Methods of integrating data to uncover genotype–phenotype interactions

Marylyn D. Ritchie[1], Emily R. Holzinger[2], Ruowang Li[1], Sarah A. Pendergrass[1] and Dokyoon Kim[1]

Abstract | Recent technological advances have expanded the breadth of available omic data, from whole-genome sequencing data, to extensive transcriptomic, methylomic and metabolomic data. A key goal of analyses of these data is the identification of effective models that predict phenotypic traits and outcomes, elucidating important biomarkers and generating important insights into the genetic underpinnings of the heritability of complex traits. There is still a need for powerful and advanced analysis strategies to fully harness the utility of these comprehensive high-throughput data, identifying true associations and reducing the number of false associations. In this Review, we explore the emerging approaches for data integration — including meta-dimensional and multi-staged analyses — which aim to deepen our understanding of the role of genetics and genomics in complex outcomes. With the use and further development of these approaches, an improved understanding of the relationship between genomic variation and human phenotypes may be revealed.

Complex traits
Characteristics that arise from interactions among multiple molecular factors, with the potential influence of environmental and behavioural factors. Complex traits do not conform to the inheritance pattern of Mendelian traits.

[1]*Department of Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.*
[2]*National Human Genome Research Institute, Inherited Disease Research Branch, Baltimore, Maryland 21224, USA.*
Correspondence to M.D.R.
e-mail: marylyn.ritchie@psu.edu
doi:10.1038/nrg3868
Published online 13 January 2015

Understanding the genetic basis of complex traits has been an ongoing quest for many researchers. Technological advances in data generation from multiple levels of biological systems — including DNA sequence data[1], RNA expression levels[2,3], methylation patterns[4], other epigenetic markers[5], proteomics[6] and metabolomics[7] (FIG. 1) — have driven the field of translational bioinformatics for the past decade, producing ever-increasing amounts of data as researchers strive to develop complementary analysis tools. In addition to generating these data from whole-blood or specific tissue samples, the ability to generate these data from single cells is rapidly advancing[8].

Various analytical approaches have been developed to identify the genetic variation that underlies complex traits. For example, DNA sequence variation can be identified through linkage analysis in family-based data[9] and through association studies in family-[10] and population-based data[11]. In addition, the association between phenotypic outcome and variation in other high-throughput omic measurements — such as gene expression (using microarrays and RNA sequencing (RNA-seq)), epigenetic variation (by methylation arrays, methylation sequencing or chromatin immunoprecipitation followed by sequencing (ChIP–seq)) and protein variation (assayed in either metabolomic or proteomic studies in various ways) — is now routinely explored. Historically, each type of data has been considered independently to look for relationships with biological processes and, using these methods, we have assembled some of the pieces of the puzzle of complex-trait genetic architecture and basic biological pathways. However, much of the genetic aetiology of complex traits and biological networks remains unexplained, which could be partly due to the focus on restrictive single-data-type study designs.

Owing to our limited understanding of many complex traits from this single-data-type approach, meta-dimensional analysis and multi-staged analysis (that is, systems genomics approaches) have been used increasingly. As reviewed previously[12–16], a systems genomics approach can achieve a more thorough and informative interrogation of genotype–phenotype associations than an analysis that uses only a single data type. Combining multiple data types can compensate for missing or unreliable information in any single data type, and multiple sources of evidence pointing to the same gene or pathway are less likely to lead to false positives. Importantly, the complete biological model is only likely to be discovered if the different levels of genetic, genomic and proteomic regulation are considered in an analysis.

Meta-dimensional analysis
An approach whereby all scales of data are combined simultaneously to produce complex models defined as multiple variables from multiple scales of data.

Multi-staged analysis
A stepwise or hierarchical analysis method that reduces the search space through different stages of analysis.

Systems genomics
An analysis approach that models the complex inter- and intra-individual variations of traits and diseases using data from next-generation omic data.

Data integration
The incorporation of multi-omic information in a meaningful way to provide a more comprehensive analysis of a biological point of interest.

In this Review, we describe the principles of meta-dimensional analysis and multi-staged analysis, and provide an overview of some of the approaches that are used to predict a given quantitative or categorical outcome, the tools available to implement these analyses, and the various strengths and weaknesses of these strategies. In addition, we describe the analytical challenges that emerge with data sets of this magnitude, and provide our perspective on how such systems genomic analyses might develop in the future.

Why integrate data?

Data integration can have numerous meanings; however, in this Review, we use it to mean the process by which different types of omic data are combined as predictor variables to allow more thorough and comprehensive modelling of complex traits or phenotypes — which are likely to be the result of an elaborate interplay among biological variation at various levels of regulation — through the identification of more informative models. Data integration methods are now emerging that aim to bridge the gap between our ability to generate vast amounts of data and our understanding of biology, thus reflecting the complexity within biological systems. The primary motivation behind integrated data analysis is to identify key genomic factors, and importantly their interactions, that explain or predict disease risk or other biological outcomes. The success in understanding the genetic and genomic architecture of complex phenotypes has been modest, and this could be due to our limited exploration of the interactions among the genome, transcriptome, metabolome and so on. Data integration may provide improved power to identify the important genomic factors and their interactions (BOX 1). In addition, modelling the complexity of, and the interactions between, variation in DNA, gene expression, methylation, metabolites and proteins may improve our understanding of the mechanism or causal relationships of complex-trait architecture. There are two main approaches to data integration: multi-staged analysis, which involves integrating information using a stepwise or hierarchical analysis approach; and meta-dimensional analysis, which refers to the concept of integrating multiple different data types to build a multivariate model associated with a given outcome[16–18].

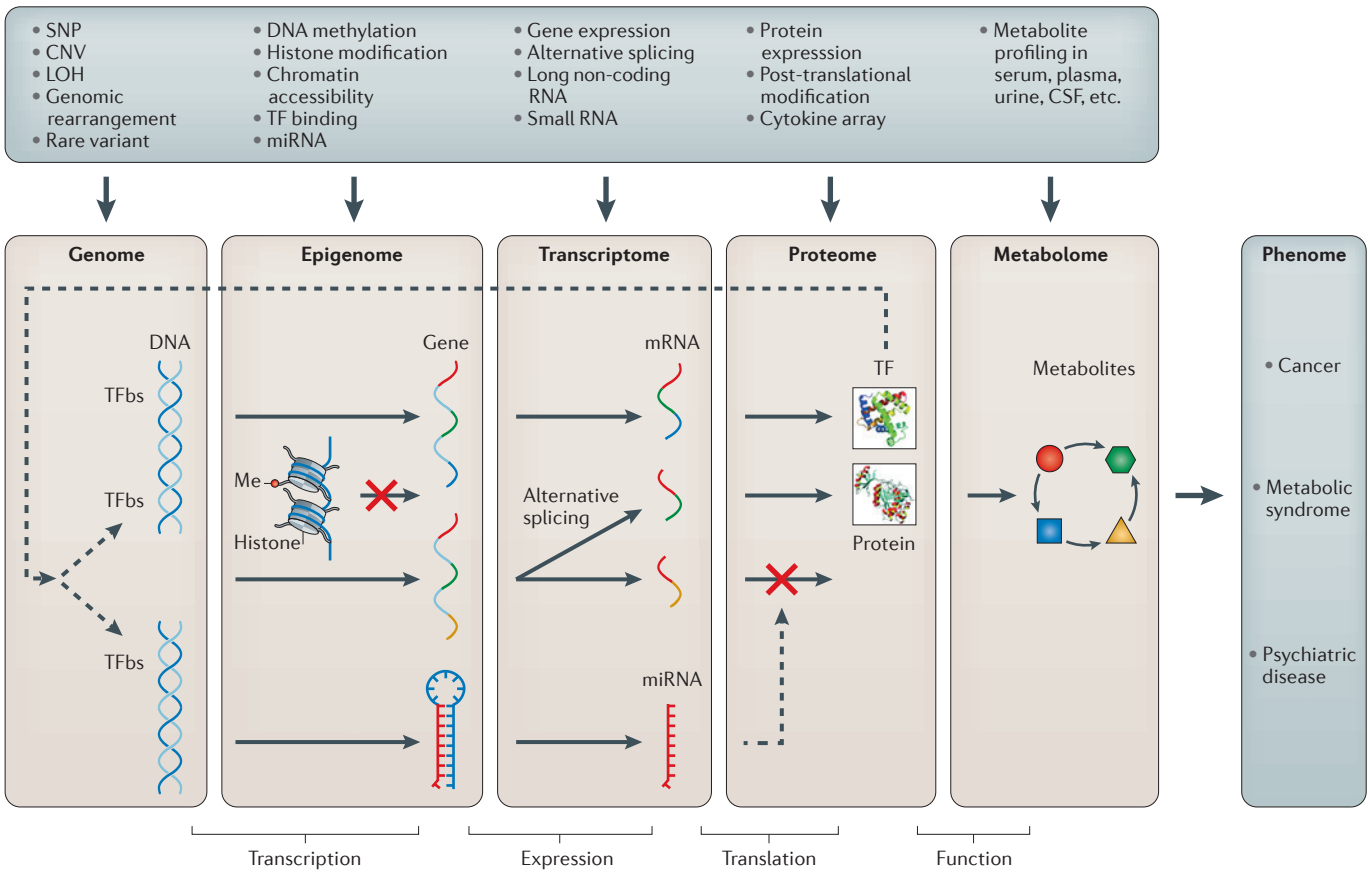


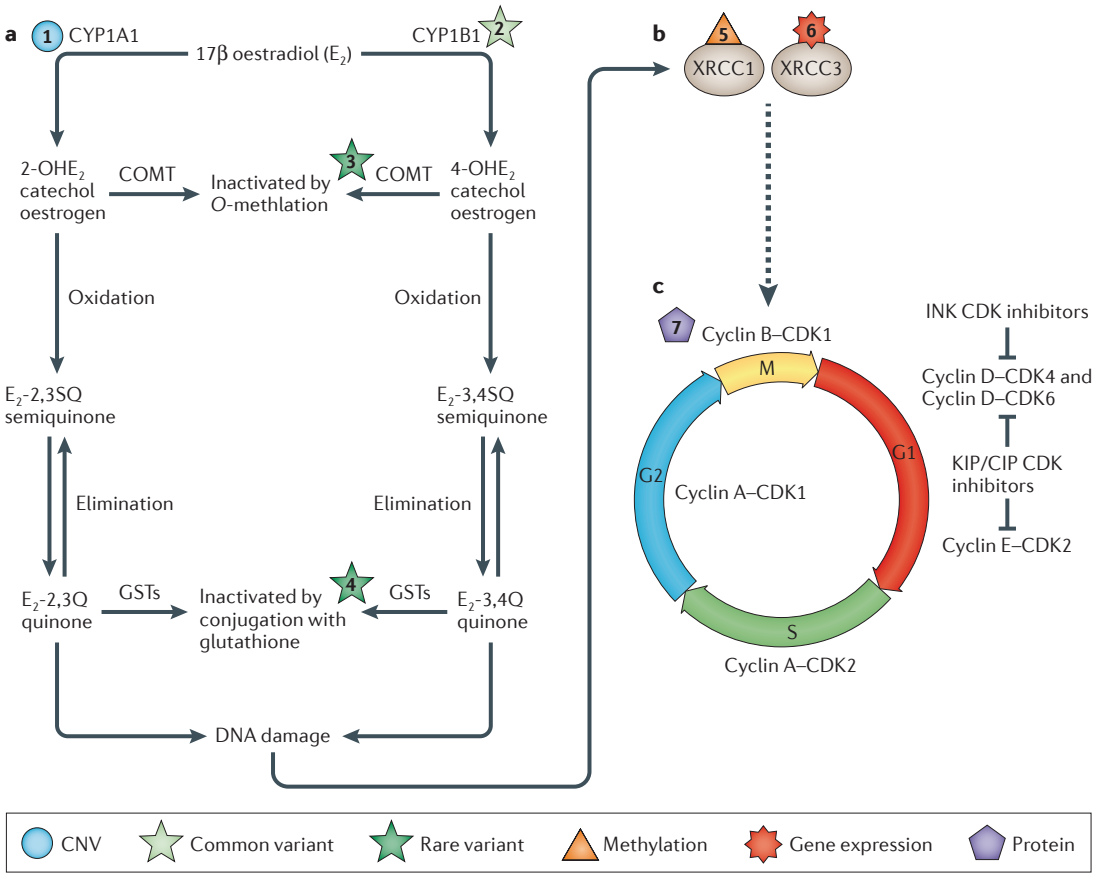
Figure 1 | **Biological systems multi-omics from the genome, epigenome, transcriptome, proteome and metabolome to the phenome.** Heterogeneous genomic data exist within and between levels, for example, single-nucleotide polymorphism (SNP), copy number variation (CNV), loss of heterozygosity (LOH) and genomic rearrangement, such as translocation, at the genome level; DNA methylation, histone modification, chromatin accessibility, transcription factor (TF) binding and micro RNA (miRNA) at the epigenome level; gene expression and alternative splicing at the transcriptome level; protein expression and post-translational modification at the proteome level; and metabolite profiling at the metabolome level. Arrows indicate the flow of genetic information from the genome level to the metabolome level and, ultimately, to the phenome level. The red crosses indicate inactivation of transcription or translation. CSF, cerebrospinal fluid; Me, methylation; TFBS, transcription factor-binding site.

Box 1 | **The reality of complex outcomes and the importance of data integration**

Assessing variation only of a single omic data type can miss complex models that require variation across multiple levels of biological regulation. Data integration approaches can provide a key to making sense of greater complexity. Here, we present a hypothetical example, using sporadic breast cancer as a complex trait, of the complexity underlying disease that should be encompassed when analysing data, and illustrate the utility of meta-dimensional analysis. Many candidate gene studies[85,86], genome-wide association studies[87–90] and pathway analyses[91,92] have been pursued, yet only a relatively small proportion of the estimated heritability of sporadic breast cancer has been explained. Consider the following example involving three well-studied pathways for breast cancer: oestrogen metabolism, DNA damage repair and the cell cycle (see the figure; points of variation are indicated by a shape, and the numbers correspond to the description in the text).

It is now established that oestrogen can cause DNA damage if it is not properly metabolized[93]. Two genes, cytochrome P450 1A1 (*CYP1A1*) and *CYP1B1*, participate in the first step of oestrogen breakdown. The metabolite created by CYP1B1 (4-OHE_2 catechol oestrogen) creates a more carcinogenic form of oestrogen by-product than that metabolized by CYP1A1. In our model, copy number variation (CNV) in *CYP1A1* (label 1 in the figure) reduces activity, and single-nucleotide polymorphisms (SNPs) in *CYP1B1* (label 2) increase activity, resulting in higher levels of carcinogenic by-products (see the figure, part **a**). Additionally, multiple rare variants in caffeic acid 3-*O*-methyltransferase (*COMT*; label 3), glutathione *S*-transferase μ1 (*GSTM1*) and glutathione *S*-transferase θ1 (*GSTT1*; label 4) reduce the metabolism of carcinogenic by-products, resulting in a higher level of DNA damage (see the figure, part **a**). Even so, these variations may not increase the risk of cancer if the DNA damage repair pathway can offset the increase in carcinogenic metabolites. However, hypermethylation of X-ray repair cross-complementing 1 (*XRCC1*; label 5) and variation in the gene expression of *XRCC3* (label 6) result in reduced transcription levels, and this repair pathway may no longer be able to adequately keep DNA repair at necessary levels (see the figure, part **b**). Finally, dysregulated protein expression of genes in the cell cycle pathway — for example, in cyclin-dependent kinase 1 (CDK1; label 7) — may result in a rate of cell replication that is higher than average and therefore DNA damage (see the figure, part **c**). The end result can lead to an abundance of damaged cells (that is, breast cancer cells). In our hypothetical model, all of the variation mentioned above is required to pass the threshold into cancer development. Therefore, only an analytical approach that integrates data from the genome, transcriptome and proteome would identify the full model.

This purely hypothetical example is used to illustrate the following point: an analysis that assesses variation of only a single omic data type can miss complex models that require variation across multiple levels of biological regulation. Disease aetiology is probably the result of the complex interplay between multiple biological pathways; thus, when exploring only one piece of data (that is, only SNPs), we cannot understand the full complexity of biological systems and outcomes. We believe meta-dimensional analysis is critical for fully realizing the scope of trait architecture.



Data integration will allow us to explore new scientific questions, although assembling all of these data types together into a more complete biological story is immensely challenging. In particular, diversity in the size of data sets, patterns of missing data and noise across the different data types, and correspondence and correlation between measurements from different technologies can create substantial challenges. Many methods to integrate data have been developed, all with strengths and weaknesses, and no single analysis approach will be optimal for all studies. Therefore, a comprehensive and expanded analysis 'toolbox' will be an important factor in the discovery and interpretation of the complexities of biology.

Challenges with individual data sets

There are unique challenges for individual data types, and it is important to consider these before implementing multi-staged or meta-dimensional analyses; these include data quality, data scale or dimensionality, and potential confounding of the data (see below). If these issues are not dealt with for each individual data type, then they could cause problems when the data types are integrated. Evaluating each data type carefully before integration is important to avoid downstream problems with the analysis. Additionally, rapid advances in data generation require substantial increases in computational power and storage capabilities of computing systems[19]. Numerous approaches and strategies, from open-source to commercial packages, are being explored to store and track these data[20–22].

Quality assurance and quality control. In the past, for more limited collections of data, data quality could be assessed at the level of individual data points. For example, genotyping using low-throughput assays such as TaqMan would be evaluated by the laboratory through assessing the genotype clusters of homozygous and heterozygous samples for each single-nucleotide polymorphism (SNP) genotyped to determine whether the SNP was of high quality and whether there are any samples that did not cluster well with the rest of the data set. However, with the large-scale nature of high-throughput data, examining data individually is not feasible, and researchers often rely on summary statistics and broad overviews of the data. For example, several quality control pipelines have been established and implemented for genome-wide SNP data, such as the Electronic Medical Records and Genomics (eMERGE) and Gene Environment Association Studies (GENEVA) networks[23–25]. Similarly, DNA sequencing[26], RNA-seq[27] and genome-wide methylation profiling[28] approaches have specific and critical quality control steps that must be implemented before analysis. These include looking for quality of the individual genomic variables, sample integrity and distributional evaluations of the genomic variables or samples, with respect to variables in a clinical or phenotypic data set. These evaluations, which need to be performed for SNP or DNA sequence data separately from RNA-seq data or metabolomic data, will ensure that high-quality data are integrated. The phrase 'garbage in, garbage out' comes to mind when deciding how rigorously to perform quality control checks before data integration. To ensure high-quality results, the goal is to start with high-quality data.

Data reduction. Data reduction can be used to limit the number of variables evaluated in a single data set, but it can also be used as an initial step for performing analyses across multiple data types. For example, when considering data with a vast number of independent variables and a substantially smaller number of samples, statistical power can be very limited[29]. Several analytical strategies, such as cross-validation and permutation testing[30], can be implemented to address this concern, although investigators usually attempt to perform some form of data reduction before their association, correlation or modelling analysis. Reducing the amount of data through some type of filtering strategy (see below) facilitates data integration analyses on a smaller, more refined subset of the data. This can lead to more efficient computations and can potentially reduce the multiple-hypothesis testing burden. Furthermore, when exploring millions of measurements within a single data set, especially for complex models that include interactions, some level of data reduction is often necessary for analysis of single data types, as well as for integrative analyses across multiple data types. For example, if exploring models with more than one variable (such as models of SNP–SNP interactions or models incorporating multiple gene expression variables for outcome prediction), the full dimensionality of the data often cannot be modelled owing to computation time, memory and sample size considerations. Consider an example in which the data set includes 5 million SNPs. Calculating statistics for combinations of SNPs in an exhaustive manner leads to a combinatorial increase in models and their respective computation times. If we construct all possible pairwise models (by choosing 2 of the 5 million variables at a time and constructing all possible 2-variable models and then repeating with different variables until all combinations have been assessed (referred to as '5 million choose 2')), we have 1.25×10^{13} pairwise models to be evaluated; this number rises dramatically as the number of variables in the model increases. For example, calculating all of the statistical models that include 3 variables results in 2.09×10^{19} models to test. At a computation rate of 1 million models per second, it would take more than 3,400 hours to perform all of the 2-variable models in this example, and more than 5.7×10^{9} hours to perform all of the 3-variable models. Even with large GPU (graphics processing unit) clusters, which are considerably faster than traditional computing processors, these computation times are reaching the limits of practicality.

Data reduction through filtering and data mining can be either extrinsic (that is, using information external to the data set itself) or intrinsic (that is, using the data set and some analytical technique for filtering). Extrinsic approaches use prior knowledge that is accessible in the public domain, such as from Biofilter[31], whereas intrinsic approaches use methods such as ReliefF[32,33], chi-square statistics, principal

Quality control
Various techniques used to remove noise and confounding factors from the data.

component analysis (PCA)[30,34], factor analysis[30] and genetic algorithms[35]. An example of intrinsic filtering for SNPs would be to use linkage disequilibrium (LD) patterns to reduce the number of SNPs, thus focusing on a smaller number of variables. An example of extrinsic filtering would be to filter a gene expression data set only for gene expression from genes known to have a relationship to the immune system when investigating an autoimmune trait. A limitation of extrinsic filtering is that we only 'know what we know', and extrinsic data reduction is therefore limited by the knowledge of the field at the time that information is used to guide data reduction. However, intrinsic filtering might remove biologically important features. In some analyses, a combination of intrinsic and extrinsic filtering can be performed. For example, with 5 million SNPs, a researcher may perform a chi-square test of association for each SNP and keep only those that are significant based on a chosen *P* value threshold, and may also select all biologically relevant variants based on a Biofilter annotation.

The nature of the chosen method for data reduction will drive the hypothesis that can be tested. For example, there are two primary molecular variability hypotheses that might explain a resulting complex trait (FIG. 2). The dominant paradigm has been that variation at the DNA level will lead to changes in gene expression, leading to changes in protein expression and finally a change in phenotype — a fundamentally linear assumption of disease aetiology (Hypothesis A). If Hypothesis A is considered, then stratifying the data by type or scale and performing initial analyses before moving on to a step of further data integration is the most powerful, easily implemented and interpretable approach. For example, this would involve first reducing the amount of SNP data to include only those SNPs associated with a particular outcome, then limiting the amount of proteomic data to only those proteomic variables associated with the outcome, before analysing the SNP and proteomic data together. Hypothesis B is the alternative possibility, in which multiple levels of molecular variation contribute to disease risk in a nonlinear, interactive and complex way. If Hypothesis B is considered, then stratifying by data type for data reduction and subsequently performing analyses would inhibit the ability to detect the true model; thus, an alternative data reduction approach that combines the multi-omics data sets prior to data reduction would be more appropriate. For example, data from copy number variation, methylation and micro RNA (miRNA) could be combined and then reduced via ReliefF[32]; the resultant filtered data set could then be analysed for association with a particular outcome or phenotype.

Confounding. Confounding is another challenge with data integration (as with some other genomic and proteomic analyses) that can lead to spurious associations and interpretations of findings. Confounding occurs when an independent variable is associated both with another independent variable and with the dependent variable; it can occur because of genetic, environmental, demographic or other technical factors. For example, population stratification is a type of confounding that can occur in genetic association studies[36]. Several methods have been developed to address population stratification, including mixed-model approaches[37] and PCA[38]. Surrogate variable analysis has been introduced as a strategy to accurately capture the relationship between variation in molecular variables (such as gene expression) and variation in other variables of interest, and to overcome the potential issues with heterogeneity and confounding[39]. Evidence of confounding needs to be addressed prior to any comprehensive data integration analyses.

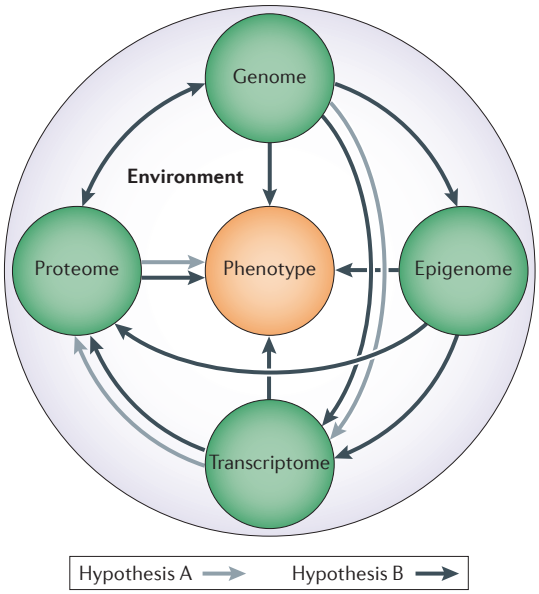


Figure 2 | **Alternative hypothesis of complex-trait aetiology.** Hypothesis A (grey arrow) is the theory that variation is hierarchical, such that variation in DNA leads to variation in RNA and so on in a linear manner. Hypothesis B (black arrow) is the idea that it is the combination of variation across all possible omic levels in concert that leads to phenotype.

An overview of data integration

Data integration methods can be broadly categorized into two types of approaches. In multi-staged analysis, models are constructed using only two different scales at a time, in a stepwise, linear or hierarchical manner. By scale, we refer to the numerical and categorical features of the data, for example, SNP variables, and gene expression variables that have either continuous values for the level of expression or a categorical variable indicating overexpressed or underexpressed genes. This approach reflects Hypothesis A of FIG. 2. Meta-dimensional analysis, or the fusion of scales, is an approach in which all scales of data are combined simultaneously to identify complex, meta-dimensional models with multiple variables from different data types. This approach reflects Hypothesis B of FIG. 2. There are several types of analysis and software tools that can be used to implement both multi-staged analysis and meta-dimensional analysis (TABLE 1).

Factor analysis
A statistical method used to describe variability among observed, correlated variables in terms of a smaller number of unobserved (latent) variables.

Multi-omics data
Multiple types of genome-scale data sets that emerged from high-throughput technologies, including genome sequencing data (genomics), genome-wide RNA-sequencing data (transcriptomics), methylation and histone modification data (epigenomics), and mass spectrometry protein data (proteomics).

Population stratification
A situation in which different subpopulations exist within a data set owing to different allele frequencies because of underlying genetic ancestry that leads to different strata being present in the data set. This can lead to spurious associations if not adjusted for appropriately.

Table 1 | **Categorization of data analysis methods**

Approach	Methods	Software and/or tools	Refs
Multi-staged analysis			
Genomic variation analysis	eQTL, mQTL and causal analysis	Matrix eQTL[94] and QTDT[95]	43,94,95
	Allele-specific expression	AlleleSeq[96] and ChIP–SNP[48]	48,96
Domain knowledge-guided analysis	Correlation and mapping variation to pathway	ANNOVAR[97], HaploReg[98] and RegulomeDB[99]	97–100
Meta-dimensional analysis			
Concatenation-based integration	Grammatical evolution neural network	ATHENA	18,56,57
	Bayesian network	WinBUGS	54
	Multivariate Cox LASSO model	Glmpath	55
Transformation-based integration	Kernel-based integration	SKMsmo	59,60
	Graph-based semi-supervised learning	Graph-based semi-supervised learning	53,61,62
Model-based integration	Majority voting	ipred	64,65
	Ensemble classifier	Weka 3	66

ATHENA, Analysis Tool for Heritable and Environmental Network Associations; ChIP, chromatin immunoprecipitation; eQTL, expression quantitative trait locus; LASSO, least absolute shrinkage and selection operator; mQTL, methylation QTL; QTDT, quantitative trait linkage disequilibrium test; SNP, single-nucleotide polymorphism.

Data integration: multi-staged analysis

Multi-staged analysis, as its name suggests, aims to divide data analysis into multiple steps, and signals are enriched with each step of the analysis. The main objective of the multi-staged approach is to divide the analysis into multiple steps to find associations first between the different data types, then subsequently between the data types and the trait or phenotype of interest. Examples of multi-staged analyses are shown in FIG. 3 and described below.

Genomic variation analysis approaches. The most commonly used genomic variation integration technique so far has been a three-stage or triangle method[16]. In the triangle method, the following steps are taken.

1. SNPs are associated with the phenotype and filtered based on a genome-wide significance threshold.
2. SNPs deemed significant from step 1 are then tested for association with another level of omic data. For example, one option is to look for the association of SNPs with gene expression levels. These SNPs are called expression quantitative trait loci (eQTLs). Alternatively, methylation QTLs (mQTLs; which are SNPs associated with DNA methylation levels), metabolite QTLs (which are SNPs associated with metabolite levels) and protein QTLs (pQTLs; which are SNPs associated with protein levels or other molecular traits such as long non-coding RNA and miRNA) could be used.
3. Omic data used in step 2 are then tested for correlation with the phenotype of interest.

Different methods of analysis can be used to implement this triangle approach, including linear or logistic regression (depending on a continuous or a binary dependent variable, respectively). The rationale of this approach is based on Hypothesis A of FIG. 2, in which genetic variations are the foundation of all other molecular variations. The triangle approach has been used, for example, in studies of chemotherapeutic drug response in HapMap cell lines, in which significant eQTLs were tested for correlation with the drug response[40–42]. The difficulty of triangle-based methods comes when a relatively arbitrary threshold, generally a *P* value, is used to identify the significant associations for further analyses. As the *P* value threshold also needs to be adjusted for the number of tests being carried out to combat multiple testing problems, there is likely to be a large number of false-negative SNPs, eQTLs, mQTLs and pQTLs being filtered out. This approach is often used to find SNPs associated with both a gene expression trait or a methylation level and the phenotype of interest to focus on functional SNPs.

Some researchers have begun to develop causal inference association approaches. For example, Schadt *et al.* have introduced a multistep approach to identify key drivers of complex traits that exploit the naturally occurring DNA variation observed in populations[43]. DNA variation is tested for association with gene expression, and gene expression traits are then ordered relative to one another. Analyses then determine whether DNA variants that lead to variation in relative transcript abundances are supported statistically as an independent, causative or reactive function[43] using maximum likelihood approaches. These causal approaches[43,44] allow the dissection of the genotype-to-phenotype process in a clear, linear manner. As long as Hypothesis A of FIG. 2 is being tested, these approaches are fairly powerful.

Allele-specific expression approaches. Another approach that links genomic variations to transcript levels is called allele-specific expression (ASE). In diploid organisms, one of the two alleles is preferentially expressed in some genes[45]. ASE variants are associated

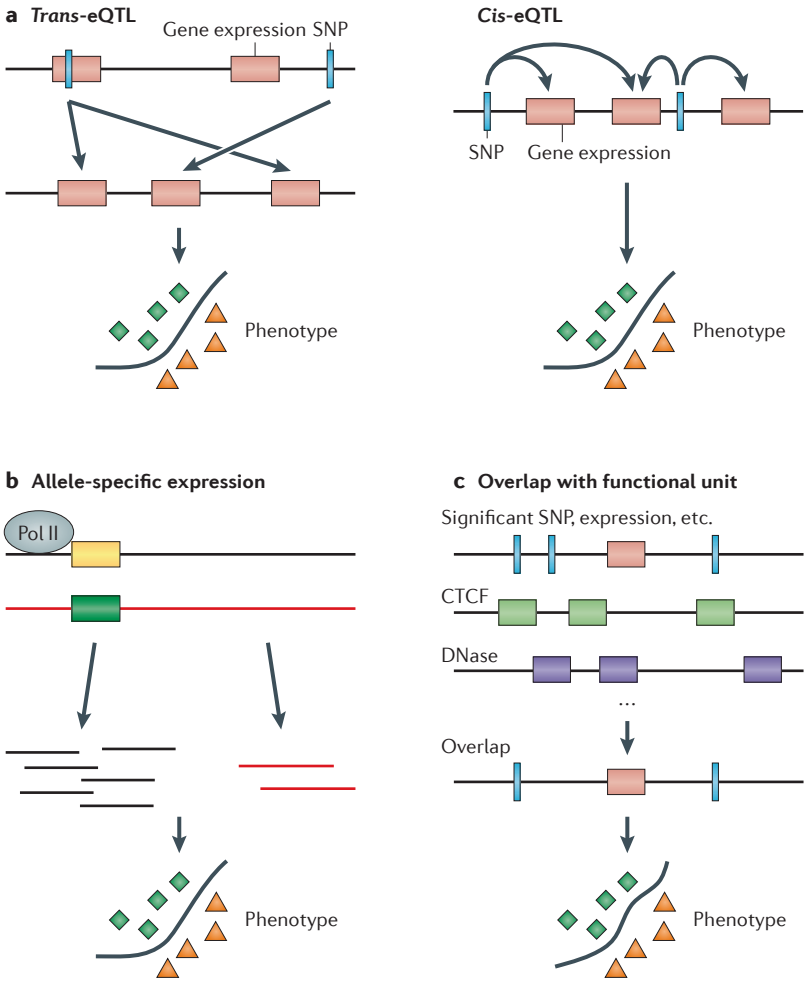


Figure 3 | **Categorization of multi-staged analysis.** Multi-staged analysis can be divided into three categories. **a** | Analysis of expression quantitative trait loci (eQTLs) analysis involves the identification of genetic variation associated with measures of quantitative gene expression. **b** | Allele-specific expression involves the analysis of whether the maternal or paternal allele is preferentially expressed, followed by the association of this allele with *cis*-element variations and epigenetic modifications. **c** | Domain knowledge overlap involves a two-step analysis in which an initial association analysis is performed at the single-nucleotide polymorphism (SNP) or gene expression variable followed by the annotation of the significant associations with knowledge generated by other biological experiments. This approach enables the selection of association results with functional data to corroborate the association. CTCF, CCCTC-binding factor; Pol II, RNA polymerase II.

with *cis*-element variations and epigenetic modifications[46]. The first step of ASE approaches is to distinguish the gene product of one parental allele from the product of the other parental allele. Next, an analysis to associate the allele with gene expression (eQTLs) or methylation (mQTLs) can be carried out to compare the two alleles. Finally, the resulting alleles can be tested for correlation with a phenotype or an outcome of interest. The practicality of this approach depends on the extra resources used for experimentally tagging the two alleles and the subsequent mapping of the alleles. ASE and other extended methods — such as allele-specific transcript structure (ASTS), which looks at the frequency of expression of splice transcripts that are allele-specific — have been used to identify functional variation[47] and protein–DNA[48] interactions in humans. This allele-specific approach has also been used in other contexts. For example, several groups have explored allele-specific analysis in chromatin state[49] and histone modification[50]. More allele-specific applications are likely to emerge as we continue to observe these allele-specific effects.

Domain knowledge-guided approaches. Other studies have integrated functional and pathway information that is generated and consolidated by initiatives such as the Encyclopedia of DNA Elements (ENCODE)[51] and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[52] to select and annotate significant results. In this approach, the genomic regions of interest are inputs. Various software and databases can be used to determine whether the regions are within pathways and/or overlapping with functional units, such as transcription factor binding, hypermethylated or hypomethylated regions, DNase sensitivity and regulatory motifs. For example, a researcher may take a collection of genotyped SNPs and annotate them with domain knowledge from multiple public database resources. The subsequent list of SNPs that have functional annotations can then be taken into the next stage, during which they are associated with other omic data, such as gene expression data (from microarray or RNA-seq) or metabolomic data. The resulting SNPs that have functional annotations and that are associated with other omic data can then be evaluated for correlation with a phenotype or an outcome of interest. This approach can be similar to the triangle approach mentioned above, with the exception that there is another step of annotating the variants and only taking those with functional annotations to the next stage of analysis. Adding information from diverse data sets can substantially increase our knowledge of our data; however, we are also limited and biased by current knowledge.

Even though multi-staged analysis uses both linear and nonlinear analytical mathematics to understand the relationship between two different types of data, there are clear limitations. For example, if complex traits are the result of a combination of DNA sequence variants, gene expression variability, methylation states and protein structure or expression changes that occurs simultaneously along with environmental perturbations (FIG. 2, Hypothesis B) rather than in a stepwise linear model (FIG. 2, Hypothesis A), the multi-staged approach will fail to effectively model the complex trait. However, when the relationship between genotype and phenotype can be modelled in a linear manner, as is the case for SNPs associated with metabolites and subsequently associated with phenotypes, for example, a multi-staged analysis would be applicable.

Data integration: meta-dimensional analysis

Meta-dimensional analysis combines multiple data types in a simultaneous analysis[16,17,53] and is broadly categorized into three approaches: concatenation-based integration, transformation-based integration and model-based integration (FIG. 4).

Multivariate Cox LASSO (least absolute shrinkage and selection operator) model
A method that performs variable selection via LASSO, followed by a multivariate Cox regression analysis.

Concatenation-based integration. Concatenation-based integration combines multiple data matrices for each sample into one large input matrix before constructing a model. One advantage of concatenation-based integration is that, after it is determined how to combine the variables into one matrix, it is relatively easy to use any statistical method for continuous and categorical data for analysis. For example, Fridley *et al.*[54] performed concatenation-based integration by incorporating multiple types of genomic data into an association analysis with a complex phenotype using a Bayesian modelling strategy. Data from SNPs and mRNA gene expression were combined into a single data matrix, and the joint relationship of mRNA gene expression and SNP genotypes was then modelled using a Bayesian integrative model to predict a quantitative phenotype (for example, drug cytotoxicity). Mankoo *et al.*[55] predicted time to recurrence and survival in ovarian cancer using copy number alteration, methylation, miRNA and gene expression data using a multivariate Cox LASSO (least absolute shrinkage and selection operator) model. This strategy involves performing variable selection via LASSO, rather than a stepwise method, and then modelling the selected set of variables in a Cox regression. The other main advantage of this approach is that concatenation-based integration is particularly useful for considering interactions between different types of genomic data. For example, if the underlying model that one is trying to detect is a SNP interacting with metabolite to explain disease risk and if the two variables are not combined into one model, then the effect may be missed. This approach has been used to combine SNP and gene expression data to predict high-density lipoprotein cholesterol levels[18,56], and to identify interactions between copy number alteration, methylation, miRNA and gene expression data associated with cancer clinical outcomes[57].

The challenge with concatenation-based integration is identifying the best approach for combining multiple matrices that include data from different scales in a meaningful way. For example, SNP data contain 0, 1 or 2 as values corresponding to the copies of a specific allele per individual; copy number data may consist of −2, −1, 0, 1 or 2 as values corresponding to copy number status in a given genetic region (although they can also

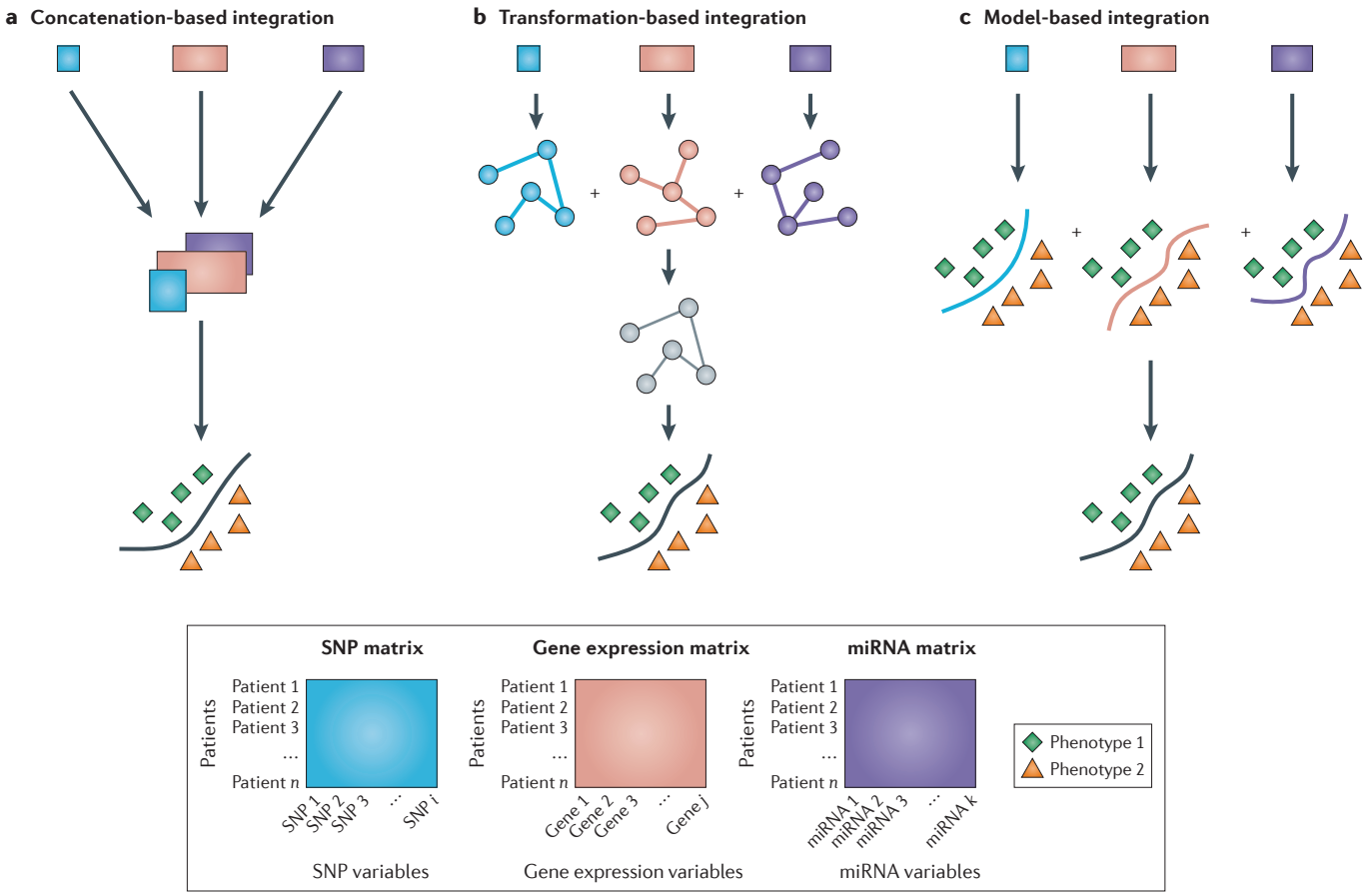


Figure 4 | **Categorization of meta-dimensional analysis.** Meta-dimensional analysis can be divided into three categories. **a** | Concatenation-based integration involves combining data sets from different data types at the raw or processed data level before modelling and analysis. **b** | Transformation-based integration involves performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices. **c** | Model-based integration is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest. miRNA, microRNA; SNP, single-nucleotide polymorphism.

be continuous-scale data); and DNA methylation profiles report between 0 and 1 for CpG loci. Identifying a way to appropriately integrate or combine these data without biases driven by data type can be challenging. Furthermore, this form of data integration can inflate high-dimensionality for the data, with the number of samples being smaller than the number of measurements for each sample[58]. Thus, concatenation-based integration is only suitable if the appropriate way to assemble the data matrix for analysis is determined. Subsequently, statistical or computational models can be used to analyse the data matrix to consider interactions between different types of genomic data. Data reduction strategies as described above may be needed, depending on the number of variables in the data matrix. If there are too many variables, the analysis may not be computationally feasible; therefore, performing data reduction to limit the number of variables would be required to make this analysis possible.

Transformation-based integration. The second approach, transformation-based integration, combines multiple data sets after transforming each data type into an intermediate form, such as a graph or a kernel matrix (a symmetrical and positive semi-definite matrix that represents the relative positions of all samples conducted by valid kernel functions). Multiple graphs or kernels can then be merged before elaborating any models (FIG. 4). The transformation-based integration approach has the advantage of preserving data-type-specific properties from each data set when each type of data is transformed into an appropriate intermediate representation. In addition, this approach can be used to integrate many types of data, including continuous or categorical values and sequence data, as long as the data contain a unifying feature, such as patient identifiers linking data types. Moreover, the transformation-based integration approach is robust to different data measurement scales.

For example, Lanckriet *et al.*[59] proposed kernel-based integration for protein function prediction with multiple types of heterogeneous data, including amino acid sequences, hydropathy profiles, gene expression data and known protein–protein interactions, and Borgwardt *et al.*[60] combined structural, sequential and chemical information into one graph model for predicting protein function via graph kernels. By contrast, Tsuda *et al.*[61] and Shin *et al.*[62] predicted protein function with multiple networks using graph-based semi-supervised learning. Kim *et al.*[53] proposed a graph-based integration framework for predicting cancer clinical outcomes using copy number alteration, methylation, miRNA and gene expression data. The disadvantage of transformation-based integration is that identifying interactions between different types of data (such as a SNP and gene expression interaction) can be difficult if the separate transformation of the original feature space changes the ability to detect the interaction effect. Each data type is transformed independently, which can make it more difficult to detect some effects. The goal is to perform a data transformation that maintains the majority of the data-type-specific properties so that these types of interaction effects are not missed. Thus, transformation-based integration is suitable if there is a relevant intermediate representation, such as a kernel or graph, for each genomic data type, and the goal is to preserve data-type-specific properties while integrating them.

Model-based integration. Model-based integration, the third meta-dimensional approach, encompasses methods in which multiple models are generated using the different types of data as training sets, and a final model is then generated from the multiple models created during the training phase, preserving data-specific properties. This approach can combine predictive models from different types of data. For example, model-based integration may allow the integration of data sets in which each data type is collected from a different set of patients but all patients have the same disease or phenotype. If the goal is to identify genetic, genomic and proteomic associations with ovarian cancer, data sets could be extracted from the public domain, where DNA sequence data may be available on five sets of patient samples, microarray data on eight sets of patient samples, and proteomic data on two sets of patient samples. Model-based integration would allow the independent analysis of each of the 15 data sets, followed by an integration of the top models from each data set to look for integrative models. This is an area of future work for the Analysis Tool for Heritable and Environmental Network Associations (ATHENA) methodology[56,57,63]. ATHENA is a suite of analysis tools for performing systems genomic analyses to integrate different omic data and look for association with clinical outcomes. Model-based integration has been performed with ATHENA to look for associations between copy number alterations, methylation, microRNA and gene expression with ovarian cancer survival[57]. A neural network model was constructed for each data type (such as copy number aberration and methylation) separately, and the four resulting models were then analysed to create an integrative model. As another example, a majority voting approach was used to predict drug resistance of HIV protease mutants[64] using structural features of the HIV protease–drug inhibitor complex and DNA sequence variants. In most cases, the variables from the top models are combined in a subsequent analysis. In addition, ensemble classifiers — such as predicted secondary structure, hydrophobicity, van der Waals volume, polarity, polarizability and pseudo-amino acid composition — have been used to predict protein fold recognition[65]. The resulting models (from each data type) were combined in a weighted voting scheme to determine the fold of the protein. Finally, network-based approaches have been developed in which a Bayesian network is constructed using gene expression data, metabolomic data and SNP genotype data, followed by integration to construct probabilistic causal networks[66–68]. In each of these model-based integration examples, a model is built on each data type individually, and the models are then combined in some meaningful way to detect integrative models.

It is important to note that model-based integration requires a specific hypothesis and analysis for each data

Kernel-based integration
The use of a valid kernel to perform a data matrix transformation before the integration of multiple data types.

Graph-based integration
The use of graphs to perform a data matrix transformation before integration. A graph is a natural method for analysing relationships between samples, as the nodes depict individual samples and the edges represent their possible relationships.

Majority voting
A method in which multiple models are constructed and subsequently evaluated to determine which performs best.

Ensemble classifiers
Classifiers constructed through the use of multiple learning methods to obtain better predictive performance than could be obtained from any of the individual learning algorithms.

Bayesian network
A type of statistical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.

type, and a mechanism to combine the resulting models in a meaningful way. Consider a data set of cancer tumour tissue and normal tissue with DNA sequence, methylation and metabolomic data measured. Each of the three data types can be analysed for association with cancer. The resultant DNA sequence model, methylation model and metabolomics model can then be integrated to identify a meta-dimensional model. As the only variables that are incorporated into the integrative analysis are the ones that are detected in the data-type-specific modelling process, it is possible to miss some of the interactions between different data types if they do not have effects to identify within the data type. For example, if there is a pattern of methylation and another pattern of protein expression that are not associated with the outcome independently but only associated through their interaction, then their effects will be missed in model-based integration. Moreover, these forms of ensemble-based approaches are well known for overfitting[69]. Therefore, model-based integration is particularly suitable if each genomic data type is extremely heterogeneous, such that combining the data matrix (concatenation-based integration) or performing data transformation to a common intermediate format (transformation-based integration) is not possible.

For the meta-dimensional analysis descriptions above, we only consider the data integration approaches from the point of view of a supervised learning strategy in which data with known labels (outcome or phenotype) are used. However, unsupervised learning is another category for data integration in which there are no known labels or phenotypes of interest, but analysis of the data set (one approach is to use clustering) might identify hidden structure in the observed data. For example, iCluster uses a joint latent variable model for integrative clustering of the meta-dimensional genomic data[70]. In addition, there are other integrative clustering methods based on the Bayesian approach in the context of exploratory or unsupervised learning[71,72]. The unsupervised learning strategies may also potentially add increased power and benefit to meta-dimensional data analysis.

Caveats and limitations

In this Review, we discuss several strategies for analysing multi-omic data, with the goal of elucidating the genetic architecture of complex traits using data integration approaches. As with any analysis, there are limitations and caveats to each method and, in addition, there are some broad limitations that should be mentioned. First, it is difficult to comprehensively assess the statistical power of many of these methods. Some approaches have theoretical distributions from which power calculations can be performed[73]. For others, simulation studies and/or permutation testing is needed to estimate empirical power[30]. In either case, these power estimates will apply only to the data set or simulation at hand, and they have limited use for interpreting the universal power of the approach. Therefore, power calculations or estimates based on these systems genomics approaches should be interpreted carefully.

Some of the analysis strategies described have potential pitfalls that could lead to limited power to identify certain associations. For example, a single variable (such as a SNP) in the genome may often be functional and associated with a trait, whereas SNPs in LD with the functional SNP may be associated but have no function. By performing data reduction, we may by chance filter out the functional SNP but keep the non-functional SNPs that are in LD with it, thereby missing the association with the functional SNP. In addition, most of the analysis techniques do not perform an exhaustive evaluation of possible statistical or computational models, as the computation time can be prohibitive. These methods rely on surrogate signals and correlation in the data that would allow model identification without exhaustively testing all possible models. Therefore, depending on the approach, true models (those that actually explain the biology) might not be assessed. Some data reduction methods, such as factor analysis, result in derived variables that extract orthogonal, or independent, relationships from the data; however, understanding which primary variables are essential can be difficult. Therefore, interpreting models composed of derived variables can be challenging.

Discussion of these limitations and caveats is not meant to discourage readers from using any of these systems genomics approaches. However, it is critical that the assumptions of the model, limitations of the analysis, and caution about inference and interpretation be taken into consideration for a successful multi-omic study.

Replication. An important consideration in large-scale analyses is the potential for false discoveries, so it is important to determine a way to identify results that are more likely to be true associations and not false positives. The 'gold standard' in human genetics is to look for replication of results using independent data[74], and seeking replication of multi-omic models is one way to identify robust predictive models. The strictest definition of replication with genetic variation data requires the same type of variation in a locus to be associated with the same trait and with the same direction of effect[74]. This ensures a more stringent protection from type I errors. However, there are problems with this definition for replication. For example, when using SNP data, this replication requirement ignores the fact that most SNP variants reported in genome-wide association studies are tag SNPs. Therefore, the tested SNPs are likely to be non-functional but correlated with the functional SNP owing to LD. As such, one would not necessarily expect the same variants to be associated in multiple data sets, especially when small differences in allele frequency variation can have a large effect on LD patterns[75]. Thus, in one data set, two SNPs might show main effects and an interaction effect, whereas in a second data set, SNPs in LD with those SNPs from the first data set might exhibit the strongest signal. When seeking replication, we recommend careful consideration of the underlying functional genomic units that are represented by each variable, and seeking replication of the genomic signals that are relevant and appropriate to the data at hand. There may be

Overfitting
Building a statistical model that explains the training data set that but does not generalize to independent data.

Type I errors
(Also known as false positives). The acceptance of the alternative hypothesis when the null hypothesis is true.

Genome-wide association studies
Studies that aim to identify disease- or trait-related genetic variations from the whole genome.

complex predictive models detected in the replication data set that do not exactly represent the initial discovery predictive models but that contain the same majority of genes or SNPs in LD with original SNPs, or genes within a same pathway. In this case, a similar model in which the same genes are present in the discovery and replication data sets with the same biological context may also be considered replication of the biological signal. The data set used to seek replication is also an important consideration. In some cases, external replication is possible, so independent data sets can be used. However, because of the cost of some molecular assays and the limited availability of tissues, independent data sets are not often readily available. In such cases, internal replication can be implemented. Several data-splitting or internal-validation approaches are available (reviewed in REF. 30). Finally, strategies are in development for using extrinsic data to develop evidence to support an association that cannot be directly replicated. For example, the diverse convergent evidence (DiCE) approach integrates information from multiple sources (omics, informatics and laboratory experiments) to estimate the strength of the available corroborating evidence supporting a given association[76].

Validation. Functional validation is a viable alternative to replication, and the focus lies in performing additional complementary or orthogonal experiments to corroborate the evidence that emerged from the original discovery experiment. For example, basic experimental bench science can be used to provide validation for statistical models[77,78]. This type of evidence has the potential to identify the biological mechanisms that underlie the statistical association. As such, this is a highly desirable validation technique.

Another validation approach is the use of text mining to find literature that supports or refutes the original findings. Many text-mining tools have been developed as a means of performing this type of informatics analysis. Gene relationships among implicated loci (GRAIL) is a commonly used tool in human genomics that allows one to search for the co-occurrence of genes in PubMed abstracts to determine potential biological connections between associated genes[79].

Finally, *in silico* modelling is an additional approach that can be useful. Based on a series of experiments that provide small portions of a complete model, mathematics can be used to integrate these different elements and make predictions about outcomes. For example, Crooke *et al.*[80] used a theoretical pathway for oestrogen metabolism, a statistical model of gene–gene interaction, a series of kinetic experiments and differential equations models to predict breast cancer risk.

Correlated variables. As mentioned above, different types of high-throughput data are likely to have highly correlated variables both within and between data types. In genome-wide SNP arrays, many SNPs are correlated with one another owing to LD. There are also varying levels of correlation, for example, between SNPs and gene expression, as well as between gene expression and methylation. These correlations can be used to help to guide, filter or interpret data; however, correlation can create problems for some analytical methods. In regression analyses, multi-collinearity (that is, high levels of correlation) might not allow matrix inversion, which is required to estimate reliable regression coefficients[81]. It is important to understand how each method handles correlated data, and whether pre-preprocessing is necessary to reduce the level of correlation (that is, pruning out correlated variables). The decision to reduce the correlation will be based on the analysis method selected.

Overfitting. Finally, overfitting is always a risk in data-driven analytical methods. This occurs when the model classifies or predicts the outcome for the samples within the data set extremely well but performs poorly on data that was not used to build the model. This often occurs when dealing with high-dimensionality problems in the data (such as small sample size and many independent variables), which leads to sparse data matrices when considering three, four or more variables in the model. Fortunately, there are many techniques to prevent overfitting. For example, cross-validation is a statistical technique in which some proportion of the data set is used to build the model and a subset is used for testing the model[30]. Another approach is the use of receiver–operating curves and the area under the curve. These approaches balance the sensitivity and specificity of the models to help to select the optimal models[82]. Additionally, Pareto optimization is a technique that is commonly used in computer science; two metrics of the models are compared: a fitness metric (that is, accuracy or area under the curve) and a parsimony metric (that is, the number of variables in the model)[83,84]. The goal in Pareto optimization is to find the fittest model with the simplest structure. This approach works relatively well because often the reason for overfitting is the inclusion of too many variables in a model.

Future directions

Our ability to generate molecular data has been improving at a rapid pace for the past decade, and this trend is likely to continue for the next decade. Most of the omic data are generated on crude tissue extracts from whole blood or other tissue types, such as lung, liver and heart tissues. However, single-cell technologies are advancing and showing promise for the future, and it is likely that we will soon have the capability to generate omic data on single cells from different tissue types of interest. The costs are also likely to continue to decrease, making the ability to generate these high-throughput omic data on very large sample sizes a reality.

To complement the continuation of data generation technologies, data analysis strategies will also experience major advancements. Computer technology for processing and storing data continues to evolve and expand, and this will enable more computational power to push analyses further than have been possible before. Additionally, the reductionist paradigm of looking for the 'low-hanging fruit' (the single variables that explain some portion of trait variability) is slowly becoming

less prevalent. Novel questions will be asked about the complex interplay of different types of omic data using new statistical and machine-learning approaches as more researchers think 'outside the box'. These emerging systems genomics approaches yield more informative results, and the pace of development will accelerate. As the tools become more readily available and affordable, such systems genomics approaches will prevail as the dominant type of study design and analytical strategy — the days of studying molecular data variability in isolation are slowly coming to an end.

Conclusions

The emergence of new statistical and computational techniques will facilitate the search for genomic factors that contribute to the architecture of complex traits and continue to unravel novel biological insights. The realization that performing all analyses within one data type has limitations led to the development of many new ideas and methods for data integration. These systems genomics approaches are still in their infancy, and gold standard methods have not yet emerged. However, there are various strategies that can currently be implemented to perform a powerful integrative analysis, although it is likely that one single method will not perform best for all data analyses. Thus, approaches need to be selected according to specific types of data, different types of scientific questions or different types of underlying genomic models. Remaining issues and challenges include strategies for combining data from multiple time points and multiple tissues; data normalization and scaling issues when considering multiple data types; consideration of complex biological processes, including feedback loops and compensatory mechanisms, which will require non-linearity models; and the possibility of combining different data types from different individuals or samples. A bottleneck in the development of new strategies is the inability to know what the true models should include and therefore what the most effective modelling strategies will be. In research, we tend to look at previous successes and the types of biological models identified to guide what we look for in the future. However, systems genomics approaches, as described here, are new; therefore, we do not have past examples of success to use as a guide. As more data are generated across multiple data types and multiple tissues, novel explorations will further our understanding of important biological processes and enable more comprehensive systems genomic strategies. It is through collaboration among statisticians, mathematicians, computer scientists, bioinformaticians and biologists that the continued development of meta-dimensional analysis methods will lead to a better understanding of complex-trait architecture and generate new knowledge about human disease and biology.

1. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
2. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Rev. Genet.* **12**, 87–98 (2011).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
4. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Rev. Genet.* **11**, 191–203 (2010).
This is a comprehensive review of DNA methylation data analysis.
5. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
6. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Rev. Genet.* **14**, 35–48 (2013).
7. Shulaev, V. Metabolomics technology and bioinformatics. *Brief. Bioinform.* **7**, 128–139 (2006).
8. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Rev. Genet.* **14**, 618–630 (2013).
9. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**, 1198–1211 (1998).
10. Horvath, S., Xu, X. & Laird, N. M. The family based association test method: strategies for studying general genotype—phenotype associations. *Eur. J. Hum. Genet.* **9**, 301–306 (2001).
11. Devlin, B., Roeder, K. & Bacanu, S. A. Unbiased methods for population-based association studies. *Genet. Epidemiol.* **21**, 273–284 (2001).
12. Reif, D. M., White, B. C. & Moore, J. H. Integrated analysis of genetic, genomic and proteomic data. *Expert Rev. Proteomics* **1**, 67–75 (2004).
13. Hamid, J. S. *et al.* Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics* **2009**, 869093 (2009).
14. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18**, 389–401 (2007).
15. Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nature Rev. Genet.* **11**, 476–486 (2010).
16. Holzinger, E. R. & Ritchie, M. D. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics* **13**, 213–222 (2012).
17. Holzinger, E. *et al.* in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (eds Giacobini, M., Vanneschi, L. & Bush, W.) **7246**, 134–143 (Springer Berlin Heidelberg, 2012).
18. Holzinger, E. R. *et al.* ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pac. Symp. Biocomput.* 385–396 (2013).
19. Stein, L. D. The case for cloud computing in genome informatics. *Genome Biol.* **11**, 207 (2010).
20. Dorff, K. C. *et al.* GobyWeb: simplified management and analysis of gene expression and DNA methylation sequencing data. *PLoS ONE* **8**, e69666 (2013).
21. Reid, J. G. *et al.* Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014).
22. Heath, A. P. *et al.* Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *J. Am. Med. Inform. Assoc.* **21**, 969–975 (2014).
23. Turner, S. *et al.* Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* **68**, 1.19.1–1.19.18 (2011).
24. Zuvich, R. L. *et al.* Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet. Epidemiol.* **35**, 887–898 (2011).
This paper provides detailed lessons learned about quality control processes in high-throughput genotype data and guides readers toward best practices when cleaning and merging genotype data.
25. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
26. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
27. Marguerat, S. & Bähler, J. RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* **67**, 569–579 (2010).
28. Hirst, M. & Marra, M. A. Next generation sequencing based approaches to epigenomics. *Briefings Funct. Genom.* **9**, 455–465 (2010).
29. Johnstone, I. M. & Titterington, D. M. Statistical challenges of high-dimensional data. *Phil. Trans. R. Soc. A.* **367**, 4237–4253 (2009).
30. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag, 2001).
31. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput.* 368–379 (2009).
32. Greene, C. S., Penrod, N. M., Kiralis, J. & Moore, J. H. Spatially uniform ReliefF (SURF) for computationally-efficient filtering of gene–gene interactions. *BioData Min.* **2**, 5 (2009).
33. Moore, J. H. & White, B. C. in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (eds Marchiori, E., Moore, J. H. & Rajapakse, J. C.) 166–175 (Springer Berlin Heidelberg, 2007).
34. Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **15**, 265–286 (2006).
35. Holland, J. H. Genetic algorithms. *Sci. Am.* **267**, 66–72 (1992).
36. Vilhjálmsson, B. J. & Nordborg, M. The nature of confounding in genome-wide association studies. *Nature Rev. Genet.* **14**, 1–2 (2013).
37. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11**, 407–409 (2014).
38. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
39. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161 (2007).
40. Hartford, C. M. *et al.* Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood* **113**, 2145–2153 (2009).
41. Huang, R. S. *et al.* A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl Acad. Sci. USA* **104**, 9758–9763 (2007).
This is one of the first papers to present an integrative analysis to identify DNA variants and gene expressions associated with chemotherapeutic drug-induced cytotoxicity.

42. Huang, R. S., Duan, S., Kistner, E. O., Hartford, C. M. & Dolan, M. E. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol. Cancer Ther.* **7**, 3038–3046 (2008).
43. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37**, 710–717 (2005).
This study used an integrative approach to use DNA variation and gene expression data to identify drivers of complex traits.
44. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotech.* **31**, 142–147 (2013).
45. Khan, Z. *et al.* Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Mol. Syst. Biol.* **8**, 602 (2012).
46. Wei, X. & Wang, X. A computational workflow to identify allele-specific expression and epigenetic modification in maize. *Genomics Proteomics Bioinformatics* **11**, 247–252 (2013).
47. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
This paper reports the sequencing and analysis of mRNA and microRNA of hundreds of multi-ethnic individuals from the 1000 Genome Project.
48. Maynard, N. D., Chen, J., Stuart, R. K., Fan, J.-B. & Ren, B. Genome-wide mapping of allele-specific protein–DNA interactions in human cells. *Nature Methods* **5**, 307–309 (2008).
49. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
50. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
51. Encode Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
52. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. Kim, D., Shin, H., Song, Y. S. & Kim, J. H. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Inform.* **45**, 1191–1198 (2012).
This study shows a graph-based approach for predicting cancer clinical outcome by integrating multi-omics data as a transformation-based integration.
54. Fridley, B. L., Lund, S., Jenkins, G. D. & Wang, L. A. Bayesian integrative genomic model for pathway analysis of complex traits. *Genet. Epidemiol.* **36**, 352–359 (2012).
55. Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A. & Sander, C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS ONE* **6**, e24709 (2011).
56. Holzinger, E. R., Dudek, S. M., Frase, A. T., Pendergrass, S. A. & Ritchie, M. D. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics* **30**, 698–705 (2014).
ATHENA is a tool for meta-dimensional integration of multi-omics data. This paper describes the software and its application for these types of analyses.
57. Kim, D., Li, R., Dudek, S. M. & Ritchie, M. D. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Min.* **6**, 23 (2013).
58. Clarke, R. *et al.* The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Rev. Cancer* **8**, 37–49 (2008).
This review addresses the properties of high-dimensional data spaces and the challenges for data analysis and interpretation.
59. Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635 (2004).
This is the first study to propose a kernel-based integration as a transformation-based integration.
60. Borgwardt, K. M. *et al.* Protein function prediction via graph kernels. *Bioinformatics* **21**, i47–i56 (2005).
61. Tsuda, K., Shin, H. & Schölkopf, B. Fast protein classification with multiple networks. *Bioinformatics* **21**, ii59–ii65 (2005).
62. Shin, H., Lisewski, A. M. & Lichtarge, O. Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics* **23**, 3217–3224 (2007).
63. Turner, S. D., Dudek, S. M. & Ritchie, M. D. ATHENA: a knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait loci. *BioData Min.* **3**, 5 (2010).
64. Drăghici, S. & Potter, R. B. Predicting HIV drug resistance with neural networks. *Bioinformatics* **19**, 98–107 (2003).
65. Shen, H.-B. & Chou, K.-C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics* **22**, 1717–1722 (2006).
66. Akavia, U. D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
This paper demonstrated a computational framework that identified drivers of melanoma using chromosomal copy number and gene expression data.
67. Zhu, J. *et al.* Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol.* **10**, e1001301 (2012).
68. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genet.* **40**, 854–861 (2008).
69. Opitz, D. & Maclin, R. Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999).
70. Shen, R. *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7**, e35236 (2012).
71. Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. & Wild, D. L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297 (2012).
72. Lock, E. F. & Dunson, D. B. Bayesian consensus clustering. *Bioinformatics* **29**, 2610–2616 (2013).
73. Dupont, W. D. & Plummer, W. D. Power and sample size calculations. A review and computer program. *Control Clin. Trials* **11**, 116–128 (1990).
74. NCI–NHGRI Working Group on Replication in Association Studies. Replicating genotype–phenotype associations. *Nature* **447**, 655–660 (2007).
75. Greene, C. S., Penrod, N. M., Williams, S. M. & Moore, J. H. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS ONE* **4**, e5639 (2009).
76. Ciesielski, T. *et al.* Diverse convergent evidence in the genetic analysis of complex disease: Coordinating omic, informatic, and experimental evidence to better identify and validate risk factors. *BioData Min.* **7**, 10 (2014).
77. Van Poucke, M., Vanhaesebrouck, A. E., Peelman, L. J. & Van Ham, L. Experimental validation of *in silico* predicted *KCNA1*, *KCNA2*, *KCNA6* and *KCNQ2* genes for association studies of peripheral nerve hyperexcitability syndrome in Jack Russell Terriers. *Neuromuscul. Disord.* **22**, 558–565 (2012).
78. Sharaf, R. N. *et al.* Computational prediction and experimental validation associating FABP-1 and pancreatic adenocarcinoma with diabetes. *BMC Gastroenterol.* **11**, 5 (2011).
79. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
80. Crooke, P. S. *et al.* Estrogens, enzyme variants, and breast cancer: a risk model. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1620–1629 (2006).
81. Farrar, D. E. & Glauber, R. R. Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* **49**, 92 (1967).
82. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874 (2006).
83. Moore, J.H., Hill, D. P., Sulovari, A & Kidd, L.C. in *Genetic Programming Theory and Practice X* 87–101 (Springer, 2013).
84. Jin, Y. & Sendhoff, B. Pareto-based multiobjective machine learn: an overview case studies. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **38**, 397– 415 (2008).
85. Kristensen, V. N. & Borresen-Dale, A. L. Molecular epidemiology of breast cancer: genetic variation in steroid hormone metabolism. *Mutat. Res.* **462**, 323–333 (2000).
86. Mitrunen, K. *et al.* Glutathione *S*-transferase M1, M3, P1, and T1 genetic polymorphisms and susceptibility to breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **10**, 229–236 (2001).
87. Kiyotani, K. *et al.* A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese. *Hum. Mol. Genet.* **21**, 1665–1672 (2012).
88. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature Genet.* **45**, 392–398, 398e1–2 (2013).
89. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genet.* **45**, 353–361, 361e1–2 (2013).
90. Zheng, W. *et al.* Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. *Hum. Mol. Genet.* **22**, 2539–2550 (2013).
91. Mogushi, K. & Tanaka, H. PathAct: a novel method for pathway analysis using gene expression profiles. *Bioinformation* **9**, 394–400 (2013).
92. Chung, R.-H. & Chen, Y.-E. A two-stage random forest-based pathway analysis method. *PLoS ONE* **7**, e36662 (2012).
93. Bailey, L. R., Roodi, N., Dupont, W. D. & Parl, F. F. Association of cytochrome P450 1B1 (CYP1B1) polymorphism with steroid receptor status in breast cancer. *Cancer Res.* **58**, 5038–5041 (1998).
94. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
95. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
96. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
97. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
98. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
99. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
100. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
This important paper presents the relationship between genetic variation, gene expression and clinical phenotypes using human blood and adipose tissue.

Acknowledgements
Support for the authors was provided by the US National Institutes of Health grants LM010040 (ATHENA) and HL065962 (the P-STAR Network Resource of the PGRN). E.R.H. was funded by grant Z01 HG00153-08-IDRB. R.L. was funded by the US National Science Foundation under Grant number DGE1255832. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US National Science Foundation.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
ATHENA: https://ritchielab.psu.edu/athena-downloads
Glmpath: http://cran.r-project.org/web/packages/glmpath/index.html
Graph-based semi-supervised learning: http://mammoth.bcm.tmc.edu/papers/lisewski2007.gz
ipred: http://cran.r-project.org/web/packages/ipred/index.html
SKMsmo: http://imagine.enpc.fr/~obozinsg/SKMsmo.tar
Weka 3: http://www.cs.waikato.ac.nz/ml/weka/
WinBUGS: http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF