenched falsehoods). For more advanced users with expertise in programming and interfacing with computational tools, tools such as IntegrOmics, SteinerNet, Omics Integrator, MixOmics are available. These tools allow customization of various parameters and settings allowing more control of data analyses. Those interested in integration of datasets driven from metabolomics can opt for online tools such as XCMSOnline, which allows multi-omics integration of metabolomics data with genomics and proteomics as well (Table 1).

Recent examples of integration in real world datasets

A majority of the current literature uses terms such as multi-omics and integrated omics to denote research efforts where only two omics datasets were integrated (e.g. transcriptomics and proteomics, or proteomics and metabolomics, etc.), and multiple cases where the omics datasets integrated were only at the level of the genome (e.g., ChipSeq and methylomics). As part of this review, we highlight recent examples of successful multi-omics integration which include at least three different omics

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Table 1 List of various tools, software, statistical approaches and databases available for integrated –omics approaches.

Name	Computational platform	User friendliness	Functionality	Availability	Reference
Omics data integration tools					
MapMan	Java	Easy	Visualize and map gene expression, metabolite or other data, displays large data sets onto diagrams of metabolic pathways	https://mapman.gabipd.org/	Thimm *et al.* (2004)
Weighted Gene Coexpression Network Analysis (WGCNA)	R	Moderate	A comprehensive collection of R functions for performing various aspects of weighted correlation network analysis	https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/	Langfelder & Horvath (2008)
iCluster	R	Difficult	Detection of novel biomarkers, their ranking and annotation with existing knowledge using corresponding Transcriptomics and Proteomics data sets	https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster	Shen *et al.* (2009)
Pathway Studio, Ariadne Genomics	License, Web, Local	Easy	Analysis and visualization of disease mechanisms, gene expression and proteomics and metabolomics data	http://www.pathwaystudio.com/	Yuryev *et al.* (2009)
IntegrOmics	R	Difficult	Efficiently performs integrative analyses of two types of 'omics' variables that are measured on the same samples	http://math.univ-toulouse.fr/biostat	Cao *et al.* (2009)
Paintomics	Web	Easy	Integrated visual analysis of transcriptomics and metabolomics data	http://www.paintomics.org	García-Alcalde *et al.* (2011)
IMPaLA	Web	Easy	Joint pathway analysis of transcriptomics or proteomics and metabolomics data that also performs over-representation or enrichment analysis	http://impala.molgen.mpg.de	Kamburov *et al.* (2011)
SteinerNet	R	Moderate	Integrating transcriptional, proteomic and interactome data by searching for the solution to the prize-collecting Steiner tree problem	https://cran.r-project.org/src/contrib/Archive/SteinerNet/	Tuncbag *et al.* (2012)
PhenoLink	Web	Easy	Phenotype links to a multitude of ~omics data, e.g., gene presence/absence (determined by e.g.: CGH or next-generation sequencing), gene expression (determined by e.g.: microarrays or RNA-Seq), or metabolite abundance (determined by e.g.: GC-MS)	http://bamics2.cmbi.ru.nl/websoftware/phenolink/	Bayjanov *et al.* (2012)
3Omics	Web	Easy	Integrating multiple inter- or intra-transcriptomic, proteomic, and metabolomic human data	http://3omics.cmdm.tw/	Kuo *et al.* (2013)
CrossPlatform Commander	R	Difficult	Detection of novel biomarkers, their ranking and annotation with existing knowledge using corresponding Transcriptomics and Proteomics data sets	http://www.ruhr-uni-bochum.de/mpc/software/xplatcom/index.html.en	Kohl *et al.* (2014)
Multi-Omics Data matcher	NA	Difficult	Identify and correct sample labeling errors in multiple types of molecular data, which can be used in further integrative analysis	http://research.mssm.edu/integrative-network-biology/Software.html	Yoo *et al.* (2014)
Ingenuity Pathway Analysis, Qiagen	License, Web, Local	Easy	Integration and mapping of genomics, transcriptomics, proteomics, and metabolomics datasets	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/	Krämer *et al.* (2014)
OncoIMPACT	R	Difficult	Algorithmic framework that nominates patient-specific driver genes by integratively modeling genomic mutations (point, structural and copy number) and the resulting perturbations in transcriptional programs via defined molecular networks	https://github.com/CSB5/OncoIMPACT	Bertrand *et al.* (2015)
GalaxyP, GalaxyM	Web	Easy	Development of a complete suite for integrated omics analysis, proteomics informed by transcriptomics analysis available to the typical bench scientist	https://usegalaxy.org/	Fan *et al.* (2015), Davidson *et al.* (2016)
Omics Integrator	Python, Web	Easy	Integrate proteomic data, gene expression data and/or epigenetic data using a protein–protein interaction network	http://fraenkel.mit.edu/omicsintegrator, https://github.com/fraenkel-lab/OmicsIntegrator	Tuncbag *et al.* (2016)
MONGKIE	Java	Easy	Multi-layered omics data such as somatic mutations, copy number variations, and gene expression data	http://yjjang.github.io/mongkie/	Jang *et al.* (2016)
MixOmics	R	Difficult	Provides a wide range of linear multivariate methods for data exploration, integration, dimension reduction and visualization of biological data sets	http://mixomics.org/	Rohart *et al.* (2017)
Statistical approaches for integration					
CAusal Modelling with Expression Linkage for cOmplex Traits (Camelot)	Matlab	Difficult	Integrates genotype, gene expression and phenotype data to build models	https://www.c2b2.columbia.edu/danapeerlab/html/camelot.html	Chen *et al.* (2009)

(*Continued*)

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Table 1 Continued.

Name	Computational platform	User friendliness	Functionality	Availability	Reference
Transcriptional Modules Discovery (TMD)	NA	NA	Network-free Bayesian approach that adopts a mixture modeling approach using hierarchical Dirichelet process to perform integrative modeling of two datasets	NA	Savage *et al.* (2010)
Sparse multiblock PLS (sMBPLS)	MATLAB	Difficult	Multi-dimensional regulatory modules from several layers of genomic datasets	http://zhoulab.usc.edu/sMBPLS/	Li *et al.* (2012)
Multiple dataset integration (MDI)	R, C++	Difficult	Integrates information from a wide range of different datasets and data types simultaneously including capabilities to model time series data using Gaussian processes	https://github.com/smason/mdipp	Kirk *et al.* (2012)
OnPLS	Python	Difficult	Multiblock data analysis with prefiltering of unique and locally joint variation	https://github.com/tomlof/OnPLS	Srivastava *et al.* (2013)
Factor Analysis (FA) and linear discriminant analysis (LDA) (FALDA)	NA	Difficult	Discriminate different classes of samples based on standardization and merger of several omics	NA	Liu *et al.* (2013)
Weighted multiplex networks	NA	NA	Weighted multiplex networks are characterized by significant correlations across layers	NA	Menichetti *et al.* (2014)
Multiple coinertia analysis (MCIA)	R	Difficult	Exploratory data analysis method that identifies co-relationships between multiple high-dimensional datasets	https://rdrr.io/bioc/omicade4/man/mcia.html	Meng *et al.* (2014)
moCluster	R	Difficult	Gene set analysis based on multiple omics data	https://www.bioconductor.org/packages/release/bioc/html/mogsa.html	Meng *et al.* (2016)
Tools for integration within the domain of genomics					
iCluster	R	Difficult	Joint latent variable model for integrative clustering that incorporates flexible modeling of the associations between different data types and the variance–covariance structure within data types	https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster	Shen *et al.* (2009)
Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction (COALESCE)	Web	Easy	Genomic data and Bayesian integration to predict co-regulated gene modules	http://quantbio-tools.princeton.edu/cgi-bin/COALESCE	Huttenhower *et al.* (2009)
Multiple Concerted Disruption method	R	Difficult	Integrates CNV, DNA methylation, and allelic loss of heterozygosity status to find genes representing key nodes in the pathways and significant genes	CRAN	Chari *et al.* (2010)
PARADIGM	Commercial	Difficult	Identifies pathway-level activities from multi-dimensional cancer genomics datasets	http://five3genomics.com/technologies/paradigm	Vaske *et al.* (2010)
COpy Number and EXpression In Cancer (CONEXIC)	Java	Medium	Integrates matched copy number, amplifications and deletions, and gene expression data from tumor samples	https://www.c2b2.columbia.edu/danapeerlab/html/conexic.html	Akavia *et al.* (2010)
CNAmet	R	Difficult	Integrative analysis of high-throughput copy number, DNA methylation and gene expression data	http://csbi.ltdk.helsinki.fi/CNAmet/	Louhimo and Hautaniemi (2011)
Patient-specific Data Fusion (PSDF)	MATLAB	Difficult	Bayesian nonparametric modeling that integrates copy number and expression data to jointly classify patients into cancer subgroups	https://sites.google.com/site/patientspecificdatafusion/	Yuan *et al.* (2011)
PLRS	R	Difficult	flexible modeling of the association between DNA copy number and mRNA expression	http://bioconductor.org/	Leday & van de Wiel (2013)
NuChart	R	Difficult	Annotation and statistical analysis of a list of input genes with information relying on high-throughput sequencing data, integrating knowledge about genomic features that are involved in the chromosome spatial organization	NF	Merelli *et al.* (2013)
In-Trans Process Associated and Cis-Correlated (iPAC)	NA	NA	Multi-step method to identify genes that are in-cis correlated through integrating gene expression and CNV data, as well as genes that are in-trans associated to the biological processes	CRAN	Aure *et al.* (2013)
Multi-objective optimization (MOO)	R	Difficult	Generates networks of biological components that incorporate multi-omics information, such as transcriptomics data from two different sources	CRAN	Mosca and Milanesi (2013)
Network smoothed T-statistic SVMs (stSVM)/netClass	R	Difficult	Integrates network information and other kinds of experimental data into one classifier, by smoothing t-statistics of individual genes or miRNAs over the structure of a combined protein–protein interaction and miRNA-target gene network	https://sourceforge.net/projects/netclassr/	Cun & Fröhlich (2014)
BioMiner	Web	Easy	BioMiner incorporates transcriptomic and cross-omics high-throughput data sets, with a focus on cancer	http://systherDB.microdiscovery.de/	Bauer *et al.* (2015)

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

miRTarVis	Java	Easy	miRNA–mRNA integration	http://hcil.snu.ac.kr/~rati/miRTarVis/index.html	Jung *et al.* (2015)
Omics Pipe	Python	Difficult	Integration of RNA-Seq, miRNA-seq, Exome-seq, Whole-Genome sequencing, ChIP-seq analyses	http://sulab.scripps.edu/omicspipe	Fisch *et al.* (2015)
CPAS	C, R		Trans-omics pathway analysis of genome-wide CNVs and mRNA expression profiles data, a gene set enrichment analysis algorithm	NA	Zhang *et al.* (2015)
CPAS	R, C	Difficult	Recognizes disease relevant biological pathways through joint pathway analysis of genome-wide copy numbers variants (CNVs) and mRNA expression profile data	https://sourceforge.net/projects/%20cpasv1/files/	Zhang *et al.* (2015)
Galaxy Integrated Omics (GIO)	Galaxy	Easy	Transcriptomics, proteomics	https://usegalaxy.org/	Fan *et al.* (2015)
BioVLAB-mCpG-SNP-EXPRESS	Web	Easy	Integrated analysis of gene expression, DNA methylation, and genetic variations	http://bhi2.snu.ac.kr:3000/	Chae *et al.* (2016)
GeneTrail2	Web	Easy	Integrated analysis of transcriptomic, miRNomic, genomic and proteomic datasets	https://genetrail2.bioinf.uni-sb.de/	Stöckel *et al.* (2016)
Mergeomics	R	Difficult	Genetic association (e.g., GWAS or exome sequencing), transcriptome-wide association (e.g., TWAS from microarray or RNA sequencing studies), and epigenetic association (e.g., EWAS from methylome association studies), functional genomics (such as eQTLs and ENCODE annotations), biological pathways, and gene networks	http://mergeomics.research.idre.ucla.edu/	Shu *et al.* (2016)
MultiAssayExperiment	R, Bioconductor	Difficult	Storage, and operation on multiple diverse genomic data, i.e., Cancer Genome Atlas data followed by scalable, reproducible statistical analysis of multiomics data	https://bioconductor.org/packages/release/bioc/html/MultiAssayExperiment.html	Ramos *et al.* (2017)
Databases and resources for integration					
NCBI – multiple databases	Web	Easy	Genomics and transcriptomics	https://www.ncbi.nlm.nih.gov/	Pruitt *et al.* (2006)
ProteomeXchange Consortium	Web	Easy	Proteomics	http://www.proteomexchange.org/	Vizcaíno *et al.* (2014)
MOPED	Web	NF	Database with a multi-omics resource portal that combines 250 publicly available protein and mRNA abundance profiles of four different model organisms, human, mouse, worm and yeast	http://moped.proteinspire.org	Montague *et al.* (2014)
OMICtools	Web	Easy	Repositories which aids in integration of omics datasets	https://omictools.com/	Henry *et al.* (2014)
MetabolomicsWorkBench	Web	Easy	Metabolomics datasets archiving	http://www.metabolomicsworkbench.org/	Sud *et al.* (2015)
CardioGenBase	Web	Easy	Database contains more than ~1500 CVD associated genes and relevant information from about ~24000 publications	http://www.cardiogenbase.com/faq.php	Alexandar *et al.* (2015)
Global Natural Products Social 242 Molecular Networking (GNPS)	Web	Easy	Metabolomics (untargeted datasets) archiving	http://gnps.ucsd.edu	Wang *et al.* (2016)
MetaboLights	Web	Easy	Metabolomics datasets archiving	http://www.ebi.ac.uk/metabolights/	Kale *et al.* (2016)
Methods for Integrated analysis of multiple Omics datasets (MIMOmics)	Web	NA	Innovation project funded by the European Commission, running from October 2012 till October 2017, coordinated by the Leiden University Medical Center	http://www.mimomics.eu	Auffray *et al.* (2016)
Ecomics	Web	Easy	Multi-omics compendium for Escherichia coli with cohesive meta-data information semi-supervised normalization pipelines and perform experimental characterization, growth, transcriptome, proteome	http://prokaryomics.com/	Kim *et al.* (2016)
Omics Database Generator	Clojure, Java	Difficult	Uses genome files and output from various programs to create a graph database for querying genomic data across domains	https://github.com/jguhlin/odg	Guhlin *et al.* (2017)
PeptideAtlas repository	Web	Easy	Proteomics	http://www.peptideatlas.org/PASS/PASS00512	Desiere *et al.* (2006)
Proteomics Identifications Database (PRIDE)	Web	Easy	Proteomics	https://www.ebi.ac.uk/pride/archive/	Vizcaíno *et al.* (2013)
WikiPathways	Web, R	Easy	Collage of pathways amenable to automated and manual workflows for mapping of genes, proteins, and metabolites	http://wikipathways.org/	Slenter *et al.* (2018)
G6G Directory of Omics and Intelligent Software	Web	Easy	Directory to obtain an array of commercially available and free tools for omics analysis	http://g6g-softwaredirectory.com/apps/bio/cross-omics/pathway-dbs-kbs/ListingsByAppCOPathwayKBDB.php	NA
XCMS Online	Web	Easy	Systems biology scale workflow, that allows rapid metabolic pathway mapping from raw metabolomics data to integration of genomic and proteomics data for mechanistic insights	https://xcmsonline.scripps.edu/	Forsberg *et al.* (2018)

The table provides the computational platform in which tools are available (web based or programming language), degree of user friendliness, functionalities, availability, and associated cited literature in their chronological order of appearance. Ease of use definitions: difficult – requires in-depth understanding and knowledge of the specified programming language; medium – modest level of proficiency or programming skills; andeasy – requires minor level of skill to implement the tool.

CGH, comparative genomic hybridization; ChIP-seq, chromatin immunoprecipitation sequencing; ENCODE, Encyclopedia Of DNA Elements; eQTL, expression quantitative trait loci; NA, not available; NF, not found; WGCNA, Weighted Gene Coexpression Network Analysis.

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain
Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

platforms and allow the discovery of novel biological factors and/or processes through this approach.

Complex diseases

Williams *et al.* (2016) integrated sequential window acquisition of all theoretical mass spectra (SWATH MS) generated proteomics data with metabolomics and genomics datasets for a systems level assessment of liver mitochondrial function. This study included 386 mice from the BXD recombinant inbred strain and used three omics datasets – transcriptome (25,136 transcripts), proteome (2622 proteins) and metabolome (981 metabolites). They validated interactions of key molecules nominated from this approach and showed that sequence variants in the Cox7a2l gene alter the encoded protein's activity, leading to downstream differences in mitochondrial super complex formation. This study demonstrates the utility of omics integration for identification of functional variants underlying complex diseases.

Zierer *et al.* (2016) integrated epigenomics, transcriptomics, glycomics and metabolomics, with disease traits from 510 participants of the TwinsUK cohort to find molecular pathways underlying age-related diseases. Using network analysis where the mixed graphical model was inferred using the *Graphical Random Forest* (*GRaFo*) method, they identified seven modules representative of distinct aspects of aging. Their findings demonstrate interconnectivity in age-related diseases and that use of integrated omics can reveal novel molecular networks relevant to complex phenotypes.

Krishnan *et al.* (2018) used adipose and liver tissue gene expression analysis by microarray, bioenergetics measurements in cell lines and mitochondria followed by GWAS and eQTL analyses to integrate various omics datasets via an advanced multiscale embedded gene coexpression network analysis (Song & Zhang 2015) that was preferred over WGCNA analyses for identification of networks. Clearly, the authors concluded that network modeling from a large dataset and *in vitro* approaches helped predict key driver genes regulating non-alcoholic fatty liver disease.

Recently, using a BXD mouse cohort as sources for multi-omics analysis (including (expression-based) phenome-wide association, transcriptome-/proteome-wide association and (reverse-) mediation analysis), Li *et al.* (2017) demonstrated the feasibility for identification of gene–gene, gene–phenotype links that are translatable to cross populations and species in their multi-omics framework.

Immunity and infection

The integrative personal omics profile (iPOP) is a pioneering study that combined genomics, transcriptomics, proteomics, metabolomics and autoantibody profiles from a single individual over a 14-month period. In this approach, pathways enriched for differentially expressed molecules were computed at each time point, while taking into account pathway structure and longitudinal design (Stanberry *et al.* 2013). Similar endeavors in organ-specific multi-omic integrated tools include kidney and urinary pathway knowledge base for kidney diseases where as an example of the utility of this integrated database to facilitate rapid hypothesis generation, the authors identified calreticulin as a protein central to human interstitial fibrosis and tubular atrophy in chronic kidney transplant rejection and validated the importance of this protein *in vitro* and *in vivo* (Klein *et al.* 2012). Another study characterized response of ferrets to pandemic H1N1 influenza viral infection using an integrated omics approach with lipidomic, metabolomic and proteomic datasets and discovered that pro-inflammatory lipid precursors impact virus pathogenesis (Tisoncik-Go *et al.* 2016). These studies highlight the power of integrated omics approaches to identify novel molecules that influence immune function and infection.

Cancer

Liu *et al.* (2013) used both an integrated and a non-integrated approach to analyze the NCI-60 cancer cell line panel to identify potential molecular mechanisms dysregulated in cancer. They performed joint analysis of the small transcriptome (miRNAs), transcriptome and proteome using factor analysis with linear discriminant analysis (LDA) and demonstrated that the integrated approach provides a more complete picture of miRNA/gene interactions in the *Wnt* signaling pathway, which is a surrogate marker of melanoma progression. Liu *et al.* (2016) generated and integrated data on genomic CNVs, genomic methylation, transcriptome and small transcriptome datasets to characterize subtypes of hepatocellular carcinoma. Using 256 hepatocellular carcinoma samples, they identified five hepatocellular carcinoma subgroups with distinct molecular signatures, and each with a distinct survival rate. Other studies have used this approach obtaining high quality and comprehensive omics measurements, followed by integrated omics analysis to describe molecular variation in other cancer types (Jiang *et al.* 2016, Kamoun *et al.* 2016). Further, MiRbooking algorithms provide vital insights

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

into integration of miRNA–mRNA in hybridization competition that occurs in a given cellular condition (Weill *et al.* 2015). An integrated omics approach in cancer may provide information for improved diagnosis of carcinoma subtype pathogenesis. Recently, Muqaku *et al.* (2017), using both label-free and targeted proteomics, lipidomics and metabolomics efforts followed by data integration in human serum samples from patients with metastatic melanoma, proposed a model on reprogramming of organ functions induced by metastatic melanoma through formation of platelet activating factors from long-chain polyunsaturated phosphatidylcholines under oxidative conditions.

Host microbiome interactions

Heintz-Buschart *et al.* (2016) used a multi-omics approach integrating metagenomic, metatranscriptomic and meta-proteomic data from the gastrointestinal microbiome to identify intra- and inter-individual variation in subjects with type 1 diabetes mellitus (T1DM). The study revealed several microbial populations contributing to functional differences among T1DM individuals. Thaiss *et al.* (2016) used integrated omics of the transcriptome, methylome, metagenome and metabolome with imaging data to quantify the global programming of the host circadian transcriptional, epigenetic and metabolite oscillations by intestinal microbiota. They found that the gut microbiome and host circadian activities are tightly linked and showed that disruption of microbiome rhythmicity abrogates normal host genome, epigenome and transcriptional oscillations in both intestine and liver, influencing host diurnal fluctuations. These integrated omics studies are beginning to reveal the complex interactions between the host and the gut microbiome, and the resulting impact on host metabolism.

Quinn *et al.* (2016) implemented an integrated omics pipeline for human and environmental omic samples, 16S rRNA gene sequencing, inferred gene function profiles and LC-MS/MS metabolomics, in less than 48 h using Qiita, Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) and Global Natural Product Social Molecular Networking (GNPS) pipelines. This study demonstrated feasibility for using an omics approach to assess human health status in a time frame that matches traditional clinically relevant culture-based approaches. Additional studies of this type may provide more feasible and accurate methods for microbe identification in the clinic and eliminate the need for culturing microbes.

Statistical approaches for current challenges

The 1930s graph theory and 1950s holistic general system theory form the basis of diverse mathematical tools for network analysis across various scientific disciplines including integration of omics datasets. Graph theory defines a graph as a set of nodes with each pair joined by an edge, and each edge associated with two nodes that form an unordered pair. Holistic general system theory defines a system as an entity with interrelated and interdependent parts, and changing any one part affects other parts and affects the entire system in predictable patterns. This theory has, since then formed the corner stone of large-scale high-throughput and high-dimensional data set oriented omics studies.

Number of samples vs number of molecules

Optimal statistical analyses are central to the computational framework for omics data integration. Each omics layer, and the underlying analytical methodology, harbors different levels of noise (Arakawa & Tomita 2013). Sampling directly impacts the appropriate statistical tools employed and must be defined prior to sampling for a given study. Bayesian network-based analyses have been used to robustly integrate multiple high-dimensional datasets, even with small sample sizes (Mukherjee & Speed 2008, Wang *et al.* 2015). The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 2011) and the Elastic Net (ENET) (Zou & Hastie 2005) approaches are penalized regression methods that, after appropriate standardization, can model more than one type of omics data, which all must deal with multi-collinearity issues and mitigate the '$n<<p$' problem, i.e., the number of independent samples (n) is much smaller than the number of measurements per sample (p). Statistical solutions include the orthogonal partial least squares (O2PLS), multivariate regression methods, regularized generalized CCA (RGCCA), principal component analysis extensions (STATIS, dual-STATIS, DISTATIS, ANISOSTATIS etc.), multiblock redundancy analysis (mbRA) and multiblock continuum redundancy (MCR) (Rajasundaram & Selbig 2016).

Dimension reduction

Multi-omics-derived datasets are high-dimensional in nature, and their handling can be computationally intensive. Dimension reduction is one strategy to reduce the computational burden while also addressing multiple testing concerns. Tools for dimension reduction that deal

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

with data heterogeneity are essential, but currently limited. Popular data dimensionality reduction approaches lack value ratio calculations, low variance and high correlation filters and random forest, PCA, and backward or forward feature elimination approaches. PCA is currently the most widely used dimension reduction approach for omics studies, as discussed by Meng *et al.* (2016). Essentially, dimension reduction techniques for integrative analysis include multiple coinertia analysis (MCIA), generalized CCA (gCCA), regularized generalized CCA (rGCCA), sparse generalized CCA (sGCCA), structuration des tableaux á trois indices de la statistique (STATIS) X-statis family of methods (STATIS), higher order generalizations of SVD and PCA (CANDECOMP/PARAFAC/Tucker3) and partial triadic analysis, and CIA (statico), all of which are available as R-packages at CRAN. These methods extract linear relationships that best explain correlated structures across datasets and variability both within and between observations. In addition, they may reveal issues such as batch effects or outliers in a given dataset. For a more detailed view on the predictive modeling and analytics approaches, the readers are advised to consult a recent review by Kim and Tagkopoulos (2018).

Data integration

Most methods implemented for data integration have relied on PCA, correlation or Bayesian or non-Bayesian network-based methods. All approaches estimate instability, model over-fitting and local convergence. Large standard errors compromise the predictive advantage provided by multiple measures. Also, it is difficult to reliably estimate many parameters and correctly infer associations from multiple hypotheses tested simultaneously. As a result, analysis of both single and integrated omics data is prone to high rates of false positives due to chance events. Thus, multiple testing must be addressed in the analytical pipeline to control for both type I error rate (e.g. Bonferroni corrections, Westfall and Young permutation) and false-positive rate (e.g. Benjamini–Hochberg).

Bersanelli *et al.* (2016) provided a detailed review of current integration tools and underlying mathematics. They defined four classes of integrative methods for reduction of multi-omics data: network-free non-Bayesian (NF-NBY), network-free Bayesian (NF-BY), network-based non-Bayesian (NB-NBY) and network-based Bayesian (NB-BY) methods. Based on current knowledge and tools, they conclude that for network-based applications, Bayesian network approaches are a useful compromise between network analysis and probability theory, where the Bayesian framework addresses noise, and errors from noise can be taken into account at the beginning of analyses. Huang *et al.* (2017) provides a review on currently available computational resources and algorithms for genomic data – i.e., genomics, transcriptomics, miRNAomics, ChIP-sequencing and gene arrays. These genomic tools are less than ideal for all types of omics datasets. However, the methods summarized here are critical for future development of more robust and less error-prone tools for integration of diverse omic datasets.

Table 1 summarizes current tools, software and approaches including the computational platform in which they can be implemented, their user friendliness, functionalities, current availability status and links and associated cited literature.

Current challenges and looking to future

We highlight five essential areas in the integrated omics workflow which are (i) experimental challenges, (ii) individual omics datasets, (iii) integration issues, (iv) data issues and (v) biological knowledge. Figure 2 summarizes the current challenges posed by integrated omics approaches.

Experimental challenges

Challenges in sample preparation

Numerous reviews have underscored the challenges for efficient sample preparation from diverse samples for individual omics studies, ranging from plants, animals and microbes for genomics (van Dijk *et al.* 2014), transcriptomics (Chomczynski & Sacchi 2006), proteomics (Wiśniewski *et al.* 2009, Erickson *et al.* 2017) and metabolomics (Villas-Bôas *et al.* 2005, Bruce *et al.* 2009, Kim & Verpoorte 2010). More focused efforts are also available such as sample preparation for fecal metabolomics (Deda *et al.* 2017), lipidomics (Teo *et al.* 2015), single-cell genomics (Vitak *et al.* 2017, Zahn *et al.* 2017) among others. However, with multi-omics, the sample amount becomes one of the major bottlenecks, further challenged by unified extraction strategies amenable for simultaneous extraction of nucleic acids, proteins and metabolites from a given matrix without significant loss. Thus, single tube extraction methods were proposed to allow for multi-phasic extraction of the three types of biomolecules as well (Valledor *et al.* 2014). Not only academic efforts, but commercial kits are currently being made available to address sample preparation for integrated omics analysis. For instance,

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

metabolite, protein and lipid extraction (MPLEx) protocol was proposed to be a robust method that is potentially applicable to a diverse set of sample types, including cell cultures, microbial communities and tissues (Nakayasu *et al.* 2016). Recently, a simultaneous metabolite, protein, lipid extraction (SIMPLEX) procedure was proposed as a novel strategy for the quantitative investigation of lipids, metabolites and proteins that allowed quantification of 360 lipids, 75 metabolites, and 3327 proteins from only 10^6 cells (Coman *et al.* 2016). Some of these methods have been optimized to yield data from samples to multi-omics under 48 h (Quinn *et al.* 2016). However, the unified sample preparation workflows are in their infancy with current methods typically providing unequal sample quality, such as combined extractions of DNA, RNA and proteins; significant work is required to achieve universal applications for diverse biological matrices.

Optimizing, documenting and sharing workflows

The success of an integrated omics workflow depends on a robust experimental design and execution. This includes the sample handling workflow with optimized sample collection and preparation protocols that allow analysis of a given material in a single step for generating multiple omics datasets. This increases comparability of multiple omics datasets and limits batch effects and technical variation issues that often plague high-dimensional data generation workflows.

It is essential to define and document all steps in the data handling workflow, including generation of individual omics datasets and integration of omics datasets. Handling, storage and analysis of multi-omics data is computationally intensive, and each step in an analytical pipeline generates new output files. A critical aspect of every pipeline is determining which output files to save and share. These decisions, which take into account the time and effort required to generate output files for each step in the workflow, impact the types and amount of data that must be processed and stored. In turn, data analytical pipeline implementation requires decisions on use of cloud infrastructures or local hardware/software for data storage and processing.

To this end, The Konstanz Information Miner (KNIME)-based modular environment workflow incorporates steps from data preprocessing to statistical analysis and visualization of omics scale data (Berthold *et al.* 2009). BioMart, Taverna and the BII Infrastructure are other workflow management systems, which help in the omics data integration and streamlining of the process. BioMart (http://www.biomart.org) is a query-oriented database management system developed jointly by the Ontario Institute for Cancer Research and the European Bioinformatics Institute. The Taverna workbench (http://taverna.sourceforge.net) is a free software tool for designing and executing workflows, created by the myGrid project (http://www.mygrid.org.uk/tools/taverna). Toward workflow sharing, myExperiment (https://www.myexperiment.org/home) is another growing collaborative environment where scientists can safely publish their workflows and *in silico* experiments, share them with groups and view workflows constructed by others (Goble *et al.* 2010). Journals publishing metabolomics studies may soon require inclusion of the workflow with the submitted manuscript (Sarpe & Schriemer 2017), which would significantly improve quality, reproducibility and utility of datasets.

Data processing

The decision to integrate 'raw' datasets to yield a merged dataset for further processing, or to first process each independent omics dataset and then merge significant results for further interpretation, has a significant impact on the final results obtained. Analysis tools chosen for integrative efforts also have a significant impact on outcomes. While several single-omics scale imputation methods are known (e.g. KNN impute imputation using k-nearest neighbors, BPCA, singular value decomposition (SVDimpute), local least squares and iterated local least squares (iLLS)), missing data imputation is challenging for multi-omics datasets. Iterative processes for imputation in a data-dependent manner are also needed. To improve imputation accuracy, a recent novel multi-omics imputation approach that integrates multiple correlated omics datasets by combining estimates of missing values from individual omics datasets, and concomitantly imputing multiple missing omics data points by an iterative algorithm was put forward (Lin *et al.* 2016).

Time course studies

Sampling time courses are important for understanding integrated network dynamics. However, this poses additional issues as response times differ for transcriptomic, proteomic and metabolomic changes. No tools exist to compensate for these differences even if assuming a steady state of -omes within a cell at a given time.

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

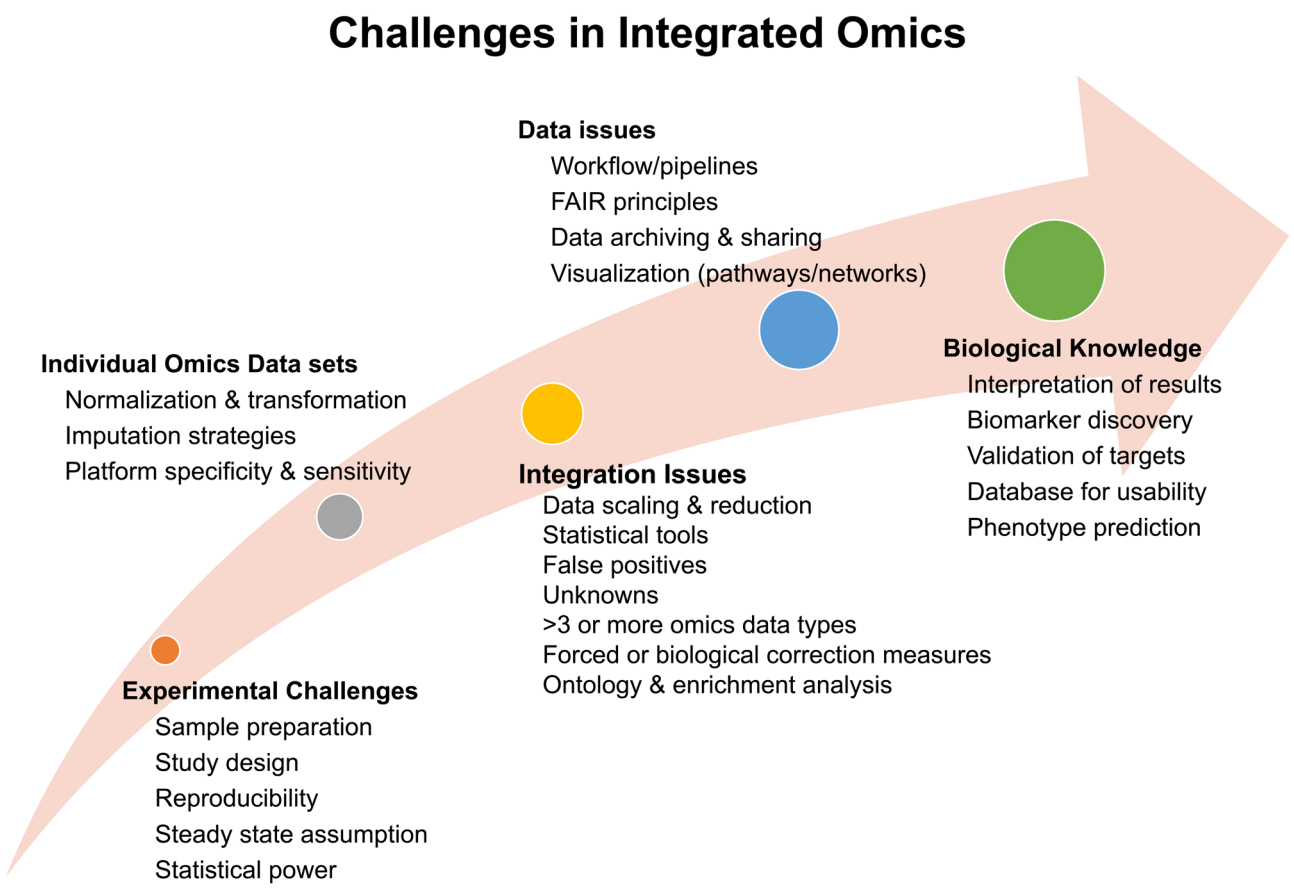

Figure 2
Five challenges associated with integrated omics which encompass (A) experimental challenges, (B) individual omics datasets, (C) integration issues, (D) data issues and (E) biological knowledge.

Individual omics datasets – normalization, transformation of different omics data types

As stated previously, each omics platform has unique limitations. Normalization, transformation and scaling approaches in the three major omics fields, i.e., transcriptomics, proteomics and metabolomics, are very different due to differences in the information included in a dataset. For example, a zero value in a RNA-Seq-based transcriptome dataset is treated as non-expression for that transcript, whereas a zero value in a proteomic or metabolomic dataset may represent either non-expression or simply missing data (e.g. for technical reasons, owing to the complexity of MS-based analyses). Consequently, imputation of missing values must be addressed differently for the different types of datasets.

Integration issues – data scaling, false positives and unknowns

Tools for scaling datasets and addressing false positives from three or more independent platforms for integration and subsequent analysis have not yet been developed. Integration is challenging for a wide array of reasons. Platforms for genomics and transcriptomics vs MS-based proteomics and metabolomics platforms operate in different numerical scales, different dynamic ranges of detection and quantification and different time scales, for example, the variation in turnover rates of transcripts, proteins and metabolites. In addition, integration of data from multiple sources increases difficulty accounting for false positives in the combined datasets. The decision to address false positives in individual omics datasets dramatically impacts results. For instance, until recently, FDR estimation methods have not been available for metabolomics datasets due to variability in spectral matching scoring and non-consensus in the MS databases (Scheubert *et al.* 2017), whereas FDR statistical methods have been available for genomics and transcriptomics for more than a decade. Additional issues include stringency of correction where stringent approaches typically rely on data structure and statistical models compared with less stringent approaches that include biologically guided

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

integration using tools such as pathway or ontology enrichment analyses (Khatri *et al.* 2012).

Currently, there is no consensus for adopting a single workflow for data integration. Some investigators have used the WGCNA approach (Langfelder & Horvath 2008) adopted from transcriptomics/microarray analyses for integrated omics workflows. While this has been useful, it does not provide a means to address unique data structures of different omics datasets in biomedical research (Smith *et al.* 2007) To this end, recently, an R-package, MultiDataSet was proposed for encapsulating multiple data sets with application to -omics data integration, keeping in mind the different data structure (list of matrices) generated from individual omics datasets (Hernandez-Ferrer *et al.* 2017).

A key strength of unbiased omics approaches is the ability to identify novel molecules that impact biological function. A major limitation of omics analyses is the ability to annotate unknowns. Results from omics workflows are very generic and some filter out unknowns early in the analytical pipeline. While some workflows confer annotation and functional assignment of unknowns based on coexpression, structural and chemical similarities or abundance, or all of these, this has not been effective mapping the majority of unknowns. Moreover, there is a lack of harmonization, standardization and consensus among the data analysis communities affiliated with individual omics platforms for annotation of unknowns, for instance in case of metabolomics (Spicer *et al.* 2017*a*,*b*). Whereas the genomics and transcriptomics domains have circumvented these issues with vendor-neutral data formats, the MS-based -omics efforts suffer from these challenges. It is noteworthy that such 'gold standard' data sets are being generated and shared for the entire proteomics research community. For example, MS1-based label-free proteomics quadrupole Orbitrap mass spectrometer data for *Escherichia coli* digest spiked into a HeLa digest in four different concentrations is made available, deposited to ProteomeXchange with identifier PXD001385 at PRoteomics IDEntifications database (PRIDE DB) (Shalit *et al.* 2015). A Sigma UPS1 48 protein mix (all equimolar proteins) spiked into a yeast digest background at different concentrations using a Orbitrap Velos platform running High-Low (FT for MS1 and CID ion trap MS/MS scans) is made available as PXD001819 at PRIDE DB (Ramus *et al.* 2016). Similarly, a state-of-the-art data-independent acquisition carried out via SWATH MS Gold Standard data set was made available (Röst *et al.* 2014) to the proteomics research community. Examples of such efforts in lipidomics include harmonization and interoperability of metabolomics standards (Bowden *et al.* 2017) and data sharing, the description, storage and exchange of NMR-based metabolomics efforts (Schober *et al.* 2018). In the absence of robust statistical treatment and measures, investigators are liable to employ 'p-hacking' (investigators select data or statistical analyses until non-significant results become significant). Omics efforts are highly susceptible to such practices owing to the lack of clearly defined 'gold standard' analytical pipelines (Chiu 2017), reiterating the need for appropriate use of statistical tools and publication of analytical pipelines with manuscripts.

Data issues – data archiving and sharing

There is a growing urgency for reproducible research using integrated omics, similar to all disciplines in science. Data archiving is very important for reproducibility of singular omics and integrated omics data, including adherence to Findability, Accessibility, Interoperability and Reusability (FAIR) principles (Wilkinson *et al.* 2016). Part of the solution is a requirement for open sharing of scripts and codes for these analyses (e.g. R, Python, MATLAB, Java) using platforms such as GitHub (https://www.github.com) where developers can share code, review code, manage projects and build software in collaboration with other developers. For instance, cBioPortal for Cancer Genomics (http://cbioportal.org) provides a web resource for exploring, visualizing and analyzing multi-dimensional cancer genomics and clinical data (Gao *et al.* 2013). The Cancer Genome Atlas (TCGA, https://tcgadata.nci.nih.gov/tcga/) has been generating multimodal genomics, epigenomics and proteomics data for thousands of tumor samples from >20 types of cancers (Tomczak *et al.* 2015). The Gene Expression Omnibus (GEO) repository at the National Center for Biotechnology Information (NCBI) archives and freely distributes high-throughput molecular abundance data, predominantly gene expression data (Barrett & Edgar 2006). The most current omics-driven data are archived at OmicsDI (www.omicsdi.org) that houses 149,702 datasets, covering 3926 diseases, 2773 tissues from 6428 species (Perez-Riverol *et al.* 2018).

Thus, although public databases for archiving individual omics datasets exist (Table 1), no such archive exists for integrated omics datasets. Data sharing, especially for large multi-omics studies, can facilitate availability of resources for further exploratory, training and post-publication analyses. To this end, sharing of large datasets using tools like DRYAD (http://datadryad.org; White *et al.* 2008) and Fig Share (https://figshare.com;

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Thelwall & Kousha 2016) are very useful for the research community. Cloud computing technologies may facilitate dataset sharing where a large number of users can easily access and process data from a given dataset and share workflows. Some investigators have begun these efforts for omics datasets (Pavlovich 2017, Warth *et al.* 2017). In addition to routine data sharing, there is also a need for sharing of 'gold standard' datasets from different model systems such as *E. coli*, yeast, *Arabidopsis*, nematode, mouse, non-human primate models, humans, and so forth to clearly define strengths and limitations of each of type of dataset and to provide guidelines on appropriate analyses. In summary, NCBI, SRA, GEO, TCGA in genomics, PRIDE DB, PeptideAtlas repository for proteomics, and MetaboLights, MetabolomicsWorkbench and GNPS for metabolomics have taken center stage for data archiving, although standard databases that allow for submission and retrieval of three or more integrated omic datatypes from a single repository or single interface are lacking.

Hurdles in implementing multi-omics approaches in the clinic for diagnostic/prognostic purposes

Multifactorial and polygenic diseases such as cancer, cardiovascular diseases, neurodegenerative diseases, cardio-metabolic diseases, autoimmune disorders and psychiatric disorders are caused by variation in multiple genes, proteins and metabolites and often influenced by environmental factors such as life-style and diet. The promise of the Human Genome Project was that by identifying all genetic variants in an individual, it would be possible to identify variants that caused complex diseases and provide targets for therapeutic interventions. Based on this premise, the majority of studies focused on identifying biological variation that influences complex disease risk have investigated genetic and epigenetic variation. In addition, these studies typically measure variation in only one omic dataset, e.g. DNA sequence variants, variation in transcript abundance, etc. Despite these research efforts, the promise is largely still unfulfilled. We now know this is in large part due to contributions to health and disease by additional biological variation such as post-translational modifications of proteins and metabolite abundance. Thus, there is the need for not only quantification of different types of biological variants, but integration of these data in ways that inform our understanding health and disease which will translate to clinical practice.

There are examples of single metabolite tests (e.g. glucose, creatinine, bilirubin, lactate and ammonia) routinely performed in the clinic, such as in newborn screening that has established worldwide (Kayton 2007). However, capturing an extensive number of biomolecules for clinical application still presents multiple hurdles ranging from standardized sample collection to current costs per sample. As mentioned previously, sample processing is different for different omic analyses. It is not practical for clinicians to rapidly process samples creating multiple aliquots for different types of analyses. In addition, if limited amounts of patient samples are available, it may not be feasible to perform multiple omic analyses using a single sample. Translation of integrated omics approaches to the clinic will require streamlining processes for sample collection and storage, reducing technical variation, improving reproducibility, standardizing analytical methods, reducing costs and reducing time for sample analyses (Kopczynski *et al.* 2017, Wilson *et al.* 2017). Given that genetic testing is now beginning to be routinely used in the clinic, as well as many examples of small molecule assays, it is only a matter of time before these challenges are addressed for newer and more comprehensive omics platforms to contribute essential clinical data that will provide insights in prognosis and diagnosis of diseases.

Biological knowledge – data interpretation

The largest hurdle for any omics dataset remains 'making sense of the data', which is the 5th 'V' of big data – value. One major objective of multi-dimensional omics approaches is biomarker discovery – no matter from which omics layer the key molecules are derived, sensitivity and specificity of molecular biomarkers are essential for usefulness in biomedical research and clinical translation of findings. Interpretation and curation of complex multi-layered networks is challenging, computationally and time intensive, and requires detailed knowledge of the biological system being studied. Studies using an integrated omics approach without applying biological knowledge of the system frequently end with nomination of key molecules and networks for hypothesis testing that are not biologically relevant. Because validation of key molecules, inclusion of validation cohorts and networks (e.g. genes, proteins and/or metabolites) is time consuming and often challenging, biologically informed nomination of candidates is essential.

Conclusions

Currently, no single approach exists for processing, analyzing and interpreting all data from different -omes.

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055
© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

The need for multimodal data amalgamation strategies and development of reproducible, high throughput, user friendly and effective frameworks must be addressed for this field to advance. Each standard model organism and non-standard model organism poses different challenges due to the uniqueness of metabolite abundance, gene expression bias, epigenetic regulation and cell-type specificity of a given omics dataset. Additionally, with rapid advancement of technologies for genomics, transcriptomics, proteomics and metabolomics, the community needs to embrace challenges posed from these complex datasets to standardize sample quality, sample analysis pipelines, data analysis pipelines and data formats for public data availability. Furthermore, as tools evolve, they must become user friendly, interoperable and effective for computationally intensive analyses. Integrated omics is not just a collage of tools, but a cohesive paradigm for insightful biological interpretation of multi-omics datasets that will potentially reveal novel insights into basic biology, as well as health and disease.

Glossary

Concepts and key terms in this treatise encountered during implementation of integrated omics workflows.

Omics platform terms

Multi-/integrated-/pan-/poly-/trans-/omics: Driven by high-dimensional data generated from >2 –omics technology platforms (usually from multiple types, i.e., genomics, transcriptomics, proteomics, metabolomics) for addressing a biological questions in a seamless manner using bioinformatics and computational workflows and resources. Steps include-sample preparation, -omics data acquisition, raw data preprocessing, filtering and quality control measures, accounting for confounders and analytical challenges: all of it at individual –omics level, followed by their integration.

Systems biology: Uses mathematical modeling for analysis of experimental data to predict the behavior of biological systems, mostly using high-throughput omics technology to quantify the cell functionality using mRNA, proteins and metabolites (but not limited to these) as an *in silico* output using computational models.

NGS: Next generation, massively parallel or deep sequencing encompass modern sequencing efforts that are high throughput, low cost and accurate and are conducted via a single experiment reading millions of nucleotides, as compared to classical Sanger sequencing methods.

Copy number variation: Structural variation in the genome where specific regions of DNA are duplicated, with varying number of repeats.

iTRAQ/TMT tags: Are isobaric peptide labeling methods used in quantitative proteomics using tandem MS for determination of protein abundance in multiple samples in a single experiment.

Untargeted (label free) proteomics/ metabolomics: Unbiased and comprehensive analysis of all measurable analytes (proteins and metabolites) in a given sample including unknowns.

SWATH MS: Is a data-independent acquisition method that complements traditional proteomics experiments by allowing a complete recording of all fragment ions detectable in peptide precursors in a given sample.

Statistical terms

Bayesian: Developed by Thomas Bayes, this statistical logic is applied in decision making and inferential statistics, which deals with probability inference to predict future events based on the knowledge of previous events.

Multivariate: Statistical analysis of data collected on more than one variable that needs to be analyzed simultaneously where dependence is taken into account.

Dimensionality reduction: Commonly used in big data, machine learning and statistical approaches, it allows for reducing the number of dimensions (i.e., the number of random variables under consideration) by providing a set of principal variables.

Principal component analysis: Statistical procedure that summarizes high-dimensional data i.e., constituted of tens of thousands of features where variables are correlated into a lower-dimensional set of uncorrelated variables known as principle components.

WGCNA: To allow screening for genes or modules that are biologically significant, WGCNA defines a gene significance measure, and thus, enables study of biological networks based on pairwise correlations between the variables.

R: Is an open source software environment for statistical computing and graphics that runs on wide variety of operating systems.

Declaration of interest
The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this review.

https://jme.bioscientifica.com
https://doi.org/10.1530/JME-18-0055

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Funding
This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

References

Aebersold R & Mann M 2016 Mass-spectrometric exploration of proteome structure and function. *Nature* **537** 347–355. (https://doi.org/10.1038/nature19949)

Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA & Pe'er D 2010 An integrated approach to uncover drivers of cancer. *Cell* **143** 1005–1017. (https://doi.org/10.1016/j.cell.2010.11.013)

Alexandar V, Nayar PG, Murugesan R, Mary B, Darshana P & Ahmed SS 2015 CardioGenBase: a literature based multi-omics database for major cardiovascular diseases. *PLoS ONE* **10** e0143188. (https://doi.org/10.1371/journal.pone.0143188)

Anand S, Samuel M, Ang CS, Keerthikumar S & Mathivanan S 2017 Label-based and label-free strategies for protein quantitation. In *Proteome Bioinformatics*. Methods in Molecular Biology, vol. **1549**. Eds S Keerthikumar & S Mathivanan. New York, NY: Humana Press. (https://doi.org/10.1007/978-1-4939-6740-7_4)

Arakawa K & Tomita M 2013 Merging multiple omics datasets *in silico*: statistical analyses and data interpretation. *Systems Metabolic Engineering: Methods and Protocols* **985** 459–470. (https://doi.org/10.1007/978-1-62703-299-5_23)

Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, Bernal-Delgado E, Blomberg N Bock C, Conesa A, *et al*. 2016 Making sense of big data in health research: towards an EU action plan. *Genome Medicine* **8** 1. (https://doi.org/10.1186/s13073-016-0323-y)

Aure MR, Steinfeld I, Baumbusch LO, Liestøl K, Lipson D, Nyberg S, Naume B, Sahlberg KK, Kristensen VN, Børresen-Dale AL, *et al*. 2013 Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PLoS ONE* **8** 53014. (https://doi.org/10.1371/journal.pone.0053014)

Bakalarski CE & Kirkpatrick DS 2016 A biologist's field guide to multiplexed quantitative proteomics. *Molecular and Cellular Proteomics* **15** 1489–1497. (https://doi.org/10.1074/mcp.O115.056986)

Bantscheff M, Lemeer S, Savitski MM & Kuster B 2012 Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry* **404** 939–965. (https://doi.org/10.1007/s00216-012-6203-4)

Barrett T & Edgar R 2006 Gene Expression Omnibus: microarray data storage, submission, retrieval, and analysis. *Methods in Enzymology* **411** 352–369. (https://doi.org/10.1016/S0076-6879(06)11019-8)

Bauer C, Stec K, Glintschert A, Gruden K, Schichor C, Or-Guil M, Selbig J & Schuchhardt J 2015 BioMiner: paving the way for personalized medicine. *Cancer Informatics* **14** 55. (https://doi.org/10.4137/CIN.S20910.)

Bayjanov JR, Molenaar D, Tzeneva V, Siezen RJ & van Hijum SA 2012 PhenoLink-a web-tool for linking phenotype to~ omics data for bacteria: application to gene-trait matching for *Lactobacillus plantarum* strains. *BMC Genomics* **13** 1. (https://doi.org/10.1186/1471-2164-13-170)

Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G & Milanesi L 2016 Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* **17** 67. (https://doi.org/10.1186/s12859-015-0857-9)

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K & Wiswedel B 2009 KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explorations Newsletter* **11** 26–31. (http://dx.doi.org/10.1145/1656274.1656280)

Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, Huang SK, Hoon DS, Liu ET, Hillmer A, *et al*. 2015 Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Research* **43** e44. (https://doi.org/10.1093/nar/gku1393.)

Bowden JA, Heckert A, Ulmer CZ, Jones CM, Koelmel JP, Abdullah L, Ahonen L, Alnouti Y, Armando A, Asara JM, *et al*. 2017 Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using standard reference material 1950 metabolites in frozen human plasma. *Journal of Lipid Research* **58** 2275–2288. (https://doi.org/10.1194/jlr.M079012)

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, *et al*. 2017 Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35** 725–731. (https://doi.org/10.1038/nbt.3893)

Box GE & Cox DR 1964 An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* **26** 211–252.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al*. 2001 Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics* **29** 365–371. (https://doi.org/10.1038/ng1201-365)

Bruce SJ, Tavazzi I, Parisod V, Rezzi S, Kochhar S & Guy PA 2009 Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Analytical Chemistry* **81** 3285–3296. (https://doi.org/10.1021/ac8024569)

Buescher JM & Driggers EM 2016 Integration of omics: more than the sum of its parts. *Cancer and Metabolism* **4** 1. (https://doi.org/10.1186/s40170-016-0143-y)

Chae H, Lee S, Seo S, Jung D, Chang H, Nephew KP & Kim S 2016 BioVLAB-mCpG-SNP-EXPRESS: a system for multi-level and multi-perspective analysis and exploration of DNA methylation, sequence variation (SNPs), and gene expression from multi-omics data. *Methods* **111** 64–71. (https://doi.org/10.1016/j.ymeth.2016.07.019)

Chari R, Coe BP, Vucic EA, Lockwood WW & Lam WL 2010 An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology* **4** 67–10. (https://doi.org/10.1186/1752-0509-4-67)

Chen B-J, Causton HC, Mancenido D, Goddard NL, Perlstein EO & Pe'er D 2009 Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular Systems Biology* **5** 310. (https://doi.org/10.1038/msb.2009.69)

Chiu K, Grundy Q & Bero L 2017 'Spin'in published biomedical literature: a methodological systematic review. *PLoS Biology* **15** e2002173. (https://doi.org/10.1371/journal.pbio.2002173)

Chomczynski P & Sacchi N 2006 The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nature Protocols* **1** 581. (https://doi.org/10.1038/nprot.2006.83)

Coman C, Solari FA, Hentschel A, Sickmann A, Zahedi RP & Ahrends R 2016 Simultaneous metabolite, protein, lipid extraction (SIMPLEX): a combinatorial multimolecular omics approach for systems biology. *Molecular and Cellular Proteomics* **15** 1453–1466. (https://doi.org/10.1074/mcp.M115.053702)

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

Cox LA, Glenn JP, Spradling KD, Nijland MJ, Garcia R, Nathanielsz PW & Ford SP 2012 A genome resource to address mechanisms of developmental programming: determination of the fetal sheep heart transcriptome. *Journal of Physiology* **590** 2873–2884. (https://doi.org/10.1113/jphysiol.2011.222398)

Cun Y & Fröhlich H 2014 Netclass: an R-package for network based, integrative biomarker signature discovery. *Bioinformatics* **30** 1325–1326. (https://doi.org/10.1093/bioinformatics/btu025)

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Davidson RL, Weber RJ, Liu H, Sharma-Oates A & Viant MR 2016 Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience* **5** 1. (https://doi.org/10.1186/s13742-016-0115-8)

De Tayrac M, Lê S, Aubry M, Mosser J & Husson F 2009 Simultaneous analysis of distinct Omics datasets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics* **10** 32. (https://doi.org/10.1186/1471-2164-10-32)

Deda O, Chatziioannou AC, Fasoula S, Palachanis D, Raikos N, Theodoridis GA & Gika HG 2017 Sample preparation optimization in fecal metabolic profiling. *Journal of Chromatography B* **1047** 115–123. (https://doi.org/10.1016/j.jchromb.2016.06.047)

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN & Aebersold R 2006 The PeptideAtlas project. *Nucleic Acids Research* **34** D655–D658. (https://doi.org/10.1093/nar/gkj040)

Deutsch EW, Orchard S, Binz PA, Bittremieux W, Eisenacher M, Hermjakob H, Kawano S, Lam H, Mayer G, Menschaert G, *et al*. 2017 Proteomics standards initiative: fifteen years of progress and future work. *Journal of Proteome Research* **16** 4288–4298. (https://doi.org/10.1021/acs.jproteome.7b00370)

Erickson BK, Rose CM, Braun CR, Erickson AR, Knott J, McAlister GC, Wühr M, Paulo JA, Everley RA & Gygi SP 2017 A strategy to combine sample multiplexing with targeted proteomics assays for high-throughput protein signature characterization. *Molecular Cell* **65** 361–370. (https://doi.org/10.1016/j.molcel.2016.12.005)

Fan J, Saha S, Barker G, Heesom KJ, Ghali F, Jones AR, Matthews DA & Bessant C 2015 Galaxy Integrated Omics: web-based standards-compliant workflows for proteomics informed by transcriptomics. *Molecular and Cellular Proteomics* **14** 3087–3093. (https://doi.org/10.1074/mcp.O115.048777)

Fiehn O, Robertson D, Griffin J, van der Werf M, Nikolau B, Morrison N, Sumner LW, Goodacre R, Hardy NW, Taylor C, *et al*. 2007 The metabolomics standards initiative (MSI). *Metabolomics* **3** 175–178. (https://doi.org/10.1007/s11306-007-0070-6)

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, *et al*. 2008 The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology* **26** 541–547. (https://doi.org/10.1038/nbt1360)

Fisch KM, Meißner T, Gioia L, Ducom JC, Carland TM, Loguercio S & Su AI 2015 Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics* **31** 1724–1728. (https://doi.org/10.1093/bioinformatics/btv061)

Fondi M & Liò P 2015 Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiological Research* **171** 52–64. (https://doi.org/10.1016/j.micres.2015.01.003)

Forsberg EM, Huan T, Rinehart D, Benton HP, Warth B, Hilmers B & Siuzdak G 2018 Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature Protocols* **13** 633. (https://doi.org/10.1038/nprot.2017.151)

Fukushima A & Kusano M 2013 Recent progress in the development of metabolome databases for plant systems biology. *Frontiers in Plant Science* **4** 73. (https://doi.org/10.3389/fpls.2013.00073)

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, *et al*. 2013 Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling* **6** pl1. (https://doi.org/10.1126/scisignal.2004088)

García-Alcalde F, García-López F, Dopazo J & Conesa A 2011 Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics* **27** 137–139. (https://doi.org/10.1093/bioinformatics/btq594)

Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, *et al*. 2010 Visualization of omics data for systems biology. *Nature Methods* **7** S56–S68. (https://doi.org/10.1038/nmeth.1436)

Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, *et al*. 2010 myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research* **38** W677–W682. (https://doi.org/10.1093/nar/gkq429)

Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A & Tegnér J 2014 Data integration in the era of omics: current and future challenges. *BMC Systems Biology* **8** 1. (https://doi.org/10.1186/1752-0509-8-S2-I1)

Guhlin J, Silverstein KA, Zhou P, Tiffin P & Young ND 2017 ODG: Omics database generator-a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. *BMC Bioinformatics* **18** 367. (https://doi.org/10.1186/s12859-017-1777-7)

Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, *et al*. 2016 Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology* **2** 16180. (https://doi.org/10.1038/nmicrobiol.2016.180)

Henry VJ, Bandrowski AE, Pepin AS, Gonzalez BJ & Desfeux A 2014 OMICtools: an informative directory for multi-omic data analysis. *Database* **2014** bau069. (https://doi.org/10.1093/database/bau069)

Hernandez-Ferrer C, Ruiz-Arenas C, Beltran-Gomila A & González JR 2017 MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. *BMC Bioinformatics* **18** 36. (https://doi.org/10.1186/s12859-016-1455-1)

Huang S, Chaudhary K & Garmire LX 2017 More is better: recent progress in multi-omics data integration methods. *Frontiers in Genetics* **8** 84. (https://doi.org/10.3389/fgene.2017.00084)

Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, Forman JJ, Troyanskaya OG & Coller HA 2009 Detailing regulatory networks through large scale data integration. *Bioinformatics* **25** 3267–3274. (https://doi.org/10.1093/bioinformatics/btp588)

Jang Y, Yu N, Seo J, Kim S & Lee S 2016 MONGKIE: an integrated tool for network analysis and visualization for multi-omics data. *Biology Direct* **11** 10. (https://doi.org/10.1186/s13062-016-0112-y)

Jiang Y, Shi X, Zhao Q, Krauthammer M, Rothberg BEG & Ma S 2016 Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics* **107** 223–230. (https://doi.org/10.1016/j.ygeno.2016.04.005)

Jung D, Kim B, Freishtat RJ, Giri M, Hoffman E & Seo J 2015 miRTarVis: an interactive visual analysis tool for microRNA-mRNA expression profile data. *BMC Proceedings* **9** 1. (https://doi.org/10.1186/1753-6561-9-S6-S2)

Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, Nainala VC, Spicer RA, Williams M, Li X, *et al*. 2016 MetaboLights: an open-access database repository for metabolomics data. *Current Protocols in Bioinformatics* **53** 1–18. (https://doi.org/10.1002/0471250953.bi1413s53)

Kamburov A, Cavill R, Ebbels TM, Herwig R & Keun HC 2011 Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **27** 2917–2918. (https://doi.org/10.1093/bioinformatics/btr499)

Kamoun A, Idbaih A, Dehais C, Elarouci N, Carpentier C, Letouzé E, Colin C, Mokhtari K, Jouvet A, Uro-Coste E, *et al*. 2016 Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. *Nature Communications* **7** 11263. (https://doi.org/10.1038/ncomms11263)

Kato H, Takahashi S & Saito K 2011 Omics and integrated omics for the promotion of food and nutrition science. *Journal of Traditional and Complementary Medicine* **1** 25–30. (https://doi.org/10.1016/S2225-4110(16)30053-0)

Kayton A 2007 Newborn screening: a literature review. *Neonatal Network* **26** 85–95. (https://doi.org/10.1891/0730-0832.26.2.85)

Khatri P, Sirota M & Butte AJ 2012 Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology* **8** e1002375. (https://doi.org/10.1371/journal.pcbi.1002375)

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Kim HK & Verpoorte R 2010 Sample preparation for plant metabolomics. *Phytochemical Analysis* **21** 4–13. (https://doi.org/10.1002/pca.1188)

Kim M & Tagkopoulos I 2018 Data integration and predictive modeling methods for multi-omics datasets. *Molecular Omics* **14** 8–25.(https://doi.org/10.1039/C7MO00051K)

Kim M, Rai N, Zorraquino V & Tagkopoulos I 2016 Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nature Communications* **7** 13090. (https://doi.org/10.1038/ncomms13090)

Kirk P, Griffin JE, Savage RS, Ghahramani Z & Wild DL 2012 Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28** 3290–3297. (https://doi.org/10.1093/bioinformatics/bts595)

Klein J, Jupp S, Moulos P, Fernandez M, Buffin-Meyer B, Casemayou A, Chaaya R, Charonis A, Bascands JL, Stevens R, *et al*. 2012 The KUPKB: a novel Web application to access multiomics data on kidney disease. *FASEB Journal* **26** 2145–2153. (https://doi.org/10.1096/fj.11-19438)

Kohl M, Megger DA, Trippler M, Meckel H, Ahrens M, Bracht T, Weber F, Hoffmann AC, Baba HA, Sitek B, *et al*. 2014 A practical data processing workflow for multi-OMICS projects. *Biochimica et Biophysica Acta (BBA): Proteins and Proteomics* **1844** 52–62. (https://doi.org/10.1016/j.bbapap.2013.02.029)

Kopczynski D, Coman C, Zahedi RP, Lorenz K, Sickmann A & Ahrends R 2017 Multi-OMICS: a critical technical perspective on integrative lipidomics approaches. *Biochimica et Biophysica Acta (BBA): Molecular and Cell Biology of Lipids* **1862** 808–811. (https://doi.org/10.1016/j.bbalip.2017.02.003)

Krämer A, Green J, Pollard J & Tugendreich S 2014 Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30** 523–530. (https://doi.org/10.1093/bioinformatics/btt703)

Krishnan KC, Kurt Z, Barrere-Cain R, Sabir S, Das A, Floyd R, Vergnes L, Zhao Y, Che N, Charugundla S, *et al*. 2018 Integration of multi-omics data from mouse diversity panel highlights mitochondrial dysfunction in non-alcoholic fatty liver disease. *Cell Systems* **6** 103–115. (https://doi.org/10.1016/j.cels.2017.12.006)

Kuo TC, Tian TF & Tseng YJ 2013 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology* **7** 64. (https://doi.org/10.1186/1752-0509-7-64)

Langfelder P & Horvath S 2008 WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9** 559. (https://doi.org/10.1186/1471-2105-9-559)

Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, *et al*. 2013 Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501** 506–511. (https://doi.org/10.1038/nature12531)

Lê Cao KA, González I & Déjean S 2009 integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **25** 2855–2856. (https://doi.org/10.1093/bioinformatics/btp515)

Leday GG & van de Wiel MA 2013 PLRS: a flexible tool for the joint analysis of DNA copy number and mRNA expression data. *Bioinformatics* **29** 1081–1082. (https://doi.org/10.1093/bioinformatics/btt082)

Li W, Zhang S, Liu C-C & Zhou XJ 2012 Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* **28** 2458–2466. (https://doi.org/10.1093/bioinformatics/bts476)

Li Y, Wu FX & Ngom A 2016 A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics* **19** 325–340. (https://doi.org/10.1093/bib/bbw113)

Li H, Wang X, Rukina D, Huang Q, Lin T, Sorrentino V, Zhang H, Sleiman MB, Arends D, McDaid A, *et al*. 2017 An integrated systems genetics and omics toolkit to probe gene function. *Cell Systems* **6** 90–102. (https://doi.org/10.1016/j.cels.2017.10.016)

Libbrecht MW & Noble WS 2015 Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16** 321–332. (https://doi.org/10.1038/nrg3920)

Lin D, Zhang J, Li J, Xu C, Deng HW & Wang YP 2016 An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics* **17** 247. (https://doi.org/10.1186/s12859-016-1122-6)

Liu Y, Devescovi V, Chen S & Nardini C 2013 Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Systems Biology* **7** 14. (https://doi.org/10.1186/1752-0509-7-14)

Liu G, Dong C & Liu L 2016 Integrated multiple '-omics' data reveal subtypes of hepatocellular carcinoma. *PLoS ONE* **11** e0165457. (https://doi.org/10.1371/journal.pone.0165457)

Louhimo R & Hautaniemi S 2011 Cnamet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* **27** 887–888. (https://doi.org/10.1093/bioinformatics/btr019)

Meng C, Kuster B, Culhane AC & Gholami AM 2014 A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15** 162. (https://doi.org/10.1186/1471-2105-15-162)

Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM & Culhane AC 2016 Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* **17** 628–641. (https://doi.org/10.1093/bib/bbv108)

Menichetti G, Remondini D, Panzarasa P, Mondragón RJ & Bianconi G 2014 Weighted multiplex networks. *PLoS ONE* **9** e97857. (https://doi.org/10.1371/journal.pone.0097857)

Menon V 2017 Clustering single cells: a review of approaches on high- and low-depth single-cell RNA-seq data. *Briefings in Functional Genomics*. (https://doi.org/10.1093/bfgp/elx044)

Merelli I, Lió P & Milanesi L 2013 Nuchart: an R package to study gene spatial neighbourhoods with multi-omics annotations. *PLoS ONE* **8** 75146. (https://doi.org/10.1371/journal.pone.0075146)

Min S, Lee B & Yoon S 2016 Deep learning in bioinformatics. *Briefings in Bioinformatics* **18** 851–869. (https://doi.org/10.1093/bib/bbw068)

Misra BB & van der Hooft JJ 2016 Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* **37** 86–110. (https://doi.org/10.1002/elps.201500417)

Misra BB 2018 New tools and resources in metabolomics: 2016–2017. *Electrophoresis* **39** 909–923. (https://doi.org/10.1002/elps.201700441)

Misra BB, Fahrmann JF & Grapov D 2017 Review of emerging metabolomic tools and resources: 2015–2016. *Electrophoresis* **38** 2257–2274. (https://doi.org/10.1002/elps.201700110)

Mochida K & Shinozaki K 2011 Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant and Cell Physiology* **52** 2017–2038. (https://doi.org/10.1093/pcp/pcr153)

Montague E, Stanberry L, Higdon R, Janko I, Lee E, Anderson N, Choiniere J, Stewart E, Yandl G, Broomall W, *et al*. 2014 MOPED 2.5. An integrated multi-omics resource: multi-omics profiling expression database now includes transcriptomics data. *OMICS: A Journal of Integrative Biology* **18** 335–343. (https://doi.org/10.1089/omi.2014.0061)

Mosca E & Milanesi L 2013 Network-based analysis of omics with multi-objective optimization. *Molecular Biosystems* **9** 2971–2980. (https://doi.org/10.1039/c3mb70327d)

Mukherjee S & Speed TP 2008 Network inference using informative priors. *PNAS* **105** 14313–14318. (https://doi.org/10.1073/pnas.0802272105)

Muller EE, Pinel N, Laczny CC, Hoopmann MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, *et al*. 2014 Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Communications* **5** 5603. (https://doi.org/10.1038/ncomms6603)

Muqaku B, Eisinger M, Meier SM, Tahir A, Pukrop T, Haferkamp S, Slany A, Reichle A & Gerner C 2017 Multi-omics analysis of serum samples demonstrates reprogramming of organ functions via systemic calcium mobilization and platelet activation in metastatic

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

melanoma. *Molecular and Cellular Proteomics* **16** 86–99. (https://doi.org/10.1074/mcp.M116.063313)

Nakayasu ES, Nicora CD, Sims AC, Burnum-Johnson KE, Kim YM, Kyle JE, Matzke MM, Shukla AK, Chu RK, Schepmoes AA, *et al*. 2016 MPLEx: a robust and universal protocol for single-sample integrative proteomic, metabolomic, and lipidomic analyses. *MSystems* **1** e00043-16. (https://doi.org/10.1128/mSystems.00043-16)

Nibbe RK, Koyutürk M & Chance MR 2010 An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Computational Biology* **6** e1000639. (https://doi.org/10.1371/journal.pcbi.1000639)

Oveland E, Muth T, Rapp E, Martens L, Berven FS & Barsnes H 2015 Viewing the proteome: how to visualize proteomics data? *Proteomics* **15** 1341–1355. (https://doi.org/10.1002/pmic.201400412)

Palermo G, Piraino P & Zucht HD 2009 Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and Applications in Bioinformatics and Chemistry* **2** 57. (https://doi.org/10.2147/AABC.S3619)

Parkhomenko E, Tritchler D & Beyene J 2009 Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8** 1–34. (https://doi.org/10.2202/1544-6115.1406)

Pathak RR & Dave V 2014 Integrating omics technologies to study pulmonary physiology and pathology at the systems level. *Cellular Physiology and Biochemistry* **33** 1239–1260. (https://doi.org/10.1159/000358693)

Pavel AB, Sonkin D & Reddy A 2016 Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. *BMC Systems Biology* **10** 1. (https://doi.org/10.1186/s12918-016-0260-9)

Pavlopoulos GA, Malliarakis D, Papanikolaou N, Theodosiou T, Enright AJ & Iliopoulos I 2015 Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience* **4** 1–27. (https://doi.org/10.1186/s13742-015-0077-2)

Pavlovich M 2017 Computing in biotechnology: omics and beyond. *Trends in biotechnology* **35** pp.479–480. (https://doi.org/10.1016/j.tibtech.2017.03.011)

Perez-Riverol Y, Zorin A, Dass G, Glont M, Vizcaino JA, Jarnuczak A, Petryszak R, Ping P & Hermjakob H 2018 Quantifying the impact of public omics data. *BioRxiv* 282517. (https://doi.org/10.1101/282517)

Proffitt JM, Glenn J, Cesnik AJ, Jadhav A, Shortreed MR, Smith LM, Kavanagh K, Cox LA & Olivier M 2017 Proteomics in non-human primates: utilizing RNA-Seq data to improve protein identification by mass spectrometry in vervet monkeys. *BMC Genomics* **18** 877. (https://doi.org/10.1186/s12864-017-4279-0)

Pruitt KD, Tatusova T & Maglott DR 2006 NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35** D61–D65. (https://doi.org/10.1093/nar/gkl842)

Quinn RA, Navas-Molina JA, Hyde ER, Song SJ, Vázquez-Baeza Y, Humphrey G, Gaffney J, Minich JJ, Melnik AV, Herschend J, *et al*. 2016 From sample to multi-omics conclusions in under 48 hours. *mSystems* **1** e00038-16. (https://doi.org/10.1128/mSystems.00038-16)

Rajasundaram D & Selbig J 2016 More effort – more results: recent advances in integrative 'omics' data analysis. *Current Opinion in Plant Biology* **30** 57–61. (https://doi.org/10.1016/j.pbi.2015.12.010)

Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, Chan T, Chapman P, Davis SR, Gomez-Cabrero D, *et al*. 2017 software for the integration of multiomics experiments in bioconductor. *Cancer Research* **77** e39–e42. (https://doi.org/10.1158/0008-5472.CAN-17-0344)

Ramus C, Hovasse A, Marcellin M, Hesse AM, Mouton-Barbosa E, Bouyssié D, Vaca S, Carapito C, Chaoui K, Bruley C, *et al*. 2016 Benchmarking quantitative label-free LC–MS data processing workflows using a complex spiked proteomic standard dataset. *Journal of Proteomics* **132** 51–62. (https://doi.org/10.1016/j.jprot.2015.11.011)

Ritchie MD, Holzinger ER, Li R, Pendergrass SA & Kim D 2015 Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics* **16** 85–97. (https://doi.org/10.1038/nrg3868)

Rohart F, Gautier B, Singh A & Lê Cao K-A 2017 mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* **13** e1005752. (https://doi.org/10.1371/journal.pcbi.1005752)

Ronan T, Qi Z & Naegle KM 2016 Avoiding common pitfalls when clustering biological data. *Science Signaling* **9** re6. (https://doi.org/10.1126/scisignal.aad1932)

Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, Wolski W, Collins BC, Malmström J, Malmström L, *et al*. 2014 OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* **32** 219–223. (https://doi.org/10.1038/nbt.2841)

Sarpe V & Schriemer DC 2017 Supporting metabolomics with adaptable software: design architectures for the end-user. *Current Opinion in Biotechnology* **43** 110–117. (https://doi.org/10.1016/j.copbio.2016.11.001)

Savage RS, Ghahramani Z, Griffin JE, de la Cruz BJ & Wild DL 2010 Discovering transcriptional modules by Bayesian data integration. *Bioinformatics* **26** 158–167. (https://doi.org/10.1093/bioinformatics/btq210)

Scheubert K, Hufsky F, Petras D, Wang M, Nothias LF, Dührkop K, Bandeira N, Dorrestein PC & Böcker S 2017 Significance estimation for large scale metabolomics annotations by spectral matching. *Nature Communications* **8** 1494. (https://doi.org/10.1038/s41467-017-01318-5)

Schober D, Jacob D, Wilson M, Cruz JA, Marcu A, Grant JR, Moing A, Deborde C, de Figueiredo LF, Haug K, *et al*. 2018 nmrML: a community supported open data standard for the description, storage, and exchange of NMR data. *Analytical Chemistry* **90** 649–656. (https://doi.org/10.1021/acs.analchem.7b02795)

Shalit T, Elinger D, Savidor A, Gabashvili A & Levin Y 2015 MS1-based label-free proteomics using a quadrupole Orbitrap mass spectrometer. *Journal of Proteome Research* **14** 1979–1986. (https://doi.org/10.1021/pr501045)

Shen R, Olshen AB & Ladanyi M 2009 Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912. (https://doi.org/10.1093/bioinformatics/btp543)

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA & Waterston RH 2017 DNA sequencing at 40: past, present and future. *Nature* **550** 345–353. (https://doi.org/10.1038/nature24286)

Shu L, Zhao Y, Kurt Z, Byars S, Tukiainen T, Kettunen J, Ripatti S, Zhang B, Inouye M, Makinen VP, *et al*. 2016 Mergeomics: integration of diverse genomics resources to identify pathogenic perturbations to biological systems. *BMC Genomics* **17** 874. (https://doi.org/10.1186/s12864-016-3198-9)

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, *et al*. 2018 WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46** D661–D667 (https://doi.org/10.1093/nar/gkx1064)

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, *et al*. 2007 The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **25** 1251–1255. (https://doi.org/10.1038/nbt1346)

Song WM & Zhang B 2015 Multiscale embedded gene co-expression network analysis. *PLoS Computational Biology* **11** e1004574. (https://doi.org/10.1371/journal.pcbi.1004574)

Spicer RA, Salek R & Steinbeck C 2017*a* Comment: a decade after the metabolomics standards initiative it's time for a revision. *Scientific Data* **4** 170138. (https://doi.org/10.1038/sdata.2017.138)

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Spicer RA, Salek R & Steinbeck C 2017*b* Compliance with minimum information guidelines in public metabolomics repositories. *Scientific Data* **4** 170137. (https://doi.org/10.1038/sdata.2017.137)

Srivastava V, Obudulu O, Bygdell J, Löfstedt T, Rydén P, Nilsson R, Ahnlund M, Johansson A, Jonsson P, Freyhult E, *et al*. 2013 OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI-superoxide dismutase *Populus* plants. *BMC Genomics* **14** 1. (https://doi.org/10.1186/1471-2164-14-893)

Stanberry L, Mias GI, Haynes W, Higdon R, Snyder M & Kolker E 2013 Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites* **3** 741–3760. (https://doi.org/10.3390/metabo3030741)

Stöckel D, Kehl T, Trampert P, Schneider L, Backes C, Ludwig N, Gerasch A, Kaufmann M, Gessler M, Graf N, *et al*. 2016 Multi-omics enrichment analysis using the GeneTrail2 journal service. *Bioinformatics* **32** 1502–1508. (https://doi.org/10.1093/bioinformatics/btv770)

Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, *et al*. 2015 Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research* **44** D463–D470. (https://doi.org/10.1093/nar/gkv1042)

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL *et al*. 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3** 211–221. (https://doi.org/10.1007/s11306-007-0082-2)

Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, *et al*. 2007 The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* **25** 887–893. (https://doi.org/10.1038/nbt1329)

Teo CC, Chong WP, Tan E, Basri NB, Low ZJ & Ho YS 2015 Advances in sample preparation and analytical techniques for lipidomics study of clinical samples. *Trends in Analytical Chemistry* **66** 1–8. (https://doi.org/10.1016/j.trac.2014.10.010)

Thaiss CA, Levy M, Korem T, Dohnalová L, Shapiro H, Jaitin DA, David E, Winter DR, Gury-BenAri M, Tatirovsky E, *et al*. 2016 Microbiota diurnal rhythmicity programs host transcriptome oscillations. *Cell* **167** 1495–1510. (https://doi.org/10.1016/j.cell.2016.11.003)

Thelwall M & Kousha K 2016 Figshare: a universal repository for academic resource sharing?. *Online Information Review* **40** 333–346. (https://doi.org/10.1108/OIR-06-2015-0190)

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY & Stitt M 2004 Mapman: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant Journal* **37** 914–939. (https://doi.org/10.1111/j.1365-313X.2004.02016.x)

Tibshirani R 2011 Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 273–282. (https://doi.org/10.1111/j.1467-9868.2011.00771.x)

Tisoncik-Go J, Gasper DJ, Kyle JE, Eisfeld AJ, Selinger C, Hatta M, Morrison J, Korth MJ, Zink EM, Kim YM, *et al*. 2016 Integrated omics analysis of pathogenic host responses during pandemic H1N1 influenza virus infection: the crucial role of lipid metabolism. *Cell Host and Microbe* **19** 254–266. (https://doi.org/10.1016/j.chom.2016.01.002)

Tomczak K, Czerwińska P & Wiznerowicz M 2015 The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology* **19** A68.

Tuncbag N, McCallum S, Huang S-SC & Fraenkel E 2012 Steinernet: a journal server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic Acids Research* **40** 505–509. (https://doi.org/10.1093/nar/gks445)

Tuncbag N, Gosline SJ, Kedaigle A, Soltis AR, Gitter A & Fraenkel E 2016 Network-based interpretation of diverse high-throughput datasets through the Omics Integrator software package. *PLoS Computational Biology* **12** e1004879. (https://doi.org/10.1371/journal.pcbi.1004879)

Valledor L, Escandón M, Meijón M, Nukarinen E, Cañal MJ & Weckwerth W 2014 A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. *Plant Journal* **79** 173–180. (https://doi.org/10.1111/tpj.12546)

van Dijk EL, Jaszczyszyn Y & Thermes C 2014 Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research* **322** 12–20. (https://doi.org/10.1016/j.yexcr.2014.01.008)

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D & Stuart JM 2010 Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* **26** 237–245. (https://doi.org/10.1093/bioinformatics/btq182)

Villas-Bôas SG, Højer-Pedersen J, Åkesson M, Smedsgaard J & Nielsen J 2005 Global metabolite analysis of yeast: evaluation of sample preparation methods. *Yeast* **22** 1155–1169.

Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, Carbone L, Steemers FJ & Adey A 2017 Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods* **14** 302. (https://doi.org/10.1038/nmeth.4154)

Vizcaíno JA, Côté RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, *et al*. 2013 The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*. **41** D1063–D1069. (https://doi.org/10.1093/nar/gks1262)

Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, *et al*. 2014 ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32** 223–226. (https://doi.org/10.1093/bioinformatics/btq182)

Wang D & Gu J 2016 Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology* **4** 58–67. (https://doi.org/10.1007/s40484-016-0063-4)

Wang J, Zuo Y, Man YG, Avital I, Stojadinovic A, Liu M, Yang X, Varghese RS, Tadesse MG & Ressom HW 2015 Pathway and network approaches for identification of cancer signature markers from omics data. *Journal of Cancer* **6** 54–65. (https://doi.org/10.7150/jca.10631)

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, *et al*. 2016 Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34** 828–837. (https://doi.org/10.1038/nbt.3597)

Wanichthanarak K, Fahrmann JF & Grapov D 2015 Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights* **10** 1–6. (https://doi.org/10.4137/BMI.S29511)

Warth B, Levin N, Rinehart D, Teijaro J, Benton HP & Siuzdak G 2017 Metabolizing data in the cloud. *Trends in Biotechnology* **35** 481–483. (https://doi.org/10.1016/j.tibtech.2016.12.010)

Weill N, Lisi V, Scott N, Dallaire P, Pelloux J & Major F 2015 MiRBooking simulates the stoichiometric mode of action of microRNAs. *Nucleic Acids Research* **43** 6730–6738. (https://doi.org/10.1093/nar/gkv619)

White HC, Carrier S, Thompson A, Greenberg J & Scherle R 2008 The Dryad Data Repository: a Singapore framework metadata architecture in a DSpace Environment. In *Dublin Core Conference*, pp 157–162.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, *et al*. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3** 160018. (https://doi.org/10.1038/sdata.2016.18)

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access

Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, Houten SM, Amariuta T, Wolski W, Zamboni N, *et al*. 2016 Systems proteomics of liver mitochondria function. *Science* **352** 6291. (https://doi.org/10.1126/science.aad0189)

Wilson J, Turna NS, Banks R, Pappin DJ & Zougman A 2017 Clinical applications of universal S-Trap sample processing. *Molecular and Cellular Proteomics* **16** S56–S56.

Wiśniewski JR, Zougman A, Nagaraj N & Mann M 2009 Universal sample preparation method for proteome analysis. *Nature Methods* **6** 359. (https://doi.org/10.1038/nmeth.1322)

Xu D, Keller JM, Popescu M & Bondugula R 2008 *Applications of Fuzzy Logic in Bioinformatics*, vol. **9**. World Scientific.

Yang JY, Karr JR, Watrous JD & Dorrestein PC 2011 Integrating '-omics' and natural product discovery platforms to investigate metabolic exchange in microbiomes. *Current Opinion in Chemical Biology* **15** 79–87. (https://doi.org/10.1016/j.cbpa.2010.10.025)

Yoo S, Huang T, Campbell JD, Lee E, Tu Z, Geraci MW, Powell CA, Schadt EE, Spira A & Zhu J 2014 MODMatcher: multi-omics data matcher for integrative genomic analysis. *PLoS Computational Biology* **10** e1003790. (https://doi.org/10.1371/journal.pcbi.1003790)

Yuan Y, Savage RS & Markowetz F 2011 Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Computational Biology* **7** 1002227. (https://doi.org/10.1371/journal.pcbi.1002227)

Yugi K, Kubota H, Hatano A & Kuroda S 2016 Trans-Omics: how to reconstruct biochemical networks across multiple 'omic' layers. *Trends in Biotechnology* **34** 276–290. (https://doi.org/10.1016/j.tibtech.2015.12.013)

Yuryev A, Kotelnikova E & Daraselia N 2009 Ariadne's ChemEffect and Pathway Studio knowledge base. *Expert Opinion on Drug Discovery* **4** 1307–1318. (https://doi.org/10.1517/17460440903413488)

Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, Aparicio S & Hansen CL 2017 Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods* **14** 167. (https://doi.org/10.1038/nmeth.4140)

Zhang F, Xiao X, Hao J, Wang S, Wen Y & Guo X 2015 CPAS: a trans-omics pathway analysis tool for jointly analyzing DNA copy number variations and mRNA expression profiles data. *Journal of Biomedical Informatics* **53** 363–366. (https://doi.org/10.1016/j.jbi.2014.12.012)

Zierer J, Pallister T, Tsai PC, Krumsiek J, Bell JT, Lauc G, Spector TD, Menni C & Kastenmüller G 2016 Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model. *Scientific Reports* **6** 37646. (https://doi.org/10.1038/srep37646)

Zou H & Hastie T 2005 Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320. (http://www.jstor.org/stable/3647580)

Received in final form 2 July 2018
Accepted 12 July 2018
Accepted Preprint published online 12 July 2018

© 2019 Society for Endocrinology
Published by Bioscientifica Ltd.
Printed in Great Britain

Downloaded from Bioscientifica.com at 03/05/2020 08:09:09PM
via free access