Biochimica et Biophysica Acta 1861 (2017) 335–353

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Review

Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions

Patrick Y.P. Kao [a], Kim Hung Leung [b], Lawrence W.C. Chan [b], Shea Ping Yip [b,*], Maurice K.H. Yap [a]

[a] Centre for Myopia Research, School of Optometry, The Hong Kong Polytechnic University, Hong Kong SAR, China
[b] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 17 October 2016
Accepted 19 November 2016
Available online 23 November 2016

Keywords:
Pathway analysis
Genome-wide association study (GWAS)
Complex disease
Multi-omics
Interaction
Rare variants

ABSTRACT

Background: Genome-wide association studies (GWAS) is a major method for studying the genetics of complex diseases. Finding all sequence variants to explain fully the aetiology of a disease is difficult because of their small effect sizes. To better explain disease mechanisms, pathway analysis is used to consolidate the effects of multiple variants, and hence increase the power of the study. While pathway analysis has previously been performed within GWAS only, it can now be extended to examining rare variants, other "-omics" and interaction data.
Scope of review: 1. Factors to consider in the choice of software for GWAS pathway analysis. 2. Examples of how pathway analysis is used to analyse rare variants, other "-omics" and interaction data.
Major conclusions: To choose appropriate software tools, factors for consideration include covariate compatibility, null hypothesis, one- or two-step analysis required, curation method of gene sets, size of pathways, and size of flanking regions to define gene boundaries. For rare variants, analysis performance depends on consistency between assumed and actual effect distribution of variants. Integration of other "-omics" data and interaction can better explain gene functions.
General significance: Pathway analysis methods will be more readily used for integration of multiple sources of data, and enable more accurate prediction of phenotypes.

1. Introduction – why pathway analysis?

The growth in knowledge of our genome and new development in genomic technologies have enabled the identification of risk factors of complex diseases using genome-wide association studies (GWAS). "Complex" diseases are so called because they are caused by multiple genetic and environmental risk factors. In complex diseases, the causative genetic factors usually have small effect sizes. In GWAS, a huge number of genetic variants are tested simultaneously. To account for multiple testing, the *p*-value threshold of a single-variant test for declaring genome-wide significance was suggested to be 5×10^{-8} [1]. Despite opinions to relax this threshold to the order of 10^{-7} [2], the threshold is still very stringent. Given these factors (small effect sizes of variants, and required stringent statistics because of multiple testing), it is a challenging task to perform GWAS powerful enough to map disease genes for complex diseases successfully.

To increase the power of a GWAS, one method is to take missing heritability into account. Missing heritability refers to the inability for the disease-susceptible variants found from GWAS to explain the complete genetic component contributing to the increased risk of a phenotype [3]. One reason is that the genetic variants of complex diseases, each only having a small effect, cannot all be detected by single-variant statistical analyses. To address this issue, it is better to consider collectively effects of interesting variants *together* in a meaningful way in order to increase statistical power and reduce the burden of multiple testing [4].

Pathway analysis complements single-variant analysis in two ways. First, by combining weaker but related single-variant signals, the resulting statistics could be improved if these variants are collectively

Abbreviations: ARTP, adaptive rank truncated product; BMI, body mass index; (r)BEE, (robust) brain-expressed enhancer; CNV, copy number variation; eQTL, expression quantitative trait locus; ESI-MS/MS, electrospray ionization mass spectrometry; eSNP, expression single-nucleotide polymorphism; GCN, gene coexpression network; G-E, gene-environment; G-G, gene-gene; GEWIS, genome-wide interaction study; GO, gene ontology; GRN, gene regulatory network; GSAA, gene set association analysis; GSEA, gene set enrichment analysis; GWAS, genome-wide association study; HLA, human leukocyte antigen; iGSEA, improved gene set enrichment analysis; iGWAS, integrative GWAS; IPA, Ingenuity Pathway Analysis; KEGG, Kyoto Encyclopaedia of Genes and Genomes; lincRNA, long intervening non-coding RNA; LD, linkage disequilibrium; MAF, minor allele frequency; mGWAS, GWAS on metabolic traits; MIAME, minimum information about a microarray experiment; MS, mass spectrometry; NGS, next-generation sequencing; NMR, nuclear magnetic resonance; PPI, protein-protein interaction; RV, rare variant; SKAT, sequence kernel association test; SNP, single-nucleotide polymorphism; WGCNA, weighted gene coexpression network analysis; WKS, weighted Kolmogorov–Smirnov (statistics).

* Corresponding author at: Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China.
E-mail address: shea.ping.yip@polyu.edu.hk (S.P. Yip).

http://dx.doi.org/10.1016/j.bbagen.2016.11.030

related to the phenotype (a situation described as searching a "string" of needles in the haystack rather than needle-by-needle) [5]. It is particularly useful for pilot studies with small sample sizes (which have small single-marker statistics even for the most significant loci) to allow investigators to prioritise variants for follow-up analysis. Second, pathway-based studies can allow the discovery of novel *sets* of genetic variants with related functions, which helps explain the observed data. In both cases, we hope to increase the power of the hypothesis-free GWAS by providing functional annotations, and combine effects of variants within appropriate functional units.

Pathway analysis combines signals of multiple variants. However, what is the biological meaning of such analysis? There are two main goals in biomedical research: understanding of molecular mechanisms underlying a phenotype or disease on the one hand, and discovery and design of drugs for disease treatment on the other hand [6]. To achieve these aims, effects on the body caused by inherited genetic background and external (environmental) changes have to be considered collectively. In the past, experiments were analysed in a reductionist manner, for which only a single level of data was considered at a time because of the lack of tools for analysis [7]. Take GWAS as an example, a set of variants can be obtained by extracting variants passing a pre-defined *p*-value threshold in association tests. However, the functions and biological meaning of this set of variants or genes cannot be inferred by *p*-values alone. The retrieval of such information requires yet another layer of evaluation separate from the association study [7]. Pathway analysis can serve as a proxy in filling the gap here to infer the relationship among the observed set of selected genes represented by significant variants, and the strengths of the relationship. As a result, the association findings could be interpreted more easily.

There are other reviews focusing on pathway analysis of GWAS [5,8–11]. This review is broadly divided into two parts. The first part discusses technical aspects that researchers may find useful before carrying out pathway analysis for GWAS data. It aims to describe how to carry out pathway analysis for common variants in GWAS, and discuss the aspects that researchers (especially those with experience in common variant analysis only) may consider if they wish to carry out the analyses. The second part discusses possible steps that enable prediction of phenotypes more accurately by using extra -omics data. We deliberate on how pathway analysis is extended to integrating rare variants, other "-omics" data, and gene-environmental interactions. We hope this review article will enable researchers of GWAS to get started with pathway analysis right away. Meanwhile, they will also appreciate the possibility and value of expanding the analysis paradigm to other data types. Ultimately, this would help us understand the aetiology of diseases better, and could possibly shed light on more effective therapeutic measures.

Readers should note that, throughout the text, pathway analysis is referred to as having almost the same meaning as network analysis unless otherwise specified, which both mean a broader sense of multi-SNP analysis based on certain information. However, we would like to draw readers' attention to the fact that, in a narrower sense, pathway and network analyses are not the same based on the relationship of the genes included for analysis (Fig. 1; see Box 1 for a detailed description of pathway definitions).

1.1. Steps involved in GWAS pathway analysis – how pathway analysis is done: the big picture

There are three basic steps in pathway analysis of GWAS data (Fig. 2). First, users need to choose and determine the gene set definitions of the pathways to be used for pathway analysis. Second, input variants are mapped onto the genes they belong to for preparing the calculation of gene and/or pathway-based statistics. Finally, pathway statistics are calculated, either by a one-step approach, which only reports the pathway-based statistics; or by a two-step approach, which calculates pathway-based statistics using intermediate gene-based statistics. Various aspects that will affect the choice of analysis software tools will be discussed below.

2. How should we choose analysis software? Aspects to be considered for choosing gene set definition and analysis software

2.1. Input data – do you have covariates?

Table 1 lists software packages for pathway analysis (and interaction analysis). Pathway analysis software packages accept various input data formats, including *p*-values of single-marker association tests [12–16], keywords/gene list [17–19], or raw genotype data [9,20,21]. If covariates are to be considered in pathway analysis, it is better to control for it at an early stage of generating individual variant-level statistics. If raw data are available, obtaining covariate-adjusted statistics from raw data is straightforward. Genetic analysis software packages, such as PLINK [20] (which reports association statistics using genotype data as input) and SNPTEST [22] (which reports association statistics using allele dosage results from imputation software IMPUTE [23]), are custom-made to generate covariate-adjusted test statistics for single-marker association analysis from genotype data. The covariate-adjusted *p*-values can then be used in downstream pathway analysis. However, it should be noted that covariates usually cannot be incorporated into pathway analysis algorithms directly. Therefore, if covariate adjustment is crucial to analysis, it is advised that adjustment of covariates is first carried out in single-marker analysis. Pathway analysis is then carried out using methods that allow *p*-values as the input data.

2.2. Input data affects choice of software: competitive vs self-contained

Based on the difference in the hypothesis being tested for generating pathway *p*-values, pathway analysis methods can be divided into either self-contained or competitive (Fig. 3) [5]. For the self-contained approach, we test the hypothesis that the observed pathway is associated with a phenotype by comparing against a null genetic background (assuming to have no association). For the competitive approach, we test the hypothesis that the statistics of genes within a pathway is significantly different from that not within the pathway. To reflect the difference, the competitive approach is named as "enrichment" methods while self-contained approach is named as "association" methods [24].

What data are available limits the choice of appropriate analysis approach and hence analysis software tools and methodology. To carry out analysis using the competitive approach, data of genes not within the pathway of interest must also be available. In contrast, the self-contained approach does not require such data (as it can be obtained by permutations). Therefore, the competitive approach is not applicable to candidate-gene data (since there are no data for other genes not within the candidate gene set), while the self-contained approach is applicable to both genome-wide and candidate-gene data [11]. In some studies, the competitive approach is used in the discovery stage of GWAS, and then followed by the self-contained approach for replication. A recent evaluation of the statistical properties of gene-set enrichment methods suggests that competitive approach have an advantage over self-contained approach in that self-contained approach fails to take into consideration information from other biological pathways [25].

2.3. Sources of gene set definition

Defining gene sets is an essential step in pathway analysis. Table 2 lists some common databases for annotating pathways. The gene set information of these data sources can be classified into functional pathways, networks, gene ontology, and associated gene sets (see Box 1). Some software packages allow multiple sources of gene set definitions for more comprehensive analysis. For example, i-GSEA4GWAS [13,14] (see Table 1) allows users to choose among online datasets including

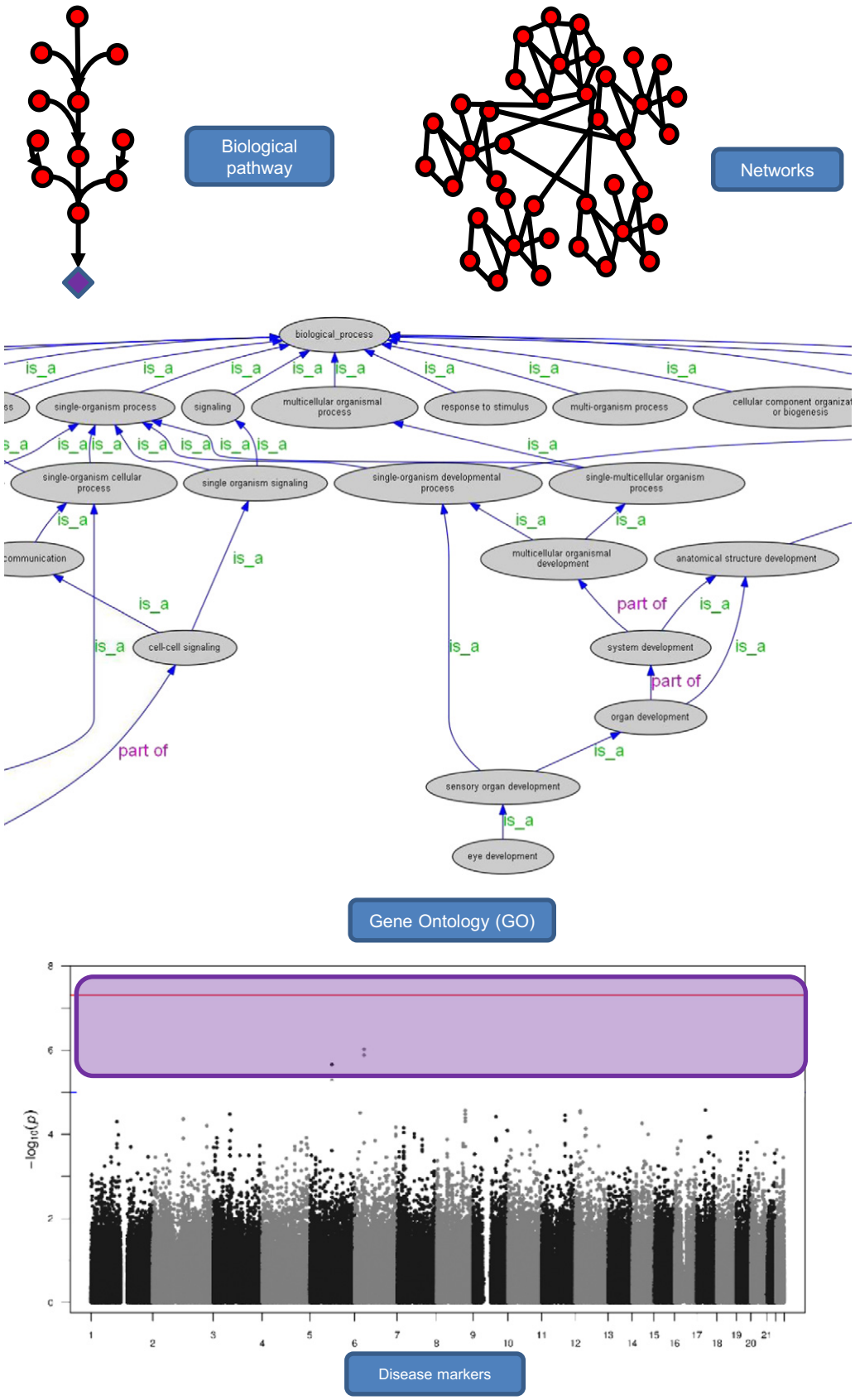

Fig. 1. Types of gene set definitions. There are four main categories of gene set definitions. Biological pathways have a specific starting point and an outcome, with the outcome being molecule or certain cellular status. Networks do not necessarily have an end point, and only describe relationships between genes and/or their products. Gene ontologies (GO) describe gene properties from a hierarchical class structure of three main aspects: molecular functions, biological processes, and cellular components. Illustrated as an example is the partial graph (produced from OBO-Edit, an open-source ontology editing tool) using the GO term "eye development" (GO:0001654). The last type of gene set definition is disease biomarkers. These markers may not have functional relevance, and may be created, for instance, by extracting significant variants from association studies.

Box 1
Definitions of pathways.

A "pathway" does not *necessarily* mean a visualised network of genes. It can refer to a group of genes that are related to each other according to certain definitions [5]. According to the nature of the gene sets defined, these pathways can be grouped into biological pathways, interacting networks, gene ontologies, and biomarkers (significant variants) (Fig. 1) [29].

Biological pathway

Biological pathway has the "strictest" definition among different types of pathways. Each biological pathway is a means for one "endpoint" (being a product or a biological function) reaching another (Fig. 1). To describe pathways precisely and to distinguish one pathway from another, it is crucial to know the endpoints (final outcomes of the pathways) as well as the intermediate steps involved. According to the nature of the "endpoints", biological pathways can be further categorised into molecular, cellular, organ/system and disease/intervention pathways. For molecular pathways, the endpoints are molecules or basic biochemical reactions. Endpoints of cellular and organ/system pathways are more complex. They include global cell status and higher tissue organisations, such as cell apoptosis and memory storage respectively. Disease pathways include events and/or reactions which can lead to disease onset. Intervention pathways, on the other hand, include those events and/or reactions which can alter disease presentation status. These endpoints can be outcomes of one or more combinations of the lower-order pathways described above. An example of disease pathway is Alzheimer's disease pathway in KEGG (entry: PATHWAY: map05010) [50], which includes lower-order molecular and cellular pathways such as calcium signalling, apoptosis and oxidative phosphorylation. Sometimes members in an intervention pathway are found to be associated with a phenotype, but without the actual mechanism known [9]. In that case, these members should only qualify as "biomarkers" (see below) because the knowledge of intermediate steps is missing [29].

Networks

Networks, on the other hand, do not necessarily have a distinct "endpoint" of biological function (Fig. 1) [29]. A network is more like a simple catalogue of all logical relationships among elements predicted or experimentally discovered. Various methodologies have been adopted in the curation for these databases [29,177], including yeast-two-hybrid system [178,179], affinity purification followed by mass spectrometry [180], protein complementation assay [181], co-immunoprecipitation, chromatin immunoprecipitation with DNA microarray, gene expression and text-mining [182]. The diversity in curation methods means that the databases are heterogeneous [29], and therefore very difficult to compare [177]. Although the exact mechanism of how genetic variants affect diseases is yet to be delineated, network data can allow us to understand the role of genetics in causing a disease [177].

Gene ontologies (GO)

GO endeavours to provide evidence-supported annotations to genomic products (including genes, proteins, non-coding RNAs and chemical complexes) to describe their biological roles [159]. Within GO are graph structures of GO classes which catalogue properties of the products (Fig. 1). These properties include molecular functions, biological processes and cellular locations (cellular components) of the products. A GO annotation of a specific product describes the relationship between it and a GO class, and all the evidence supporting the relationships [183]. However, while GO can help in discovering novel relationships, information of known relationship may not be easily inferred backwards accurately using groupings in GO. Therefore, when results are examined, it is important to also know the context of evidence – for example, whether the data are experimentally validated or computationally inferred [29].

Disease biomarkers

Another type of gene sets associated with a phenotype is disease biomarkers. Unlike other gene sets, disease biomarkers may not share similar functions or interact with each other. Biomarkers are suggested to be genomic markers that collectively associate better with a phenotype than individual genes. Such evidence usually comes from many different study populations. In addition, the hypothesis sometimes cannot hold because the marker sets may just reflect disease heterogeneity rather than a biomarker with power in disease prediction [29].

gene ontology (GO), Kyoto Encyclopaedia of Genes and Genomes (KEGG) and BioCarta pathways to define their gene sets for data analysis. Other software packages such as Ingenuity Pathway Analysis (IPA) [26] use their own curated databases to define gene sets. Some programs require users to input gene set definitions, and statistics are calculated for the input gene sets. Examples of such programs include adaptive rank truncated product (ARTP) method [27] and PLINK [20] set association.

There are both pros and cons for choosing software packages using available gene set definitions and those using user-defined definitions. The most obvious advantage of using a curated database is that users do not need to create their own gene sets. In addition, the gene sets involved are also created based on known functional knowledge, through which researchers can interpret their results easier. However, using defined pathways may deprive the users of the flexibility in defining gene sets. If researchers wish to test a customised set of genes based on their own hypotheses, then they must choose software that allows user-input gene set definitions. The web server i-GSEA4GWAS, for example, allows definition of gene sets from either curated databases or user-input gene set. Users should therefore choose appropriate software according to their hypothesis (Table 1).

2.4. One-step and two-step: which step should I make?

One method to categorise pathway analysis software is according to whether it is "one-step" or "two-step" [11] (Fig. 2). In a two-step design, *p*-values of individual variants of a gene are first considered to give a gene-based *p*-value or score. Pathway analysis is then performed using the gene-based statistics. The one-step approach, however, does not produce gene-based statistics, and pathway-based statistics are calculated from input variants directly [5,11]. There is no best answer for choosing one method over the other because there is no consensus yet on the best approach to combining single-variant statistics [28]. However, it is advisable to use all variants and analyse the data with pathways as units at an early stage so that most information can be obtained from pathway analyses [29].

2.5. Parameters for using software: size and nature of pathway

Pathway size (i.e. the number of genes in a pathway) imposes a significant impact on analysis results. Large pathways include more genes, and therefore may have a larger number of significant genes by sheer chance. On the other hand, small pathways may also lead to false positive results by including a few isolated significant variants [29]. To balance out the effects of both, the number of genes per pathway has been suggested to be 100 – 400 genes [9].

Besides the pathway size (number of genes), the composition of pathways may also affect results. Genes having large effects on a phenotype and genes involved in a number of pathways may render over-representation of pathways consisting of such genes, and therefore create a misinterpretation that other genes within the pathways also contribute to the phenotype. It is advised that if such genes exist in the test gene set, results should be compared using data without these genes to investigate whether there is a need to drop these genes as a quality control measure before pathway analysis. For example, the human leukocyte antigen (HLA) gene is a known genetic risk factor for both psoriasis and multiple sclerosis. In order to reduce the influence of HLA, a psoriasis study followed up only pathways that were significant before and after including HLA [30]. Similarly, a GWAS study of multiple sclerosis directly excluded HLA from pathway analysis to avoid complexity in interpreting results [31].

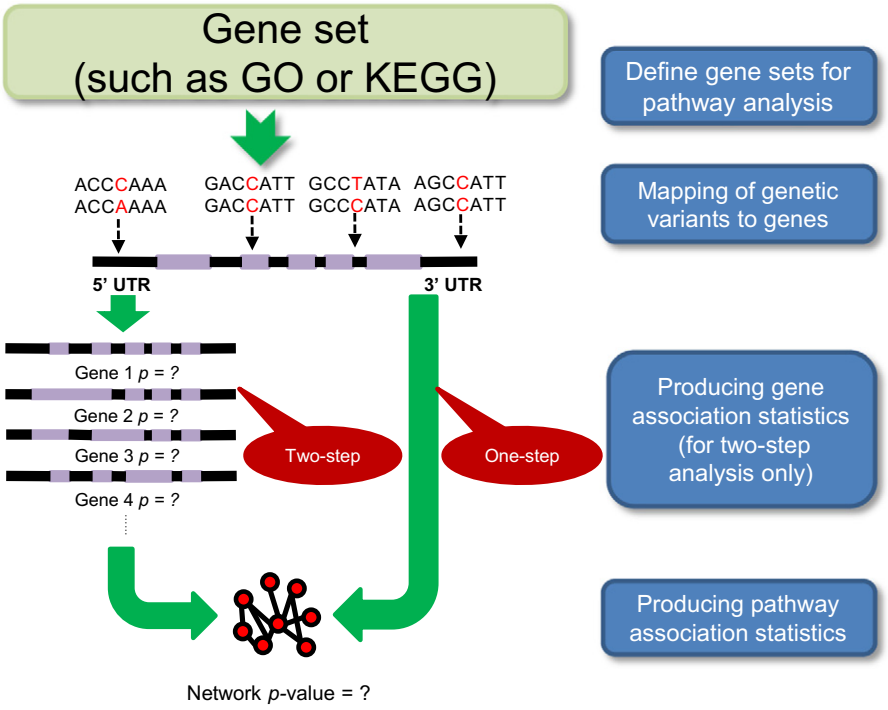

Fig. 2. Workflow of a pathway analysis for genetic studies. A typical pathway analysis includes three basic steps. First, gene set definitions are selected from various gene set databases. Second, different software tools will map variants onto the genes they belong to. Finally, for one-step analysis, pathway *p*-values are directly calculated from genotype data without returning gene *p*-values. For two-step analysis, gene *p*-values are first produced, before returning the final pathway *p*-values.

2.6. Parameter for using software: flanking size for mapping variants to genes

To produce gene- and pathway-based statistics, variants must first be assigned to their relevant genes. A simple approach is to relate only single-nucleotide polymorphisms (SNPs) within genes to their relevant genes. Nevertheless, this is not satisfactory because a large number of variants located outside gene exons will be excluded. One method to relieve this is to assign genetic variants in gene-flanking regions to their relevant genes. There is no exact answer to the covering region of interest. Despite the suggestions that most regulatory elements exist within 20-kb regions flanking a gene [28,32], values from 5 kb [33] to over 100 kb [34] have been used in different studies.

2.7. Other considerations

In addition to the technical aspects of the software, user friendliness, flexibility and expandability could improve ease of use. For example, programs such as IPA [26] and MetaCore [19] (Table 1) provide built-in options for network visualisation. Other software packages such as Cytoscape [35] (Table 1) may provide a platform which allows installation of "apps", i.e. plug-ins which can perform various tasks. Analysis and visualisation are therefore possible in one single platform with the possibility of adding new algorithms for analysis by installing new apps.

Table 3 lists some diseases for which pathway analysis has been applied to examine their genetic data. The corresponding software packages are also indicated.

3. Paradigm shift: from GWAS and beyond

In the past few years, the advancement in next-generation sequencing (NGS) technologies and multi-omics technologies has made the entire analysis paradigm walk "the extra mile". From the input-variant ("horizontal") perspective, sequencing technologies enable the detection and therefore analysis of rare variants. Multi-omics technologies provide data beyond the genetic level, thus allowing integrative ("longitudinal") analysis using other -omics data. In this section, we discuss some aspects of pathway analysis involving rare variants and other "-omics" platforms. This allows readers to compare and contrast such analyses with analysis of GWAS data, and appreciate how genetic data can be analysed together with other -omics data.

3.1. Rare variants

3.1.1. Why should we analyse multiple rare variants together?

The introduction of NGS has made deep sequencing of a large number of individual samples possible at a much lower cost. This has led to the discovery of numerous low-frequency variants (minor allele frequency (MAF) ranging from 1% to 5%) and rare variants (MAF < 1%). For ease of discussion in this article, we shall refer to all these variants collectively as "rare variants" (RVs). RV analysis has the potential to reveal novel variants predisposing to or causing diseases. Annotations of RVs are also more complete because functional units with clearer suggested roles are usually selected in targeted and exome sequencing studies. Together with the lowering sequencing cost, these factors have driven rapid growth in the number of RV analyses in the past decade [36].

In GWAS, single-variant analysis is the simplest and typical analysis method. For rare-variant analysis, single-variant analysis is also possible if the samples size is large enough to produce genome-wide significant results. However, even for a disease variant with large effect (e.g. an odds ratio of 2), it will require more than 100,000 samples for its detection with 80% power if its MAF is low (say 0.1%). Together with the multiple-testing penalty required to correct for the huge number of rare variants, obtaining adequate sample size for a powerful single-variant analysis is extremely challenging [37]. Therefore, region-based analysis

Table 1
List of software packages for pathway analysis and interaction analysis.

Software	Input	Description	URL/source
Pathway analysis software for GWAS			
Adaptive rank truncated product (ARTP) method [27]	Raw genotypes/SNP *p*-values	Two-step approach using top ranked (i.e. most significant) *p*-values. Choice of rank is by an efficient approach without the need for another layer of permutation.	http://dceg.cancer.gov/tools/analysis/artp
ALIGATOR [12]	SNP *p*-values	Users specify *p*-value cut-off for significant SNPs. Significant pathways are then counted.	http://x004.psycm.uwcm.ac.uk/~peter/
Chlibot [17]	Proteins/genes/keywords	Abstract text searching software for NCBI databases. Looks for relationships between abstracts using natural language programming (NLP) techniques.	http://www.chilibot.net/
Cytoscape [35]	(depend on "Apps")	Software platforms of "Apps" which allows integration of a variety of plug-ins.	http://www.cytoscape.org/
dmGWAS [132]	SNP *p*-values	Topology-based method using dense module searching (DMS), which tries to look for protein-protein interaction (PPI) sub-networks enriched with genes with small *p*-values.	https://bioinfo.uth.edu/dmGWAS/
GenGen [133]	Raw genotype	Gene statistics are represented by the most significant SNP signal of the genes. Pathway statistics is produced by calculating the Kolmogorov-Smirnov-like sum statistics of constituent genes.	http://gengen.openbioinformatics.org/
GRAIL [134]	SNPs/gene regions	The software tries to search similarities in published text among input genes.	https://www.broadinstitute.org/mpg/grail/
GSA-SNP [135]	SNP *p*-values	Java-based portal implementing *p*-value based pathway analysis	http://sourceforge.net/projects/gsa-snp/
GSEA-P [136]	Gene list of interest	Modified gene set enrichment analysis (GSEA) that calculates the degree of over-representation of input gene list among the most significant genes with improvements in calculating significance levels.	http://www.broadinstitute.org/gsea/index.jsp
GSEA-SNP [43]	SNP *p*-values	Modified GSEA method that allows genotype-based (recessive, dominant or additive) model for calculating SNP-based *p*-values.	https://www.nr.no/en/projects/software-genomics
HYST [137,138]	SNP *p*-values	SNP *p*-values are combined using Simes' test followed by a processing considering LD among the SNPs. PPI-based association then identifies gene pairs in which all genes are disease-susceptible.	http://statgenpro.psychiatry.hku.hk/limx/kgg/
ICSNPathway [139]	SNP *p*-values	Implements an improved GSEA algorithm.	http://icsnpathway.psych.ac.cn/
Ingenuity Pathway Analysis (IPA) [26]	Gene list of interest	Commercial software that utilises its IPA database for calculating edge statistics.	http://www.ingenuity.com/products/ipa
i-GSEA4GWAS [13] and i-GSEA4GWAS v2 [14]	SNP *p*-values or gene list	Modified GSEA with new features in v2 adding functional annotation and analyses	version 1: http://gsea4gwas.psych.ac.cn/ version 2: http://gsea4gwas-v2.psych.ac.cn/
MAGENTA [15]	SNP *p*-values	Modified GSEA	http://www.broadinstitute.org/mpg/magenta/
MAGMA [140]	Raw genotypes	Based on a multiple regression model. Can be extended to gene-environment interaction analysis.	http://ctg.cncr.nl/software/magma
MetaCore [18,19]	SNP/gene list	Calculates pathway *p*-values of constituent genes using hypergeometric distribution based on published pathways.	http://thomsonreuters.com/metacore/
PARIS [141]	SNP *p*-values	SNPs are grouped in linkage disequilibrium (LD) features and single SNP features in LD. Pathway significance are then calculated using these features by permutations.	https://ritchielab.psu.edu/software/paris-download
Pathway-PDT [142]	Raw genotypes	Performs pathway analysis based on raw genotypes in family-based GWAS.	http://sourceforge.net/projects/pathway-pdt/
PINBPA [143]	VEGAS output	A network analysis package implemented in Cytoscape	http://apps.cytoscape.org/apps/pinbpa
PLINK set based tests [20]	Raw genotypes	Enrichment score of input gene set is calculated with user specified SNPs.	http://pngu.mgh.harvard.edu/~purcell/plink/
SET SCREEN test [144]	SNP *p*-values (but raw genotypes required for PLINK)	Approximation of Fisher's statistics to partially account for SNPs' *p*-value dependence due to LD. Implemented in PLINK.	http://pngu.mgh.harvard.edu/~purcell/plink/
Seq2Pathway [145]	Raw sequences	R package for analysis of next generating sequencing (NGS) data. It provides four tests for gene set enrichment.	https://www.bioconductor.org/packages/release/bioc/html/seq2pathway.html/
SNP ratio test [21]	Raw genotypes	Ratio of significant SNPs within vs. outside of pathway is compared using permutation tests.	https://sourceforge.net/projects/snpratiotest/
VEGAS [16]	SNP *p*-values	Calculates gene-based statistics through simulation from multivariate normal distribution of *p*-values, with consideration of LD.	http://gump.qimr.edu.au/VEGAS/
PathVisio [146]		(for visualisation) Platform allowing user submitted plug-ins which can let researchers perform visualisation and pathway analysis.	http://www.pathvisio.org/
Methodologies and Software for multiple "-omics" platform			
Genotyping, eQTL and disease phenotype association [64,65]		Two-stage analysis. First, SNPs associated with gene expression pattern (eSNPs) are extracted. Then, pathway analysis is carried out among eSNPs and disease status to look for expression-associated loci associated with disease.	
DEPICT [72]		Based on expression data in tissues, it can prioritise the most likely causal genes at associated loci, highlight enriched pathways and identify tissues/cell types where genes from associated loci are highly expressed.	http://www.broadinstitute.org/mpg/depict/
GSAA [84]		Simultaneously measures genome-wide patterns of genetic and gene expression variation to identify sets of genes enriched for differential expression and/or trait-associated genetic markers.	http://gsaa.unc.edu

Table 1 (*continued*)

Software	Input	Description	URL/source
iGWAS [85]		Relations among SNPs, gene expression, and disease are modelled within a mediation analysis framework to separate genetic effects due to expression or other factors. Effects for both fractions are then tested.	
SPATIAL [147]		A multi-pathway analysis tool. It contains a set of pathway analysis methods to carry out system-wide analysis that tries to consider inter-pathway interactions and combined signals from multiple pathways.	
Transcriptome-wide association study [70]		Genotype and expression data are first obtained from a small cohort to identify information of strength of SNP-gene expression. This information is then used to impute expression profile of an independent cohort with GWAS data only for predicting gene expression or association of expression and trait.	
WGCNA [82]		Contains a collection of R scripts for building gene coexpression networks. Originally for gene expression data, it can also be applied in other contexts.	https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/
Software for association analysis considering gene-gene and gene-environment interactions			
AprioriGWAS [148]		Software for detecting gene interactions. Consists of two steps of (1) defining the best genotype pattern to be added, and (2) testing the pattern again disease status.	
BOOST [149]		Speeds up interaction test using log-likelihood ratio statistics in both screening and testing stages.	http://bioinformatics.ust.hk/BOOST.html
EPIBLASTER [150]		Difference in Pearson's correlation coefficients is first computed between controls and cases across all possible SNP pairs to flag those warranting further analysis. Pairs that are deemed potentially significant will be tested in a second-stage analysis using the likelihood ratio test.	http://www.mybiosoftware.com/epiblaster-1-0-two-locus-epistasis-detection-strategy-gpu.html
Epi2Loc [151]		Provides a convenient utility for converting, comparing, and interpreting epistatic models with flexibility.	http://cran.us.r-project.org/index.html
FastEpistasis [152]		A multi-threaded epistasis testing algorithm that supports testing interactions in parallel CPU threads.	http://www.vital-it.ch/software/FastEpistasis
INTERSNP [153,154]		Based on logistic regression framework. Allows the selection of various statistical models for analysis. Functional annotations from KEGG is incorporated for analysis. Meta-analysis of results are carried out using METAINTER [155].	http://intersnp.meb.uni-bonn.de/
Software for selecting interacting pairs for gene-gene and gene-environment interactions			
Epi2Loc [151]		An R package that compares two-SNP interaction models for selecting most proper parameters to be included for analysis	
SIXPAC [156]		A search algorithm reporting, with approximation, the most likely interacting SNP pairs	http://www.cs.columbia.edu/~snehitp/sixpac/

methods for RVs have been developed to increase power and reduce the multiple-testing penalty [36,38].

3.1.2. Can we apply pathway analysis methodologies for common variants to RVs?

There are specific methods for grouping rare variants together for analysis. Such methods have been reviewed previously [37,39,40]. One question is interesting in our current context. Given that pathway analysis methods for common variants and region-based analysis methods for RVs are both for aggregating single-variant information, are these methods applicable to pathway analyses for both rare *and* common variants?

To address this, the performances of pathway-based association methods, originally for GWAS, were compared to that of region-based association methods for RVs in a simulated dataset [41]. When common and rare variants were jointly analysed, direct application of pathway analysis software was *not* satisfactory. It was suggested that rare variants should be given higher weighting for better analysis performance [41].

Later, a direct comparison between GWAS pathway analysis software and rare-variant region-based methods was carried out [42]. In this study, a modified version of GSEA-SNP [43] using weighted Kolmogorov–Smirnov (WKS) statistics for gene-set enrichment score (as in original GSEA [44]) was chosen to represent GWAS pathway analysis software for comparison. Meanwhile, four RV region-based association methods were tested, namely weighted-sum test (WSS) [45], simple-sum test (SS) [46], collapsing test in combined multivariate and collapsing (CMC) method [47], and sequence kernel association test (SKAT) [48]. Input variants included 40,918 coding variants from 822 individuals under 1000 Genomes Project [49], after excluding indels and including biallelic variants within annotated pathways in KEGG only [50]. The effects of variants were simulated to depend on two factors: (1) increasing effect with decreasing minor allele frequency, and (2) whether it was one variant of genes from a randomly selected "causative" KEGG pathway. Four scenarios were simulated, which represented combinations of two effect-size models (as assumed by WSS and SKAT in their algorithms) and two different numbers of input causative pathways. Pathway analysis of 1000 simulated datasets was carried out using 11 methods, which included variations using the five methods mentioned above. Power (assessed by the proportion of tests that the "causative" pathway's p-value can pass Bonferroni multiple-testing threshold after correcting for the total number of pathways) and type I error (false positive) rate (number of pathways passing Bonferroni multiple-testing threshold for simulated data with no causative pathway) were evaluated to estimate the performance of the methods. Overall, no single method performed particularly better [42]. Type I error was found to be inflated in most of the pathway analysis methods. However, using all SNPs' p-values for gene-based statistics and then combined with WKS (WKS-variant method) was powerful with moderate type I error in all simulation scenarios. Moreover, pathway-based

methods had higher power than region-based methods, where their power was sensitive to whether or not the effect size of the data matched the methods' assumptions. If the model of effect sizes fitted the software's assumption, region-based software was powerful; otherwise, there could be lack of power. This is consistent with the descriptions in Lee et al. [39] that the power of RV analysis depends on the assumed underlying effects model. Furthermore, using variant-level information for pathway analysis was more powerful than collapsing variants' information into gene units first [42]. This indicates that that one-step analysis may be more powerful than two-step pathway analysis. In brief, while analysing RVs using pathway analysis software is technically feasible, the performance depends on the consistency between assumed and actual model of variants.

3.1.3. What are the pathway analysis software tools that can be applied to common variants and RVs?

Recently, there are software tools developed particularly for pathway analysis of both common and rare variants. The aSPUPath test [51] is a self-contained pathway analysis test modified from adaptive sum of powered score (aSPU) [52], which was originally developed for RV analysis. It can, by incorporating suitable weighting, cater for both common and rare variants. Parameters can be adjusted to modify the assumed direction and proportion of associated variants. This can help increase power by fitting a statistical model closer to the actual situation by which variants confer their effects. Another software tool uses smoothed functional principal component analysis [53]. In this test, genetic variants under consideration are formulated to be represented by a functional principal component score. The difference of the average scores between cases and controls are tested. Smoothed functional principal component analysis has been shown to have better power and better-controlled type I error rates than other common region-based RV analysis software. One reason for the better power is the software's ability to capture all variants' information in constructing the principal component score [53].

3.1.4. What are some challenges and points to note for carrying out pathway analysis of RVs?

Unlike GWAS, which usually cover both genes and inter-genic regions, sequencing studies currently focus more on functional regions of the genome (most notably exons) or targeted regions of particular interest. This is mainly because the cost of NGS is still high and because more deleterious mutations may be present in these regions [39]. Many RV analyses require a weight to indicate the relative importance of each variant during analysis (Table 4), which can help increase analysis power [38,54]. Assignment of weights based on functions for targeted and exome sequencing is easier because functional annotations are more likely known before experiment (Table 4). However, it remains a question as to how the weights are determined for non-coding regions since functions may not be known explicitly, and therefore only minor allele frequency [55] may be used as the most readily available information for determination. Population stratification, as in analysis of common variants, may adversely affect results. This can be captured and corrected by traditional methods used in GWAS (such as principal component analysis) [36,56]. Moreover, meta-analysis methods have also been developed for combined analysis of multiple RV studies [57]. As more sequencing studies are carried out, it will be worth investigating if they are applicable to analysis of both common and rare variants for better capture of genetic architecture for analysis.

Unlike pathway analysis for common variants, for rare variants, the concept of "pathway analysis" (involving variants of multiple genes) and "multiple-variant" analysis (involving variants of the same genes only) is not clearly distinguished now (Table 4). Further investigation is still needed whether multiple-variant analyses within genes can be directly applied to pathway context.

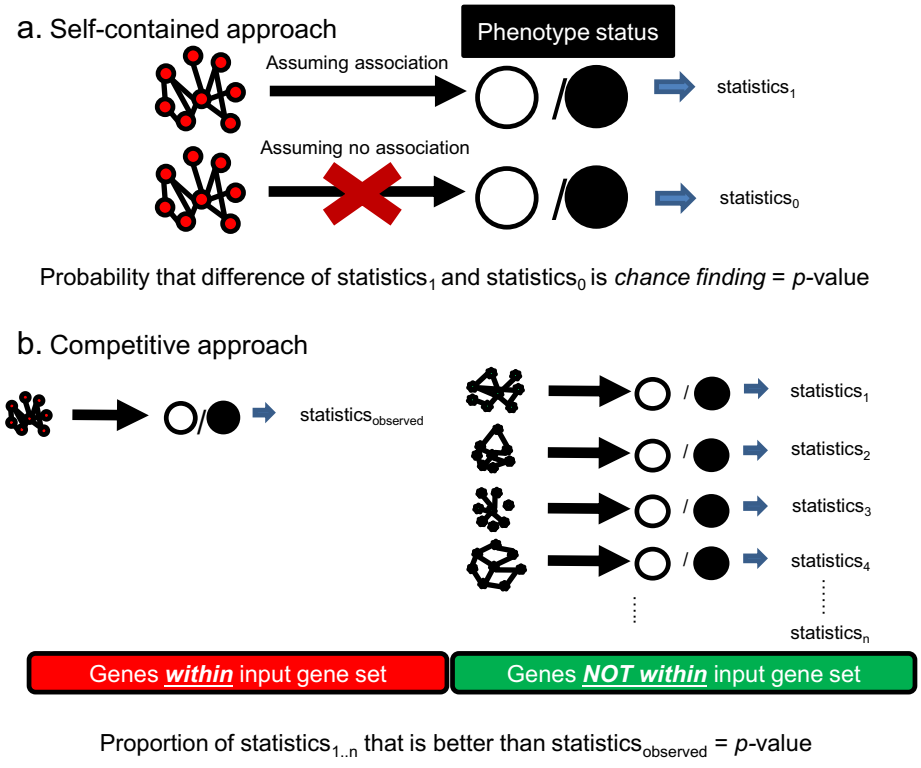

Fig. 3. Self-contained vs. competitive approaches of pathway analysis. For self-contained approach, pathway *p*-value is generated by comparing the statistics assuming there is association versus that assuming there is no association (null hypothesis). Usually, null data are assumed to follow certain standard distributions such as chi-square distribution. For competitive approach, the observed statistics of the gene set of interest is compared with those of gene sets consisting of genes not within the gene set. For competitive approach, data of gene sets not within the current gene set of interest must be available, therefore limiting its use to genome-wide analyses such as GWAS.

Table 2
Common pathway annotation databases. Besides these databases, Pathguide [157] (http://www.pathguide.org/) also provides comprehensive lists of pathway-related databases and resources catalogued according to their nature, such as metabolic or signalling pathways.

Name	Description	URL
BioCarta [158]	Users input research data to construct the knowledge base	http://www.biocarta.com
Gene Ontology (GO) [159]	Large hierarchy of terms representing biological concept	http://geneontology.org/
Kyoto Encyclopaedia of Genes and Genomes (KEGG) [50]	Provides higher-order (genomic and pathway annotations) information from input of molecular data for various organisms	http://www.genome.jp/kegg/
MetaCore [19]	Extensive pathways derived from publications. Allows users to modify pathway elements for illustration purpose	http://thomsonreuters.com/metacore/
MetaCyc [160]	Contains metabolic and enzymatic pathways from various organisms experimentally validated in literature	http://metacyc.org/
MSigDB [136]	Contains a collection of annotated gene sets for use with their gene-set enrichment analysis (GSEA) software. The collection includes various gene sets defined by biological functions, GO, KEGG, positions, sequence regulation information etc.	http://www.broadinstitute.org/gsea/msigdb/
Pathway Interaction Database (PID) [161]	A highly-structured, curated collection of information about known biomolecular interactions and key cellular processes assembled into signalling pathways	http://pid.nci.nih.gov/
REACTOME [162]	Provides a platform for annotating and visualising data from major databases such as NCBI Gene, Ensembl and UniProt databases, UCSC & HapMap Genome Browsers, KEGG Compound and ChEBI small molecule databases, PubMed, and Gene Ontology	http://www.reactome.org/

One interesting question is whether pathway analysis of RVs conveys the same biological meaning as that for common variants. "Partially correct" is a short answer to this question. The argument *against* the statement is that traditionally RVs are believed to have relatively larger effect size. Therefore, once an associated RV is identified, it is likely that the identified locus is already causative [58]. From this perspective, pathway analysis is not necessary for RVs. However, this is not entirely the whole picture for common diseases because some RVs only exert medium or small effects [59]. To investigate this, Kryukov et al. tried to estimate the proportion of mildly deleterious missense mutations as well as their fractions among human RVs in various variation datasets [60]. It was found that over half (53%) of all *de novo* missense mutations are mildly deleterious. Moreover, the majority (52–71%) of amino acid substitutions with observed frequency $<1\%$ are also mildly deleterious. The combined findings show that low-frequency missense mutations are deleterious. In addition, it has been estimated that the majority of rare missense polymorphisms in humans have small selection coefficients (0.001–0.003). This suggests that the purifying selection acting on them (i.e. the removal of these polymorphisms from population) is relatively mild. Therefore, these rare mutations can accumulate in the population, resulting in a highly heterogeneous spectrum of individual alleles with very low frequencies [60]. A better model to study genetics of such phenotypes would be to consider the cumulative frequencies of all RVs (instead of individual ones) in interested genes and compare them between cases and controls. Previously, the high cost of sequencing and the difficulty in selecting deleterious missense RVs (instead of neutral ones) had limited the use of the method [60]. With the advancement in technologies, the cost of sequencing keeps going down. Meanwhile, pathway analysis methodologies may help differentiating deleterious RVs from neutral ones. For example, *using pathway analysis approach*, structural differences were identified in multiple genes involved in signalling networks controlling neurodevelopment [61], and this approach identified rare structural variants in neurodevelopmental pathways to be associated with schizophrenia. This example demonstrated that rare variants involving multiple genes could be discovered using pathway analysis approach.

Table 3
Disease study examples that adopted pathway analysis methodologies.

Disease	Reference	Software used
Alzheimer's disease	[163]	WebGestalt
Esophageal squamous cell carcinoma	[164]	ICSNPathway server applies the i-GSEA (improved GSEA)
Multiple sclerosis	[31]	Cytoscape
Olfactory behaviour (*Drosophila*)	[165]	R spider
Biliary cirrhosis	[166]	LRT (first stage), i-GSEA4GWAS (second stage)
Bladder cancer	[167]	GSEA and ARTP, using union of the results
Bipolar disorder	[168]	IPA (first stage), GSEA-SNP (Second stage)
Parkinson disease	[169]	ALIGATOR and GSEA
Schizophrenia	[170]	ICSNPathway server applies the i-GSEA
Major depressive disorder	[171]	GSEA, hypergeometric test, sum-square statistic and sum-statistic.
Schizophrenia	[172]	MAGENTA, ALIGATOR, INRICH and Set Screen
Coronary heart disease	[173]	Variable set enrichment analysis (VSEA) in genome-wide association studies
Preterm birth	[33]	GSEA (gene set enhancement analysis)
Kawasaki disease	[174]	MetaCore
Testicular germ cell tumour	[34]	iGSEA4GWAS, MAGENTA, GSA-SNP

3.2. From genetics and beyond: multi-omics analysis and non-genetic analyses

One goal of genetic studies is to predict the outcome of a disease or phenotype. However, when genetic information is passed from DNA to RNA and then to protein through the central dogma of molecular biology, variable factors may interfere with the intermediate steps and therefore affect the final outcome. These factors could be "intrinsic", i.e. regulatory events that happen inside an organism without external stimuli, such as post-transcriptional and post-translational modifications or gene-gene interactions. The factors may also be "extrinsic", where environmental factors and external stimuli play important roles. In this section, how information other than DNA genotype data may be integrated with genetic pathway analysis will be briefly discussed.

3.2.1. Why do we need multi-omics data analysis?

Because of the complexity in biological systems, integrating information of multiple "-omics" platforms (including genomic, transcriptomic, proteomic and, more recently reactomic data) can provide extra insight into how genetic information is conveyed to the formation of phenotypes [62]. Although the idea of pathway analysis methods for GWAS originated from analysis of expression data, traditionally data of different "-omics" platforms were analysed separately. Recently, because of the availability of high-throughput expression and proteomic data, data integration has gained much attention.

3.2.2. How integration is done? What software packages can be used?

Integration of genetic and other data can be divided into "multi-stage" and "meta-dimensional" approaches [63]. For multi-stage analyses, two different types of data (e.g. genotype and expression data, expression and phenotype data, etc.) are considered at each stage. A linear pipeline that uses results from a previous analysis step carries out integration of data. On the other hand, meta-dimensional analyses

Table 4
Comparison of pathway analysis methods for common and rare/low-frequency variants.

Aspect	Common variants	Rare variants*
Frequency of minor alleles	Common (≥5%)	Rare (<1%) / low-frequency (1–<5%)
Role of pathway analysis	Usually secondary to single-marker analysis	Usually as "conventional" testing as single-marker analysis usually does not have enough power
Importance of weights of individual variants for analysis	Optional as usually density of SNPs is not high and the number of variants per gene is smaller. Dilution effects of non-causative variants are not as obvious	Crucial as there is high density of variants. More obvious dilution effects may be resulted from non-causative variants, and therefore weights can help reducing such effects.
Availability of functional annotations	Not readily available for non-exonic regions.	More readily available as study designs mostly focus on regions with known functions. However, for whole-genome sequencing, functional data may not be available for non-exonic regions.

* The boundaries of frequency for rare and low-frequency variants vary in the literature.

try to combine all data types and predict phenotype outcome using the combined data in one step [63].

3.2.3. Examples of multi-omics data analyses: coexpression analysis

One good example of integration between genetics, gene expression and phenotype outcome is obesity. Emilsson et al. [64] tried to explain this using a two-step approach. First, they analysed over 23,000 transcripts in blood and adipose tissue in 470 individuals to look for expression traits, i.e. gene transcripts with good correlation with clinical phenotypes for obesity. Then, linkage analysis was carried out using 1732 microsatellite markers near to genes corresponding to the transcripts from the same individuals to estimate "heritability" of the expression traits. It was found that expression traits with high heritability in blood and adipose tissues were highly reproducible between the two tissues. For the expression traits that were within the top 25th percentile for heritability in blood, 70% of them had a significant cis-acting expression quantitative trait locus (eQTL) in both adipose tissue and blood. This showed that expression of genes had a high genetic component. This study is important because it linked gene expression with clinical phenotypes, where such evidence was previously given by studies of cell lines only.

Later, Zhong et al. tried to achieve integration for type 2 diabetes using pathway analysis approach [65]. They first obtained SNPs associated with gene expression (eSNPs) in 707 liver, 916 omental adipose and 870 subcutaneous adipose tissues. A total of 20,563 eSNPs were identified in 9,964 genes. Association of these eSNPs with type 2 diabetes phenotype was then assessed using a GWAS of over 3,400 individuals, after imputing eSNPs present in expression analysis but not in the GWAS. Pathway analysis using modified GSEA [44] was then used to identify significant pathways with representing eSNPs. Nine pathways were finally identified, which were successfully validated using an independent cohort [65]. This pipeline of analysis has further shown that integration of genetic and expression data is possible with the use of pathway analysis. Similar approaches have been adopted for other phenotypes, including basal cell carcinoma [66], allergic rhinitis [67], coronary artery disease [68] and blood pressure [69].

This idea was extended by Gusev et al. for transcriptome-wide association study [70] (Table 1). In short, both genetic and gene expression data were available from a small set of individuals. In a larger set of individuals with GWAS data only, expression data were obtained by imputation, and association between imputed expression data and phenotype was then carried out. The main advantage for this approach is that expression data is hard to obtain for all samples under study. This study will allow expression-phenotype association analysis with expression data being generated using an indirect approach. Using this approach, 69 loci significantly associated with obesity-related phenotypes were found [70].

Recently, Locke et al. [71] carried out a large-scale GWAS of body mass index (BMI) using nearly 334,000 individuals. In this study, 97 significant loci were successfully identified. Different sources of evidence were used to identify significant SNPs associated with BMI. These sources included genes having or close to significant SNPs, results from pathway analysis software DEPICT [72] and MAGENTA [15] (see Table 1), cis-eQTL and literature search to identify overlapping SNPs. They have successfully found overlapping pathways, including those related to central nervous system, obesity, insulin secretion and/or adipogenesis.

3.2.4. Gene coexpression network (GCN): undirected coexpression networks. How does it help in identifying functions of non-coding RNA?

Gene coexpression networks (GCN) and gene regulatory networks (GRN, see Section 3.2.5) are related and yet conceptually different types of networks. Both networks consist of edges that connect genes with certain "relationships". In GCN, this relationship refers to the coexpression pattern observed between two (or more) genes. An edge can be established when the correlation of the genes' expression (represented by statistical correlation measures such as Pearson's correlation) exceeds a defined threshold. This simple definition does not imply any causal relationship. In other words, GCNs are *undirected*. On the other hand, GRNs describe the explicit causal relationships of developmental processes [73]. GRNs explain how genomic sequences can regulate the expression of a set of genes, which in turn gives rise to the collective developmental pattern and state of differentiation.

GCN is a versatile and powerful method. For example, it was used to investigate the conservation of gene expression patterns among different organisms. In Stuart et al.'s study, GCN was used to study the expression patterns of humans, flies, worms, and yeast [74]. First, 6307 "metagenes" were defined using gene sets with similar protein sequences across the different species. The aim was to find out pairs of metagenes with coexpression. To achieve this, the coexpression of each pair of genes between two organisms was represented by Pearson's correlation. The correlations of all genes were ranked. A probabilistic method was then used to determine how likely to see the combination of ranks across all organisms by chance. Connected by 22,163 edges, 3416 metagenes were obtained using a p-value cutoff at 0.05. Five metagenes with previously unknown functions were selected for investigation of their biological functions using information from their GCNs. These metagenes showed conserved coexpression with the genes involved in cell proliferation and cell cycle. Biological experiments confirmed the functions of the metagenes in cell proliferation and cell cycle. This example shows that GCN constructed across multiple species can be used to infer functions of genes with previously unknown functions in addition to coexpression patterns.

Depending on the context of transcripts used for building the networks, GCN can also be extended to study the functions of non-coding transcripts. In Yao et al.'s study, GCN was used to study enhancers expressed in the brain and their gene targets [75]. Enhancers are non-coding DNA sequences that can carry out regulatory functions. Active enhancers have signature chromatin marks. Their transcription results in non-coding enhancer RNAs. In this study, 908 enhancer regions were first identified using RNA-seq of cell and tissue samples. Of these, 673 were intronic/intergenic. By comparing RNA-seq results from adult human frontal, temporal and occipital cortices, and cerebellum, 131 brain-expressed enhancers (BEEs) were identified and 103 of these, defined as robust BEEs (rBEEs), were found to overlap with enhancer-specific histone marks H3K4me1 or H3K27ac. In order to locate the targets of rBEEs, a GCN was constructed between rBEEs and gene expression data of the brain. The authors drew several conclusions from

the GCN. First, out of all 19 coexpression interaction modules found, 12 showed brain region-specific or developmental stage-specific expression. Most obvious variation in spatiotemporal gene expression occurred in the transition from fetal to postnatal brain. Moreover, the largest GCN node contained genes more highly expressed in fetal brain than in all regions of adult brain. Other GCN nodes were specific to brain regions. This indicated the importance of brain enhancers in regulating the stage of brain development. Second, in the GCN modules, there was higher topological overlap consisting of rBEE-closest gene pairs. This indicated that rBEEs were more likely to coexpress, and therefore regulate nearby genes. Third, among all the top genes coexpressed with each of the rBEEs and also located in cis (within 500 MB) with the corresponding rBEE, there were genes related to neuronal differentiation and autism spectrum disorders. This indicated rBEE's targets identified by GCN had functional relevance to brain cell development and brain-related clinical phenotypes.

One potential use of GCN of RNA-seq data in non-coding RNA is the annotation of long intervening non-coding RNAs (lincRNAs). Recently, a protocol was introduced to identify lincRNAs and to characterise their functions using a GCN [76]. The GCN integrates the expression of protein-coding and lincRNA genes. In short, lincRNAs were first identified using coding-noncoding index (CNCI), a tool that catalogues coding and non-coding sequence features of different species. Functions of the lincRNAs were then predicted using ncFAN. ncFAN first tries to construct a GCN between lincRNA and protein-coding genes. Then, according to the functional terms annotated for the coding genes connected in a certain hub, the function of the hub can be predicted. This example suggests another possible application of GCN in predicting the functions of non-coding sequences.

One noteworthy point is that previous expression profiles were captured mainly by microarrays. Because of the advancement in NGS, expression profiling using RNA-seq has become more popular. The debate of whether using RNA-seq or DNA microarray is beyond the scope of this paper although both technologies have their strengths and weaknesses [77–80].

3.2.5. Gene regulatory network (GRN) – integrating genomics and proteomics data

GCNs are undirected networks that only show coexpression pattern. However, if we have both protein and expression data, it is possible to construct *directed* gene regulatory networks (GRN), which is able to explain more about the causal relationship between genes. This idea was used for identifying GCNs and GRNs in maize [81]. The study included three types of datasets for 23 tissue samples spanning across the vegetative and reproductive stages of maize. The three types of data included messenger RNA sequencing (mRNA-seq), electrospray ionization tandem mass spectrometry data of unmodified protein, and that of phosphorylated protein. Weighted gene coexpression network analysis (WGCNA) [82], an R package for construction of GCN, was used to discover similarly expressed genes. In total, 36 genes with similar expression patterns in at least 4 tissues were discovered. The phosphorylation patterns of genes were similar to their mRNA profile. Their phosphorylation also occurred in tissues known to be related to developmental phenotype. These suggested that the phosphorylation of these proteins was important in determining their functions. GRN, a directed regulatory network, was then used to further explore expression pattern of genes together with protein data. GRNs were constructed by observing the expression correlation between mRNA, protein, and phosphoprotein expression profile of transcription factors (TFs). It was found, using data from two previously validated TFs (KN1 and Opaque2), that GRNs constructed using protein data predicted target genes better. When this method was extended to all TFs, it was found that different data sources resulted in disparate GRN predictions. Using combinations of the data sources to build GRNs were found to have better predicting power than single-input GRNs. This study provides an example of extending analysis of genomics to proteomics data, as well as how this enables the direction of gene regulation to be discovered.

3.2.6. Some software packages

In the earlier BMI example (see Section 3.2.3), the evidence shows that BMI may be related to the control of appetite [71] because DEPICT [72] also identified brain tissues (which is related to appetite) to be a related tissue enriched in the dataset. This example has shown that integrating different sources in a meta-dimensional manner can deduce possible pathways related to a phenotype. DEPICT [72] was built on the data from a cancer expression study that tried to investigate the relationship between copy number and expression level in cancer cells [83]. In the software, there are 14,461 "reconstituted" gene sets that capture gene sets with similar functions and expression patterns. The set was curated based on the expression pattern of 77,840 samples and the functional annotation of constituent genes. In addition, tissue/cell type enrichment was carried out using another set of 37,427 microarrays of human tissues/cells to determine if genes are highly expressed

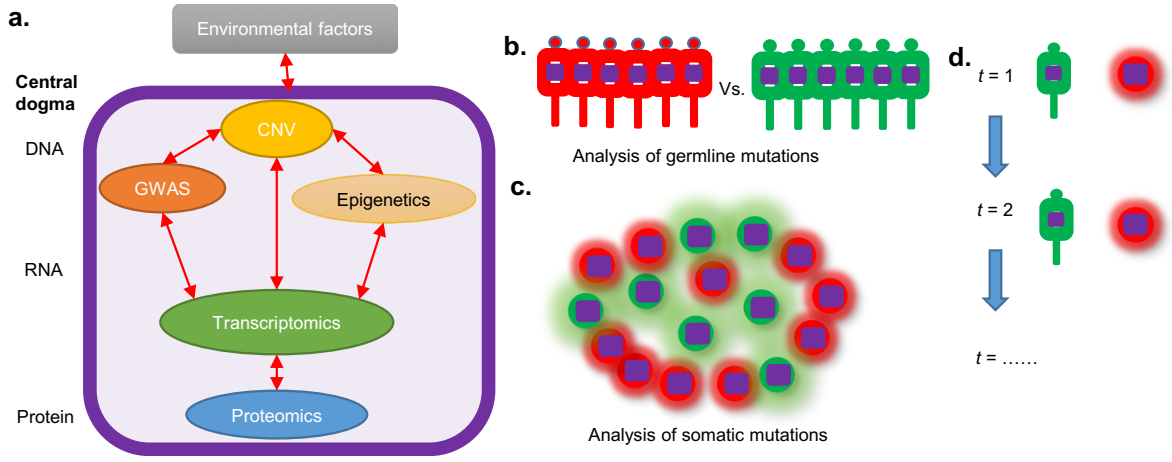

Fig. 4. Analysis paradigms for future pathway analysis. a. The expected information that can be integrated for pathway analysis. Basically, all levels of multi-omics data are expected to be combined for analysis. b. Traditionally, pathway analysis was carried out for germline mutations. c. Single-cell methodologies enable analysis of somatic mutations. d. One possible factor to be included for pathway analysis includes time point, which can capture changes in the body in response to external stress such as viral infections.

in any of the tissue and cell type annotations of the human subjects. Using this information, DEPICT can help identify genes in associated loci using input SNPs, "reconstituted" gene sets enriched in genes of associated loci, and tissue/cell types implicated by the associated loci [72].

Another integrated analysis software is Gene Set Association Analysis (GSAA) [84] (Table 1). The software tries to carry out gene set association analysis using both GWAS and expression data. GSAA is based on multi-layer association tests of gene expression and genetic association data. First, a single SNP score is produced using one of the five methods provided by the software. Then, a SNP set is defined as the SNPs within a gene and the gene's flanking region. SNP set association score of the region is given by the maximum single SNP score among all SNPs in the region. Second, a gene expression score is calculated (using one of the four methods in the software) using the difference in the means of expression between the phenotypic classes divided by standard deviation. Third, a gene association score is given by combining the SNP set association score and gene expression score using either Z-score sum, Fisher's method or rank sum method provided by the software. Finally, Kolmogorov-Smirnov test is used to determine which gene sets are associated with the phenotype most (by testing gene scores of constituent genes in the gene set). It was found that, in Crohn's disease data, GSAA was able to report more significant pathways than GSEA, which only uses gene expression data [84].

iGWAS [85] (see Table 1) is a method that uses mediation [86] to model effects of SNPs and gene expression on disease phenotype. In brief, the authors previously tried to model total genetic effects, i.e.

Table 5
Issues and solutions for pathway analysis.

Questions/issues	Suggested solution	Challenges in tackling	References
Analytical issues			
1. How to summarise SNP-level effects? 2. How to evaluate pathway-level statistics?	1. Researchers should compare and report the ability of various methods to identify the specific disease model each software performs best 2. Suggested information to be included are a. Source of gene set information b. Statistical methods and potential bias/limitations c. Report SNP- and gene-level effects, original GWAS results, and individual gene contribution d. SNP- and gene-level statistics e. Any possible gene set overlap, examination/-correction of crosstalk effects of gene sets, if any f. Biological context as supporting evidence 3. Biological Connection Markup Language (BCML), a standard method to report pathway information, was proposed		[29,175]
3. How to produce a standardised dataset to confirm/compare pathway analysis results and compare among methodologies?	There is no consensus drawn yet. There are both pros and cons for using biological system and standard simulated datasets as reference dataset.	Using well-studied biological data as reference dataset for replication can be better than using simulated data as the end-results can be accurately known. However, *how* the actual biology (i.e. the intermediate steps) leads to the end results cannot be observed. Also different gene set definitions can attribute same results to different pathways, which makes interpretation of results difficult	[10,28]
4. How should we adjust for covariates in pathway analysis?	For two-step analyses, *p*-values may be adjusted at the SNP *p*-value stage.	1. Not easy to carry out adjustment for software using raw genotypes as input due to methodology constraints 2. The dependence between covariates and pathways is not always known.	[5]
Curation issues			
1. What is the best way to include intergenic variants to genes?	New area of future work on how to include non-genic variants.	Including too many SNPs will increase multiple-testing burden. However, including too few may exclude causative variants.	[29]
2. How should we include disease and cell condition-specific information in analysis? How to improve annotation data integrity of the experiments?	1. Development of disease-specific databases 2. MIAME-like report in standard format to facilitate report of experiments		[9,10]
Other areas			
1. How can we take both genetic and non--genetic risk factors into account?	Include environmental data into consideration in predicting diseases	Commonly agreed methods in measuring non-genetic risk factors are required	[107]
2. How to account for temporal effects due to environmental change and response to stimuli?	Collection of data at different time points to capture dynamics of the genome	Need comprehensive biological data	[10,131]
3. How to extend to rare-variant analysis?	Open area of research. Studies show that rare-variant region-based analysis methods may be used for pathway analysis. However, assumption about the disease model must be carefully made before the methods can predict associations accurately. Therefore, research is needed to develop methods of how to analyse rare variants without such assumptions (as very often the actual disease model is not known prior study).	Different people can have different haplotypes, complicating situation	[176]
4. Stability of results – Are results replicable for different individuals (due to genetic heterogeneity)?	1. Study gene sets instead of individual markers 2. Using genes instead of individual variants as units of analysis		[28]

Box 2
Challenges for pathway analysis.

Analytical aspects

In terms of analytical aspects, the most important issue is the lack of reproducible results between methods. This is mainly because of the lack of consensus on summarising SNP- and gene-level statistics. There is no standard protocol or procedure in analysis because of the wide variety of analysis methodologies and gene set definitions. Therefore, studies should report enough information for users to understand the strengths and weaknesses of analysis and to choose the most appropriate software. Biological Connection Markup Language (BCML) was proposed for reporting pathway analysis [175]. This may provide a starting point for achieving more information reporting of pathway analysis data.

To compare or confirm among pathway methodologies in a fair manner, the availability of a standard replication dataset is important. However, there is no rule of thumb whether it is better to use simulated dataset or a standard biological system. Replicated dataset can be generated by simulation of random disease outcomes or random permutations of samples [28]. On the one hand, simulated/permuted data are more homogeneous because the underlying statistics for simulation or methods in permutation are based on theory. On the other hand, simulated/permuted data cannot represent all the biological events happening in a real organism. Therefore, there is still value to carry out experiments to obtain biological data for comparison of analysis methods [10]. While the final effects of certain hypotheses can be observed in the end, *how* the intermediate biological events are produced is still extremely hard to be discovered fully [10]. Therefore, confirming pathway analysis results remains a challenging mission.

Currently, many software tools do not accept adjustment for covariates in the analysis [9]. Algorithms have been developed to deal with this issue, such as supervised principal component analysis [184] and regression-based method [185]. However, covariates are still mainly controlled at the single-variant analysis stage in two-step analysis, where covariates are adjusted in producing single-variant *p*-values. It is more difficult to carry out covariate adjustment for raw-genotype-based methodologies because these methods cannot handle covariates, or strong assumptions have to be made about the independence between the covariates and the pathways [5].

Curation aspects

It still remains an issue how best to map variants into genes. Meanwhile, expression of genes and their consequences may be different between diseases and in normal tissues, different tissue types and at different time points. Therefore, it is useful to incorporate other information sources to report tissue- and disease-specific data [10] (see Section 4 "future perspectives").

Including tissue-specific data can improve quality of pathway analysis [176]. Because of the diversity in experiments, reporting how databases are curated will allow users to understand strengths and limitations of the data. The standardisation of reporting experimental data has been suggested for expression microarray data for pathway analysis [10] via the framework of minimum information about a microarray experiment (MIAME) [186]. MIAME describes six elements of microarray experiments that will allow researchers to understand results of experiments. Note that using such standard as MIAME should not limit researchers what to report according to particular methodology. Instead, it should encourage detailed catalogue of methodologies and results such that interested groups can understand how results are generated and how results are concluded by proper description and annotations [186]. This will finally help interpret, and ultimately combine results in a more meaningful and understandable manner.

the total effects of SNPs and gene expression, on a phenotype [87]. iGWAS extended the method by adopting counterfactuals [88] to separate total genetic effects on phenotype into two components: one that can be mediated through gene expression (mediation effect), and the other not mediated through gene expression (alternative effect, effected through environmental factors, for example). iGWAS can test the association of both components using an omnibus test. With asthma as an example, iGWAS has been found to be able to confirm previously reported associated genes [85].

Another example is weighted gene coexpression network analysis (WGCNA) [82]. WCGNA is widely used for the construction of GCNs. It consists of a collection of R scripts for different stages of building networks, including network construction, module detection, and calculations of topological properties [82]. Compared with Bayesian networks, WCGNA requires less time and fewer samples for training. Besides, when compared with GSEA, there is no need of a priori information as input. Multiple-testing threshold can also be alleviated because WGCNA only considers a subset of edges for network construction [89]. While originally designed for expression analysis, it has the potential to be used for other data type too [82].

GWAS results can be extended by making use of information about metabolism, known as metabolomics GWAS (mGWAS; see Section 3.2.8), to infer consequences of genetic variants at the metabolite level [90]. Software tools, such as iPEAP (integrative Pathway Enrichment Analysis Platform) [91], are available [92] for applications from "traditional" GWAS [93] to state-of-the-art NGS experiments [94], which suggests the promising prospect of analysis in the area.

3.2.7. What are the challenges in multi-omics analysis?

Although integrating data from multi-omics platforms can provide insight into the relationship between genes and phenotypes, there are issues to be resolved. Firstly, while it is relatively easy to obtain genotype data (through extracting DNA from participating subjects), expression and metabolomic data are more difficult and costly to obtain. Moreover, while obtaining both genotypes and gene expression profiles for the same individuals would be best for building disease prediction models, it is very hard to have a large sample size. Some software tools were developed using their own genotype and expression data (such as DEPICT [72]) so that users can carry out analysis without their own expression data. However, if the disease/tissue to be analysed is not in the default database of such software, producing expression data with experiments is still an inevitable step. Furthermore, the relationships between different levels of data may not be linear [95], which may render the use of simple regression models not applicable or more difficult. Further development in statistical tools may help capture such non-linear relationships [95]. Overall, increased sample size should be the most direct approach to improving the power of a study. However, validation using independent sample set, or cross-validation using sub-groups of dataset in hand may also help improve predictive ability with limited numbers of samples and resources [63].

3.2.8. Integration of data in other functional "-omics" data

While bioinformatics analysis using information from the central dogma (genomics and proteomics) allows us to understand our biology, there are yet other levels of information that allow us to further understand the molecular biology in the body. Here we wish to briefly go through a few examples of glycomics and metabolomics analysis.

DNA sequences determine the sequence of a protein. However, the structure and function of the protein can be modified by a very complex process of glycosylation involving regulation of many genes [96]. There are a number of modifications that can affect the structure of glycoproteins, and can each be represented by one level of information. These levels include glycogenomics (genomics of genes and enzymes involved in glycosylation pathways), glycoproteomics (study of glycosylated proteins and their glycosylation sites) and glycomics (identification of glycans present) [97]. Integration of various sources of data is still at an

early stage because there are still several hurdles in data analysis (e.g. how to format input data, and lack of analysis tools) [97]. Despite this, there are a few pioneering studies in glycomics. For example, Brennan et al. tried to analyse mass spectral and gene expression data [98]. In the study, they compared the glycosylation patterns of androgen-dependent and androgen-independent lymph node carcinoma of the prostate (LNCaP) cells. They utilised several layers of mathematical rules to generate networks to infer the abundance of glycan structures. An increase in H type II and Lewis Y glycan structures in the androgen-independent cells and the corresponding elevated activity of a fucosyltransferase (FUT1) could be found. However, this could not be found by single-stage analysis [98]. This example showed that a systems biology approach combining expression and mass spectrometry data could be used to discover novel findings.

Another possible multiple -omics application is in metabolomics analysis. Metabolomics is the study of chemical traces of the cell during certain cellular activities. Expression-based analysis allows the observation of the molecules present in the cells [99]. By incorporating genetics, hopefully the variations that causes such dynamics could be predicted. This is achieved by GWAS with metabolic traits, or mGWAS [100]. Similar to genetic studies, detection of metabolic traits can be divided into targeted or non-targeted methods. Targeted methods are mainly based on mass spectrometry (MS) [101] while non-targeted methods use both MS and nuclear magnetic resonance (NMR) [102]. The first mGWAS is a study of metabolite profile in serum [103]. The study included a GWAS analysis of 363 metabolites in 284 males. Four significant variants coding for enzymes were identified, where the corresponding phenotype matched the pathway in which the enzymes were involved [103]. For lipidomics, Hicks et al. carried out a GWAS with sphingolipid traits. Lipids were quantified using electrospray ionization tandem mass spectrometry (ESI-MS/MS) [104]. Thirty-two variants passed genome-wide significance threshold. The strongest signal spanned across 7 genes that function in ceramide biosynthesis and trafficking [105]. Another example is a GWAS to identify genetic risk factors for polyunsaturated fatty acids [106]. Variants of the *FADS* cluster showed the strongest association with plasma fatty acid concentration, and also a second strongest locus in *EVOVL2* associated with longer chain n-3 fatty acids [106].

3.3. Analysis of interaction and environmental factors

3.3.1. Environmental factors and interaction analysis

Besides genetics, environmental factors and gene-gene (G-G) interactions also play a crucial role in the aetiology of a phenotype. In fact, the manifestation of a disease or phenotype can be viewed as the interplay between genomics, epigenomics, and environment factors (extrinsic factors such as behaviour or stimuli from living environment) [107]. Studying genetic (as well gene expression and/or proteomic) data together with environmental variables will help us understand how the body responds to changes in external conditions. Other papers focus on study design and analysis methods [108–110]. Here, we wish to focus on giving a brief idea of how gene-environment (G-E) interaction is considered in GWAS with incorporation in the context of pathway analysis. Users may wish to refer to other references for interaction in candidate gene studies [111,112] as well as experimental designs [113].

3.3.2. Pathway analysis and gene-environmental interaction

Previously, G-E interaction analysis mainly focused on candidate gene regions with suggested functions (hypothesis-driven) [110]. One example is the interaction study between Y402H (a common coding variant in the complement factor H gene) and lifestyle factors in age-related macular degeneration [114]. It was found that individuals having the CC genotype of Y402H and higher BMI or smoking conferred the greatest risks. With the emerging number of GWAS, analysis of G-E interactions has also evolved from hypothesis-driven candidate approach to genome-wide scale (non-hypothesis-driven). Studies that carry out G-E interaction analysis on a genome-wide scale, i.e. genome-wide interaction studies (GEWIS/GWIS), can be viewed as an extension of GWAS. GEWIS has been carried out for several diseases. For example, a GEWIS of asthma investigated the interaction between genetic variants and two environmental factors, *in utero* and early childhood tobacco exposures, in 2654 cases and 3073 controls [115]. In this study, a logistic regression model containing independent variables representing genetic effects, tobacco exposures, and an interaction term was used for analysis. Variants in *EPB41L3* and *PACRG* were found to be the most significant after considering interaction with *in utero* and early childhood exposures, respectively. In a GEWIS of myopia [116], a joint meta-analysis of interaction between genetics and education level in refractive error was performed. It was found that three variants in *AREG*, *GABRR1* and *PDE10A* have strong evidence of interaction with education among Asian cohorts.

Studies of G-E and G-G interactions both involve a large number of multiple-testing penalties due to the huge number of tests for all possible interacting pairs. Some software tools can help us select more probable interacting gene pairs (Table 1). Gene-based and pathway-based interaction analyses can improve the power of GEWIS by combining signals within functional units. This can be done in a multi-stage approach, where GEWIS is carried out in the first stage, followed by gene-based and pathway-based analyses. One example is a GEWIS of lung cancer investigating the disease susceptibility in relation with asbestos exposure [117]. This study included over 300,000 SNPs with over 1100 cases and controls. Three level of analyses, namely single-variant level, gene level and pathway level, were carried out. In single-variant level of analysis (using *p*-value of interaction term reported by PLINK) and gene-level analysis (using VEGAS [16]), no significant results were found. However, in pathway-level analysis using i-GSEA [13], Fas signaling and antigen processing pathways, which are related to apoptosis and immune function regulation respectively, were found to be significant [117]. This study illustrates a relatively simple approach of how GEWIS can be combined with pathway analysis methods for the discovery of novel disease pathways.

3.3.3. Challenges in studying environmental factors and future directions

Indeed, taking cancer as an example, genome-wide G-E interaction study is only at its start-up stage, and many G-E interaction studies still adopted a candidate-gene approach [118]. The main challenge in studying G-E interaction is data collection. As in the case for multi-omics analysis, there are very few collected datasets large and wide enough for comprehensive interaction analysis [107]. This situation is slowly improving with the progress of Environmental Genome Project (http://egp.gs.washington.edu/) [119] and Toxicogenome Project [107]. How an individual develops an illness can be considered as how he/she responds to the environmentally induced stress, given the person's genetic background. Therefore, the Environmental Genome Project aims to look for genes that are related to environmentally associated diseases, and then carry out functional studies to validate the results *in vivo*. Moreover, one disease can occur simultaneously with another, a phenomenon known as comorbidity [120]. By analysing environmental interactions together, how an organism is "unwired" (i.e. to respond to external environment by changing its own metabolism) could be more clearly understood [120]. We hope that the information can be utilised to advise authorities to improve health policies. For example, if certain lifestyles are related to higher occurrence of certain diseases, preventive policies can be made to advise the public to prevent such activities in order to promote public health [107] – a paradigm shift towards precision medicine [121].

4. Further perspectives

One trend for pathway analysis is its application to other "-omics" data (Fig. 4). For example, copy number variation (CNV) data, an example of structural variations, can be used to carry out combined analysis

with expression data. In a meta-analysis of cancer transcriptomes, CNV was compared with expression data to infer *trans*-acting gene sets [122]. The correlation data were then used to look for enriched pathways to further explain the possible functional consequences of the results [122]. Another possible data type for analysis is epigenetics data [123], which include the methylation patterns of DNA. One software tool for pathway analysis of epigenetic data is LRPath [124]. It can report enriched biological concepts (from information of a database containing annotations from external sources) in input methylation data and compare methylation profiles from multiple experiments. With the expanding amount of data, the demand for software tools for multiple data types will also increase.

In addition, the expression of genes within the body is dynamic. One possible way to capture such information is to take expression measurements and analysis at different time points (Fig. 4). For example, Stanberry et al. tried to study the gene expression patterns of a person during two episodes of viral infections [125]. The study used the approach of integrative personal omics profile [126], which tried to connect dynamic, longitudinal multiple -omics data with disease status. In the study, clusters of genes having similar temporal expression patterns during viral infections could be identified. This suggested that integrating different -omics data might help model the dynamics of biological systems [125]. Moreover, traditional genomic analysis lacks cell- or tissue-specific data [10]. Recent technological improvements in whole-genome amplification and NGS have made single-cell sequencing possible [127]. This would enable us to understand genes that have effects on cell state, and possibly predict cell fate [128]. This is particularly important for cancer studies because intra-tumour heterogeneity exists among cells of the same individual [129]. Only with single-cell assays can variations in genomes among cells be detected. This has proven to be successful in breast cancer [130]. With pathway analysis, not only the genetic variations among cells could be known, but also clues to the functional background of how this happens and its consequences could be discovered.

Further improvement of accuracy of pathway analysis requires more comprehensive and replicable functional annotations, experimental data and phenotype data. To obtain replicable functional annotations, it has been suggested that functional data be reported in a standard format with minimal information, as suggested by Biological Dynamics Markup Language (BDML), a format for reporting dynamic data [131]. Optimistically, the number of replicable pathway studies will be increased by improving the quality in reporting results of experiments. However, it should be noted that even with pathway analysis, functional analysis is still needed to confirm the actual genes and variants that exert the most important effects.

Finally, we briefly summarise the issues in pathway analysis and suggested solutions in Table 5, and some issues are deliberated in detail in Box 2. This serves to inspire the readers for future development in this area.

5. Conclusions

This concise review discusses different factors to be considered in carrying out pathway analysis for GWAS to analyse complex diseases, as well as how pathway analysis could be extended to rare variants, and the possibility of including other "-omics" data and taking interaction into consideration. Along with the advancement in -omics technologies are the large amounts of data generated from multiple platforms. One strength of pathway analysis is its ability to integrate information from different sources, as well as reducing dimension of analysis into meaningful units so that the power of analysis can be improved. We foresee that pathway analysis of complex disease will be "multi-dimensional", where "-omics" and environmental factors will be considered simultaneously in analyses in order to model disease mechanisms more accurately. By learning both "intrinsic" genomic factors and external environmental factors causing diseases, better health strategies for personalised healthcare, and precise medicine based on a person's genetic and exposed environment backgrounds for the prevention and treatment of disease could be invented.

Transparency document

The Transparency document associated with this article can be found, in the online version.

Acknowledgements

This work was supported by grants from the Research Grant Council of Hong Kong [PolyU 5637/12M] and the Hong Kong Polytechnic University [87TP, 87U7 and G-YBK2]. M.K.H.Y. was also supported by the Endowed Professorship Scheme (KB Woo Family Endowed Professorship in Optometry) of the Hong Kong Polytechnic University.

References

[1] F. Dudbridge, A. Gusnanto, Estimation of significance thresholds for genomewide association scans, Genet. Epidemiol. 32 (2008) 227–234.
[2] O.A. Panagiotou, J.P. Ioannidis, What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations, Int. J. Epidemiol. 41 (2012) 273–286.
[3] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F. Mackay, S.A. McCarroll, P.M. Visscher, Finding the missing heritability of complex diseases, Nature 461 (2009) 747–753.
[4] P. Zeng, Y. Zhao, C. Qian, L. Zhang, R. Zhang, J. Gou, J. Liu, L. Liu, F. Chen, Statistical analysis for genome-wide association study, J. Biomed. Res. 29 (2015) 285–297.
[5] K. Wang, M. Li, H. Hakonarson, Analysing biological pathways in genome-wide association studies, Nat. Rev. Genet. 11 (2010) 843–854.
[6] A. Yuryev, Present and Future of Pathway Analysis in Drug Discovery, Pathway Analysis for Drug Discovery, John Wiley & Sons, Inc., Place Published, 2008 285–296.
[7] A. Yuryev, Introduction to Pathway Analysis, Pathway Analysis for Drug Discovery, John Wiley & Sons, Inc., Place Published, 2008 1–25.
[8] P. Jia, Z. Zhao, Network.assisted analysis to prioritize GWAS results: principles, methods and perspectives, Hum. Genet. 133 (2014) 125–138.
[9] V.K. Ramanan, L. Shen, J.H. Moore, A.J. Saykin, Pathway analysis of genomic data: concepts, methods, and prospects for future development, Trends Genet. 28 (2012) 323–332.
[10] P. Khatri, M. Sirota, A.J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput. Biol. 8 (2012), e1002375. .
[11] B.L. Fridley, J.M. Biernacka, Gene set analysis of SNP data: benefits, challenges, and future directions, Eur. J. Hum. Genet. 19 (2011) 837–843.
[12] P. Holmans, E.K. Green, J.S. Pahwa, M.A. Ferreira, S.M. Purcell, P. Sklar, C. Wellcome Trust Case–control, M.J. Owen, M.C. O'Donovan, N. Craddock, Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder, Am. J. Hum. Genet. 85 (2009) 13–24.
[13] K. Zhang, S. Cui, S. Chang, L. Zhang, J. Wang, i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study, Nucleic Acids Res. 38 (2010) W90–W95.
[14] K. Zhang, S. Chang, L. Guo, J. Wang, I-GSEA4GWAS v2: a web server for functional analysis of SNPs in trait-associated pathways identified from genome-wide association study, Protein Cell 6 (2015) 221–224.
[15] A.V. Segre, D. Consortium, M. investigators, L. Groop, V.K. Mootha, M.J. Daly, D. Altshuler, Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits, PLoS Genet. 6 (2010), e1001058. .
[16] J.Z. Liu, A.F. McRae, D.R. Nyholt, S.E. Medland, N.R. Wray, K.M. Brown, A. Investigators, N.K. Hayward, G.W. Montgomery, P.M. Visscher, N.G. Martin, S. Macgregor, A versatile gene-based test for genome-wide association studies, Am. J. Hum. Genet. 87 (2010) 139–145.
[17] H. Chen, B.M. Sharp, Content-rich biological network constructed by mining PubMed abstracts, BMC Bioinformatics 5 (2004) 147.
[18] M. Bessarabova, A. Ishkin, L. JeBailey, T. Nikolskaya, Y. Nikolsky, Knowledge-based analysis of proteomics data, BMC Bioinformatics 13 (2012) S13.
[19] S. Ekins, S. Andreyev, A. Ryabov, E. Kirillov, E.A. Rakhmatulin, S. Sorokina, A. Bugrim, T. Nikolskaya, A combined approach to drug metabolism and toxicity assessment, Drug Metab. Dispos. 34 (2006) 495–503.
[20] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.

[21] C. O'Dushlaine, E. Kenny, E.A. Heron, R. Segurado, M. Gill, D.W. Morris, A. Corvin, The SNP ratio test: pathway analysis of genome-wide association datasets, Bioinformatics 25 (2009) 2762–2763.
[22] J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes, Nat. Genet. 39 (2007) 906–913.
[23] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G.R. Abecasis, Fast and accurate genotype imputation in genome-wide association studies through pre-phasing, Nat. Genet. 44 (2012) 955–959.
[24] M. Evangelou, A. Rendon, W.H. Ouwehand, L. Wernisch, F. Dudbridge, Comparison of methods for competitive tests of pathway analysis, PLoS One 7 (2012) e41018.
[25] C.A. de Leeuw, B.M. Neale, T. Heskes, D. Posthuma, The statistical properties of gene-set analysis, Nat. Rev. Genet. 17 (2016) 353–364.
[26] A. Kramer, J. Green, J. Pollard Jr., S. Tugendreich, Causal analysis approaches in ingenuity pathway analysis, Bioinformatics 30 (2014) 523–530.
[27] K. Yu, Q. Li, A.W. Bergen, R.M. Pfeiffer, P.S. Rosenberg, N. Caporaso, P. Kraft, N. Chatterjee, Pathway analysis by adaptive combination of P-values, Genet. Epidemiol. 33 (2009) 700–709.
[28] L. Wang, P. Jia, R.D. Wolfinger, X. Chen, Z. Zhao, Gene set analysis of genome-wide association studies: methodological issues and perspectives, Genomics 98 (2011) 1–8.
[29] M.A. Mooney, J.T. Nigg, S.K. McWeeney, B. Wilmot, Functional and genomic context in pathway analysis of GWAS data, Trends Genet. 30 (2014) 390–400.
[30] A. Aterido, A. Julia, C. Ferrandiz, L. Puig, E. Fonseca, E. Fernandez-Lopez, E. Dauden, J.L. Sanchez-Carazo, J.L. Lopez-Estebaranz, D. Moreno-Ramirez, R.F. Vanaclocha, E. Herrera, P. de la Cueva, N. Dand, N. Palau, A. Alonso, M. Lopez-Lasanta, R. Tortosa, A. Garcia-Montero, L. Codo, J.L. Gelpi, J. Bertranpetit, D. Absher, F. Capon, R.M. Myers, J.N. Barker, S. Marsal, Genome-wide pathway analysis identifies new genetic pathways associated with psoriasis, J. Investig. Dermatol. 136 (2016) 593–602.
[31] F. Zipp, A.J. Ivinson, J.L. Haines, S. Sawcer, P. Dejager, S.L. Hauser, J.R. Oksenberg, Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls, Am. J. Hum. Genet. 92 (2013) 854–865.
[32] J.B. Veyrieras, S. Kudaravalli, S.Y. Kim, E.T. Dermitzakis, Y. Gilad, M. Stephens, J.K. Pritchard, High-resolution mapping of expression-QTLs yields insight into human gene regulation, PLoS Genet. 4 (2008), e1000214. .
[33] A. Uzun, A.T. Dewan, S. Istrail, J.F. Padbury, Pathway-based genetic analysis of preterm birth, Genomics 101 (2013) 163–170.
[34] R. Koster, N. Mitra, K. D'Andrea, S. Vardhanabhuti, C.C. Chung, Z. Wang, R.L. Erickson, D.J. Vaughn, K. Litchfield, N. Rahman, M.H. Greene, K.A. McGlynn, C. Turnbull, S.J. Chanock, K.L. Nathanson, P.A. Kanetsky, Pathway-based analysis of GWAs data identifies association of sex determination genes with susceptibility to testicular germ cell tumors, Hum. Mol. Genet. 23 (2014) 6061–6068.
[35] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
[36] A. Kiezun, K. Garimella, R. Do, N.O. Stitziel, B.M. Neale, P.J. McLaren, N. Gupta, P. Sklar, P.F. Sullivan, J.L. Moran, C.M. Hultman, P. Lichtenstein, P. Magnusson, T. Lehner, Y.Y. Shugart, A.L. Price, P.I. de Bakker, S.M. Purcell, S.R. Sunyaev, Exome sequencing and the genetic basis of complex traits, Nat. Genet. 44 (2012) 623–630.
[37] J. Asimit, E. Zeggini, Rare variant association analysis methods for complex traits, Annu. Rev. Genet. 44 (2010) 293–308.
[38] O. Zuk, S.F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M.J. Daly, B.M. Neale, S.R. Sunyaev, E.S. Lander, Searching for missing heritability: designing rare variant association studies, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) E455–E464.
[39] S. Lee, G.R. Abecasis, M. Boehnke, X. Lin, Rare-variant association analysis: study designs and statistical tests, Am. J. Hum. Genet. 95 (2014) 5–23.
[40] S. Basu, W. Pan, Comparison of statistical tests for disease association with rare variants, Genet. Epidemiol. 35 (2011) 606–619.
[41] H.W. Uh, R. Tsonaka, J.J. Houwing-Duistermaat, Does pathway analysis make it easier for common variants to tag rare ones? BMC Proc. 5 (2011) S90.
[42] G. Wu, D. Zhi, Pathway-based approaches for sequencing-based genome-wide association studies, Genet. Epidemiol. 37 (2013) 478–494.
[43] M. Holden, S. Deng, L. Wojnowski, B. Kulle, GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies, Bioinformatics 24 (2008) 2784–2785.
[44] K. Wang, M. Li, M. Bucan, Pathway-based approaches for analysis of genomewide association studies, Am. J. Hum. Genet. 81 (2007) 1278–1283.
[45] B.E. Madsen, S.R. Browning, A groupwise association test for rare mutations using a weighted sum statistic, PLoS Genet. 5 (2009), e1000384. .
[46] A.P. Morris, E. Zeggini, An evaluation of statistical approaches to rare variant analysis in genetic association studies, Genet. Epidemiol. 34 (2010) 188–193.
[47] B. Li, S.M. Leal, Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data, Am. J. Hum. Genet. 83 (2008) 311–321.
[48] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, X. Lin, Rare-variant association testing for sequencing data with the sequence kernel association test, Am. J. Hum. Genet. 89 (2011) 82–93.
[49] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (2015) 68–74.
[50] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.
[51] W. Pan, I.Y. Kwak, P. Wei, A powerful pathway-based adaptive test for genetic association with common or rare variants, Am. J. Hum. Genet. 97 (2015) 86–98.
[52] W. Pan, J. Kim, Y. Zhang, X. Shen, P. Wei, A powerful and adaptive association test for rare variants, Genetics 197 (2014) 1081–1095.
[53] J. Zhao, Y. Zhu, E. Boerwinkle, M. Xiong, Pathway analysis with next-generation sequencing data, Eur. J. Hum. Genet. 23 (2015) 507–515.
[54] P.C. Sham, S.M. Purcell, Statistical power and significance testing in large-scale genetic studies, Nat. Rev. Genet. 15 (2014) 335–346.
[55] I.P. Gorlov, O.Y. Gorlova, S.R. Sunyaev, M.R. Spitz, C.I. Amos, Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms, Am. J. Hum. Genet. 82 (2008) 100–112.
[56] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, Nat. Genet. 38 (2006) 904–909.
[57] Z.Z. Tang, D.Y. Lin, Meta-analysis for discovering rare-variant associations: statistical methods and software programs, Am. J. Hum. Genet. 97 (2015) 35–53.
[58] W. Bodmer, C. Bonilla, Common and rare variants in multifactorial susceptibility to common diseases, Nat. Genet. 40 (2008) 695–701.
[59] P.L. Auer, G. Lettre, Rare variant association studies: considerations, challenges and opportunities, Genome Med. 7 (2015) 16.
[60] G.V. Kryukov, L.A. Pennacchio, S.R. Sunyaev, Most rare missense alleles are deleterious in humans: implications for complex disease and association studies, Am. J. Hum. Genet. 80 (2007) 727–739.
[61] T. Walsh, J.M. McClellan, S.E. McCarthy, A.M. Addington, S.B. Pierce, G.M. Cooper, A.S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S.M. Stray, C.F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R.L. Findling, L. Sikich, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E.E. Eichler, P.S. Meltzer, S.F. Nelson, A.B. Singleton, M.K. Lee, J.L. Rapoport, M.C. King, J. Sebat, Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia, Science 320 (2008) 539–543.
[62] A. Califano, A.J. Butte, S. Friend, T. Ideker, E. Schadt, Leveraging models of cell regulation and GWAS data in integrative network-based association studies, Nat. Genet. 44 (2012) 841–847.
[63] M.D. Ritchie, E.R. Holzinger, R. Li, S.A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype-phenotype interactions, Nat. Rev. Genet. 16 (2015) 85–97.
[64] V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G.H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K.P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H.G. Gislason, T. Stefansson, B.G. Leifsson, U. Thorsteinsdottir, J.R. Lamb, J.R. Gulcher, M.L. Reitman, A. Kong, E.E. Schadt, K. Stefansson, Genetics of gene expression and its effect on disease, Nature 452 (2008) 423–428.
[65] H. Zhong, X. Yang, L.M. Kaplan, C. Molony, E.E. Schadt, Integrating pathway analysis and genetics of gene expression for genome-wide association studies, Am. J. Hum. Genet. 86 (2010) 581–591.
[66] M. Zhang, L. Liang, N. Morar, A.L. Dixon, G.M. Lathrop, J. Ding, M.F. Moffatt, W.O. Cookson, P. Kraft, A.A. Qureshi, J. Han, Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma, Hum. Genet. 131 (2012) 615–623.
[67] S. Bunyavanich, E.E. Schadt, B.E. Himes, J. Lasky-Su, W. Qiu, R. Lazarus, J.P. Ziniti, A. Cohain, M. Linderman, D.G. Torgerson, C.S. Eng, M. Pino-Yanes, B. Padhukasahasram, J.J. Yang, R.A. Mathias, T.H. Beaty, X. Li, P. Graves, I. Romieu, B.D.R. Navarro, M.T. Salam, H. Vora, D.L. Nicolae, C. Ober, F.D. Martinez, E.R. Bleecker, D.A. Meyers, W.J. Gauderman, F. Gilliland, E.G. Burchard, K.C. Barnes, L.K. Williams, S.J. London, B. Zhang, B.A. Raby, S.T. Weiss, Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis, BMC Med. Genet. 7 (2014) 48.
[68] V.P. Mäkinen, M. Civelek, Q. Meng, B. Zhang, J. Zhu, C. Levian, T. Huan, A.V. Segrè, S. Ghosh, J. Vivar, M. Nikpay, A.F.R. Stewart, C.P. Nelson, C. Willenborg, J. Erdmann, S. Blakenberg, C.J. O'Donnell, W. März, R. Laaksonen, S.E. Epstein, S. Kathiresan, S.H. Shah, S.L. Hazen, M.P. Reilly, A.J. Lusis, N.J. Samani, H. Schunkert, T. Quertermous, R. McPherson, X. Yang, T.L. Assimes, Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease, PLoS Genet. 10 (2014) e1004502.
[69] T. Huan, Q. Meng, M.A. Saleh, A.E. Norlander, R. Joehanes, J. Zhu, B.H. Chen, B. Zhang, A.D. Johnson, S. Ying, P. Courchesne, N. Raghavachari, R. Wang, P. Liu, C.J. O'Donnell, R. Vasan, P.J. Munson, M.S. Madhur, D.G. Harrison, X. Yang, D. Levy, Integrative network analysis reveals molecular mechanisms of blood pressure regulation, Mol. Syst. Biol. 11 (2015) 1–15.
[70] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B.W. Penninx, R. Jansen, E.J. de Geus, D.I. Boomsma, F.A. Wright, P.F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A.J. Lusis, T. Lehtimaki, E. Raitoharju, M. Kahonen, I. Seppala, O.T. Raitakari, J. Kuusisto, M. Laakso, A.L. Price, P. Pajukanta, B. Pasaniuc, Integrative approaches for large-scale transcriptome-wide association studies, Nat. Genet. 48 (2016) 245–252.
[71] A.E. Locke, B. Kahali, S.I. Berndt, A.E. Justice, T.H. Pers, F.R. Day, C. Powell, S. Vedantam, M.L. Buchkovich, J. Yang, D.C. Croteau-Chonka, T. Esko, T. Fall, T. Ferreira, S. Gustafsson, Z. Kutalik, J. Luan, R. Mägi, J.C. Randall, T.W. Winkler, A.R. Wood, T. Workalemahu, J.D. Faul, J.A. Smith, J.H. Zhao, W. Zhao, J. Chen, R. Fehrmann, Å.K. Hedman, J. Karjalainen, E.M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J.L. Bolton, J.L. Bragg-Gresham, S. Buyske, A. Demirkan, G. Deng, G.B. Ehret, B. Feenstra, M.F. Feitosa, K. Fischer, A. Goel, J. Gong, A.U. Jackson, S. Kanoni, M.E. Kleber, K. Kristiansson, U. Lim, V. Lotay, M. Mangino, I.M. Leach, C. Medina-Gomez, S.E. Medland, M.A. Nalls, C.D. Palmer, D. Pasko, S. Pechlivanis, M.J. Peters, I. Prokopenko, D. Shungin, A.

Stančáková, R.J. Strawbridge, Y.J. Sung, T. Tanaka, A. Teumer, S. Trompet, S.W. Van Der Laan, J. Van Setten, J.V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, A. Isaacs, E. Albrecht, J. Ärnlöv, G.M. Arscott, A.P. Attwood, S. Bandinelli, A. Barrett, I.N. Bas, C. Bellis, A.J. Bennett, C. Berne, R. Blagieva, M. Blüher, S. Böhringer, L.L. Bonnycastle, Y. Böttcher, H.A. Boyd, M. Bruinenberg, I.H. Caspersen, Y.D.I. Chen, R. Clarke, E.W. Daw, A.J.M. De Craen, G. Delgado, M. Dimitriou, A.S.F. Doney, N. Eklund, K. Estrada, E. Eury, L. Folkersen, R.M. Fraser, M.E. Garcia, F. Geller, V. Giedraitis, B. Gigante, A.S. Go, A. Golay, A.H. Goodall, S.D. Gordon, M. Gorski, H.J. Grabe, H. Grallert, T.B. Grammer, J. Gräßler, H. Grönberg, C.J. Groves, G. Gusto, J. Haessler, P. Hall, T. Haller, G. Hallmans, C.A. Hartman, M. Hassinen, C. Hayward, N.L. Heard-Costa, Q. Helmer, C. Hengstenberg, O. Holmen, J.J. Hottenga, A.L. James, J.M. Jeff, Å. Johansson, J. Jolley, T. Juliusdottir, L. Kinnunen, W. Koenig, M. Koskenvuo, W. Kratzer, J. Laitinen, C. Lamina, K. Leander, N.R. Lee, P. Lichtner, L. Lind, J. Lindström, K.S. Lo, S. Lobbens, R. Lorbeer, Y. Lu, F. Mach, P.K.E. Magnusson, A. Mahajan, W.L. McArdle, S. McLachlan, C. Menni, S. Merger, E. Mihailov, L. Milani, A. Moayyeri, K.L. Monda, M.A. Morken, A. Mulas, G. Müller, M. Müller-Nurasyid, A.W. Musk, R. Nagaraja, M.M. Nöthen, I.M. Nolte, S. Pilz, N.W. Rayner, F. Renstrom, R. Rettig, J.S. Ried, S. Ripke, N.R. Robertson, L.M. Rose, S. Sanna, H. Scharnagl, S. Scholtens, F.R. Schumacher, W.R. Scott, T. Seufferlein, J. Shi, A.V. Smith, J. Smolonska, A.V. Stanton, V. Steinthorsdottir, K. Stirrups, H.M. Stringham, J. Sundström, M.A. Swertz, A.J. Swift, A.C. Syvänen, S.T. Tan, B.O. Tayo, B. Thorand, G. Thorleifsson, J.P. Tyrer, H.W. Uh, L. Vandenput, F.C. Verhulst, S.H. Vermeulen, N. Verweij, J.M. Vonk, L.L. Waite, H.R. Warren, D. Waterworth, M.N. Weedon, L.R. Wilkens, C. Willenborg, T. Wilsgaard, M.K. Wojczynski, A. Wong, A.F. Wright, Q. Zhang, E.P. Brennan, M. Choi, Z. Dastani, A.W. Drong, P. Eriksson, A. Franco-Cereceda, J.R. Gådin, A.G. Gharavi, M.E. Goddard, R.E. Handsaker, J. Huang, F. Karpe, S. Kathiresan, S. Keildson, K. Kiryluk, M. Kubo, J.Y. Lee, L. Liang, R.P. Lifton, B. Ma, S.A. McCarroll, A.J. McKnight, J.L. Min, M.F. Moffatt, G.W. Montgomery, J.M. Murabito, G. Nicholson, D.R. Nyholt, Y. Okada, J.R.B. Perry, R. Dorajoo, E. Reinmaa, R.M. Salem, N. Sandholm, R.A. Scott, L. Stolk, A. Takahashi, T. Tanaka, F.M. Van T'Hooft, A.A.E. Vinkhuyzen, H.J. Westra, W. Zheng, K.T. Zondervan, A.C. Heath, D. Arveiler, S.J.L. Bakker, J. Beilby, R.N. Bergman, J. Blangero, P. Bovet, H. Campbell, M.J. Caulfield, G. Cesana, A. Chakravarti, D.I. Chasman, P.S. Chines, F.S. Collins, D.C. Crawford, L.A. Cupples, D. Cusi, J. Danesh, U. De Faire, H.M. Den Ruijter, A.F. Dominiczak, R. Erbel, J. Erdmann, J.G. Eriksson, M. Farrall, S.B. Felix, E. Ferrannini, J. Ferrières, I. Ford, N.G. Forouhi, T. Forrester, O.H. Franco, R.T. Gansevoort, P.V. Gejman, C. Gieger, O. Gottesman, V. Gudnason, U. Gyllensten, A.S. Hall, T.B. Harris, A.T. Hattersley, A.A. Hicks, L.A. Hindorff, A.D. Hingorani, A. Hofman, G. Homuth, G.K. Hovingh, S.E. Humphries, S.C. Hunt, E. Hyppönen, T. Illig, K.B. Jacobs, M.R. Jarvelin, K.H. Jöckel, B. Johansen, P. Jousilahti, J.W. Jukema, A.M. Jula, J. Kaprio, J.J.P. Kastelein, S.M. Keinanen-Kiukaanniemi, L.A. Kiemeney, P. Knekt, J.S. Kooner, C. Kooperberg, P. Kovacs, A.T. Kraja, M. Kumari, J. Kuusisto, T.A. Lakka, C. Langenberg, L. Le Marchand, T. Lehtimäki, V. Lyssenko, S. Männistö, A. Marette, T.C. Matise, C.A. McKenzie, B. McKnight, F.L. Moll, A.D. Morris, A.P. Morris, J.C. Murray, M. Nelis, C. Ohlsson, A.J. Oldehinkel, K.K. Ong, P.A.F. Madden, G. Pasterkamp, J.F. Peden, A. Peters, D.S. Postma, P.P. Pramstaller, J.F. Price, L. Qi, O.T. Raitakari, T. Rankinen, D.C. Rao, T.K. Rice, P.M. Ridker, J.D. Rioux, M.D. Ritchie, I. Rudan, V. Salomaa, N.J. Samani, J. Saramies, M.A. Sarzynski, H. Schunkert, P.E.H. Schwarz, P. Sever, A.R. Shuldiner, J. Sinisalo, R.P. Stolk, K. Strauch, A. Tönjes, D.A. Trégouët, A. Tremblay, E. Tremoli, J. Virtamo, M.C. Vohl, U. Völker, G. Waeber, G. Willemsen, J.C. Witteman, M.C. Zillikens, L.S. Adair, P. Amouyel, F.W. Asselbergs, T.L. Assimes, M. Bochud, B.O. Boehm, E. Boerwinkle, S.R. Bornstein, E.P. Bottinger, C. Bouchard, S. Cauchi, J.C. Chambers, S.J. Chanock, R.S. Cooper, P.I.W. De Bakker, G. Dedoussis, L. Ferrucci, P.W. Franks, P. Froguel, L.C. Groop, C.A. Haiman, A. Hamsten, J. Hui, D.J. Hunter, K. Hveem, R.C. Kaplan, M. Kivimaki, D. Kuh, M. Laakso, Y. Liu, N.G. Martin, W. März, M. Melbye, A. Metspalu, S. Moebus, P.B. Munroe, I. Njølstad, B.A. Oostra, C.N.A. Palmer, N.L. Pedersen, M. Perola, L. Pérusse, U. Peters, C. Power, T. Quertermous, R. Rauramaa, F. Rivadeneira, T.E. Saaristo, D. Saleheen, N. Sattar, E.E. Schadt, D. Schlessinger, P.E. Slagboom, H. Snieder, T.D. Spector, U. Thorsteinsdottir, M. Stumvoll, J. Tuomilehto, A.G. Uitterlinden, M. Uusitupa, P. Van Der Harst, M. Walker, H. Wallaschofski, N.J. Wareham, H. Watkins, D.R. Weir, H.E. Wichmann, J.F. Wilson, P. Zanen, I.B. Borecki, P. Deloukas, C.S. Fox, I.M. Heid, J.R. O'Connell, D.P. Strachan, K. Stefansson, C.M. Van Duijn, G.R. Abecasis, L. Franke, T.M. Frayling, M.I. McCarthy, P.M. Visscher, A. Scherag, C.J. Willer, M. Boehnke, K.L. Mohlke, C.M. Lindgren, J.S. Beckmann, I. Barroso, K.E. North, E. Ingelsson, J.N. Hirschhorn, R.J.F. Loos, E.K. Speliotes, Genetic studies of body mass index yield new insights for obesity biology, Nature 518 (2015) 197–206.

[72] T.H. Pers, J.M. Karjalainen, Y. Chan, H.J. Westra, A.R. Wood, J. Yang, J.C. Lui, S. Vedantam, S. Gustafsson, T. Esko, T. Frayling, E.K. Speliotes, Genetic Investigation of A.T.C., M. Boehnke, S. Raychaudhuri, R.S. Fehrmann, J.N. Hirschhorn, L. Franke, Biological interpretation of genome-wide association studies using predicted gene functions, Nat. Commun. 6 (2015) 5890.

[73] E. Davidson, M. Levin, Gene regulatory networks, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 4935.

[74] J.M. Stuart, E. Segal, D. Koller, S.K. Kim, A gene-coexpression network for global discovery of conserved genetic modules, Science 302 (2003) 249–255.

[75] P. Yao, P. Lin, A. Gokoolparsadh, A. Assareh, M.W. Thang, I. Voineagu, Coexpression networks identify brain region-specific enhancer RNAs in the human brain, Nat. Neurosci. 18 (2015) 1168–1174.

[76] H. Luo, D. Bu, L. Sun, S. Fang, Z. Liu, Y. Zhao, Identification and function annotation of long intervening noncoding RNAs, Brief. Bioinform. (2016) [Epub ahead of print].

[77] D. Gaidatzis, K. Jacobeit, E.J. Oakeley, M.B. Stadler, Overestimation of alternative splicing caused by variable probe characteristics in exon arrays, Nucleic Acids Res. 37 (2009), e107. .

[78] S. Liu, L. Lin, P. Jiang, D. Wang, Y. Xing, A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species, Nucleic Acids Res. 39 (2011) 578–588.

[79] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng, P. Khaitovich, Estimating accuracy of RNA-Seq and microarrays with proteomics, BMC Genomics 10 (2009) 161.

[80] S. Zhao, W.P. Fung-Leung, A. Bittner, K. Ngo, X. Liu, Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells, PLoS One 9 (2014), e78644. .

[81] J.W. Walley, R.C. Sartor, Z. Shen, R.J. Schmitz, K.J. Wu, M.A. Urich, J.R. Nery, L.G. Smith, J.C. Schnable, J.R. Ecker, S.P. Briggs, Integration of omic networks in a developmental atlas of maize, Science 353 (2016) 814–818.

[82] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinformatics 9 (2008) 559.

[83] R.S. Fehrmann, J.M. Karjalainen, M. Krajewska, H.J. Westra, D. Maloney, A. Simeonov, T.H. Pers, J.N. Hirschhorn, R.C. Jansen, E.A. Schultes, H.H. van Haagen, E.G. de Vries, G.J. te Meerman, C. Wijmenga, M.A. van Vugt, L. Franke, Gene expression analysis identifies global gene dosage sensitivity in cancer, Nat. Genet. 47 (2015) 115–125.

[84] Q. Xiong, N. Ancona, E.R. Hauser, S. Mukherjee, T.S. Furey, Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets, Genome Res. 22 (2012) 386–397.

[85] Y.T. Huang, L. Liang, M.F. Moffatt, W.O. Cookson, X. Lin, iGWAS: integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis, Genet. Epidemiol. 39 (2015) 347–356.

[86] D.P. MacKinnon, A.J. Fairchild, M.S. Fritz, Mediation analysis, Annu. Rev. Psychol. 58 (2007) 593–614.

[87] Y.T. Huang, T.J. Vanderweele, X. Lin, Joint analysis of SNP and gene expression data in genetic association studies of complex diseases, Ann. Appl. Stat. 8 (2014) 352–376.

[88] J.M. Robins, S. Greenland, Identifiability and exchangeability for direct and indirect effects, Epidemiology 3 (1992) 143–155.

[89] W. Zhao, P. Langfelder, T. Fuller, J. Dong, A. Li, S. Hovarth, Weighted gene coexpression network analysis: state of the art, J. Biopharm. Stat. 20 (2010) 281–300.

[90] H. Dharuri, A. Demirkan, J.B. van Klinken, D.O. Mook-Kanamori, C.M. van Duijn, P.A. t Hoen, K. Willems van Dijk, Genetics of the human metabolome, what is next? Biochim. Biophys. Acta 1842 (2014) 1923–1931.

[91] H. Sun, H. Wang, R. Zhu, K. Tang, Q. Gong, J. Cui, Z. Cao, Q. Liu, IPEAP: integrating multiple omics and genetic data for pathway enrichment analysis, Bioinformatics 30 (2014) 737–739.

[92] K. Wanichthanarak, J.F. Fahrmann, D. Grapov, Genomic, proteomic, and metabolomic data integration strategies, Biomark. Insights 10 (2015) 1–6.

[93] F. Zhang, Y. Wen, X. Guo, T. Yang, H. Shen, X. Chen, L. Tan, Q. Tian, H.W. Deng, Trans-omics pathway analysis suggests that eQTLs contribute to chondrocyte apoptosis of Kashin-Beck disease through regulating apoptosis pathway expression, Gene 553 (2014) 166–169.

[94] M. Fondi, P. Liò, Multi -omics and metabolic modelling pipelines: challenges and tools for systems microbiology, Microbiol. Res. 171 (2015) 52–64.

[95] M. Kohl, D.A. Megger, M. Trippler, H. Meckel, M. Ahrens, T. Bracht, F. Weber, A.C. Hoffmann, H.A. Baba, B. Sitek, J.F. Schlaak, H.E. Meyer, C. Stephan, M. Eisenacher, A practical data processing workflow for multi-OMICS projects, Biochim. Biophys. Acta 1844 (2014) 52–62.

[96] V. Zoldos, M. Novokmet, I. Beceheli, G. Lauc, Genomics and epigenomics of the human glycome, Glycoconj. J. 30 (2013) 41–50.

[97] S.V. Bennun, D.B. Hizal, K. Heffner, O. Can, H. Zhang, M.J. Betenbaugh, Systems glycobiology: integrating glycogenomics, glycoproteomics, glycomics, and other omics data sets to characterize cellular glycosylation processes, J. Mol. Biol. 428 (2016) 3337–3352.

[98] S.V. Bennun, K.J. Yarema, M.J. Betenbaugh, F.J. Krambeck, Integration of the transcriptome and glycome for identification of glycan cell signatures, PLoS Comput. Biol. 9 (2013), e1002813. .

[99] V. Arbona, M. Manzi, C. Ollas, A. Gomez-Cadenas, Metabolomics as a tool to investigate abiotic stress tolerance in plants, Int. J. Mol. Sci. 14 (2013) 4885–4911.

[100] J. Adamski, K. Suhre, Metabolomics platforms for genome wide association studies—linking the genome to the metabolome, Curr. Opin. Biotechnol. 24 (2013) 39–47.

[101] W.J. Griffiths, T. Koal, Y. Wang, M. Kohl, D.P. Enot, H.P. Deigner, Targeted metabolomics for biomarker discovery, Angew. Chem. Int. Ed. Eng. 49 (2010) 5426–5445.

[102] M. Malet-Martino, U. Holzgrabe, NMR techniques in biomedical and pharmaceutical analysis, J. Pharm. Biomed. Anal. 55 (2011) 1–15.

[103] C. Gieger, L. Geistlinger, E. Altmaier, M. Hrabe de Angelis, F. Kronenberg, T. Meitinger, H.W. Mewes, H.E. Wichmann, K.M. Weinberger, J. Adamski, T. Illig, K. Suhre, Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum, PLoS Genet. 4 (2008), e1000282. .

[104] G. Liebisch, W. Drobnik, M. Reil, B. Trumbach, R. Arnecke, B. Olgemoller, A. Roscher, G. Schmitz, Quantitative measurement of different ceramide species from crude cellular extracts by electrospray ionization tandem mass spectrometry (ESI-MS/MS), J. Lipid Res. 40 (1999) 1539–1546.

[105] A.A. Hicks, P.P. Pramstaller, A. Johansson, V. Vitart, I. Rudan, P. Ugocsai, Y. Aulchenko, C.S. Franklin, G. Liebisch, J. Erdmann, I. Jonasson, I.V. Zorkoltseva, C. Pattaro, C. Hayward, A. Isaacs, C. Hengstenberg, S. Campbell, C. Gnewuch, A.C. Janssens, A.V. Kirichenko, I.R. Konig, F. Marroni, O. Polasek, A. Demirkan, I. Kolcic, C. Schwienbacher, W. Igl, Z. Biloglav, J.C. Witteman, I. Pichler, G. Zaboli, T.I. Axenovich, A. Peters, S. Schreiber, H.E. Wichmann, H. Schunkert, N. Hastie, B.A. Oostra, S.H. Wild, T. Meitinger, U. Gyllensten, C.M. van Duijn, J.F. Wilson, A. Wright, G. Schmitz, H. Campbell, Genetic determinants of circulating sphingolipid concentrations in European populations, PLoS Genet. 5 (2009), e1000672. .
[106] T. Tanaka, J. Shen, G.R. Abecasis, A. Kisialiou, J.M. Ordovas, J.M. Guralnik, A. Singleton, S. Bandinelli, A. Cherubini, D. Arnett, M.Y. Tsai, L. Ferrucci, Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study, PLoS Genet. 5 (2009), e1000338. .
[107] K. Olden, N. Freudenberg, J. Dowd, A.E. Shields, Discovering how environmental exposures alter genes could lead to new treatments for chronic illnesses, Health Aff. (Millwood) 30 (2011) 833–841.
[108] D. Thomas, Gene—environment-wide association studies: emerging approaches, Nat. Rev. Genet. 11 (2010) 259–272.
[109] D. Thomas, Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies, Annu. Rev. Public Health 31 (2010) 21–36.
[110] S.J. Winham, J.M. Biernacka, Gene-environment interactions in genome-wide association studies: current approaches and new directions, J. Child Psychol. Psychiatry 54 (2013) 1120–1134.
[111] T.R. Rebbeck, M. Spitz, X. Wu, Assessing the function of genetic variants in candidate gene association studies, Nat. Rev. Genet. 5 (2004) 589–597.
[112] D.J. Hunter, Gene-environment interactions in human diseases, Nat. Rev. Genet. 6 (2005) 287–298.
[113] R. Feil, M.F. Fraga, Epigenetics and the environment: emerging patterns and implications, Nat. Rev. Genet. 13 (2011) 97–109.
[114] J.M. Seddon, S. George, B. Rosner, M.L. Klein, CFH gene variant, Y402H, and smoking, body mass index, environmental associations with advanced age-related macular degeneration, Hum. Hered. 61 (2006) 157–165.
[115] S. Scholtens, D.S. Postma, M.F. Moffatt, S. Panasevich, R. Granell, A.J. Henderson, E. Melen, F. Nyberg, G. Pershagen, D. Jarvis, A. Ramasamy, M. Wjst, C. Svanes, E. Bouzigon, F. Demenais, F. Kauffmann, V. Siroux, E. von Mutius, M.J. Ege, C. Braun-Fahrlander, J. Genuneit, G.s. group, B. Brunekreef, H.A. Smit, A.H. Wijga, M. Kerkhof, I. Curjuric, M. Imboden, G.A. Thun, N. Probst-Hensch, M.B. Freidin, E. Bragina, I.A. Deev, V.P. Puzyrev, D. Daley, J. Park, A. Becker, M. Chan-Yeung, A.L. Kozyrskyj, P. Pare, I. Marenholz, S. Lau, T. Keil, Y.A. Lee, M. Kabesch, C. Wijmenga, L. Franke, I.M. Nolte, J. Vonk, A. Kumar, M. Farrall, W.O. Cookson, D.P. Strachan, G.H. Koppelman, H.M. Boezen, Novel childhood asthma genes interact with in utero and early-life tobacco smoke exposure, J. Allergy Clin. Immunol. 133 (2014) 885–888.
[116] Q. Fan, V.J. Verhoeven, R. Wojciechowski, V.A. Barathi, P.G. Hysi, J.A. Guggenheim, R. Hohn, V. Vitart, A.P. Khawaja, K. Yamashiro, S.M. Hosseini, T. Lehtimaki, Y. Lu, T. Haller, J. Xie, C. Delcourt, M. Pirastu, J. Wedenoja, P. Gharahkhani, C. Venturini, M. Miyake, A.W. Hewitt, X. Guo, J. Mazur, J.E. Huffman, K.M. Williams, O. Polasek, H. Campbell, I. Rudan, Z. Vatavuk, J.F. Wilson, P.K. Joshi, G. McMahon, B. St Pourcain, D.M. Evans, C.L. Simpson, T.H. Schwantes-An, R.P. Igo, A. Mirshahi, A. Cougnard-Gregoire, C. Bellenguez, M. Blettner, O. Raitakari, M. Kahonen, I. Seppala, T. Zeller, T. Meitinger, E. Consortium for Refractive, Myopia, J.S. Ried, C. Gieger, L. Portas, E.M. van Leeuwen, N. Amin, A.G. Uitterlinden, F. Rivadeneira, A. Hofman, J.R. Vingerling, Y.X. Wang, X. Wang, E. Tai-Hui Boh, M.K. Ikram, C. Sabanayagam, P. Gupta, V. Tan, L. Zhou, C.E. Ho, W. Lim, R.W. Beuerman, R. Siantar, E.S. Tai, E. Vithana, E. Mihailov, C.C. Khor, C. Hayward, R.N. Luben, P.J. Foster, B.E. Klein, R. Klein, H.S. Wong, P. Mitchell, A. Metspalu, T. Aung, T.L. Young, M. He, O. Parssinen, C.M. van Duijn, J. Jin Wang, C. Williams, J.B. Jonas, Y.Y. Teo, D.A. Mackey, K. Oexle, N. Yoshimura, A.D. Paterson, N. Pfeiffer, T.Y. Wong, P.N. Baird, D. Stambolian, J.E. Wilson, C.Y. Cheng, C.J. Hammond, C.C. Klaver, S.M. Saw, J.S. Rahi, J.F. Korobelnik, J.P. Kemp, N.J. Timpson, G.D. Smith, J.E. Craig, K.P. Burdon, R.D. Fogarty, S.K. Iyengar, E. Chew, S. Janmahasatian, N.G. Martin, S. MacGregor, L. Xu, M. Schache, V. Nangia, S. Panda-Jonas, A.F. Wright, J.R. Fondran, J.H. Lass, S. Feng, J.H. Zhao, K.T. Khaw, N.J. Wareham, T. Rantanen, J. Kaprio, C.P. Pang, L.J. Chen, P.O. Tam, V. Jhanji, A.L. Young, A. Doring, L.J. Raffel, M.F. Cotch, X. Li, S.P. Yip, M.K. Yap, G. Biino, S. Vaccargiu, M. Fossarello, B. Fleck, S. Yazar, J.W. Tideman, M. Tedja, M.M. Deangelis, M. Morrison, L. Farrer, X. Zhou, W. Chen, N. Mizuki, A. Meguro, K.M. Makela, Meta-analysis of gene-environment-wide association scans accounting for education level identifies additional loci for refractive error, Nat. Commun. 7 (2016) 11008.
[117] S. Wei, L.E. Wang, M.K. McHugh, Y. Han, M. Xiong, C.I. Amos, M.R. Spitz, Q.W. Wei, Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility, Carcinogenesis 33 (2012) 1531–1537.
[118] N.I. Simonds, A.A. Ghazarian, C.B. Pimentel, S.D. Schully, G.L. Ellison, E.M. Gillanders, L.E. Mechanic, Review of the gene-environment interaction literature in cancer: what do we know? Genet. Epidemiol. 40 (2016) 356–602.
[119] K. Olden, S. Wilson, Environmental health and genomics: visions and implications, Nat. Rev. Genet. 1 (2000) 149–153.
[120] J.X. Hu, C.E. Thomas, S. Brunak, Network biology concepts in complex disease comorbidities, Nat. Rev. Genet. 17 (2016) 615–629.
[121] A. Tebani, C. Afonso, S. Marret, S. Bekri, Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations, Int. J. Mol. Sci. 17 (2016) 1555.
[122] R. Newton, L. Wernisch, A meta-analysis of multiple matched copy number and transcriptomics data sets for inferring gene regulatory relationships, PLoS One 9 (2014), e105522. .
[123] P.W. Laird, Principles and challenges of genomewide DNA methylation analysis, Nat. Rev. Genet. 11 (2010) 191–203.
[124] J.H. Kim, A. Karnovsky, V. Mahavisno, T. Weymouth, M. Pande, D.C. Dolinoy, L.S. Rozek, M.A. Sartor, LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types, BMC Genomics 13 (2012) 526.
[125] L. Stanberry, G.I. Mias, W. Haynes, R. Higdon, M. Snyder, E. Kolker, Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile, Metabolites 3 (2013) 741–760.
[126] R. Chen, G.I. Mias, J. Li-Pook-Than, L. Jiang, H.Y. Lam, R. Chen, E. Miriami, K.J. Karczewski, M. Hariharan, F.E. Dewey, Y. Cheng, M.J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J.T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A.P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M.A. Blasco, P.L. Greenberg, P. Snyder, T.E. Klein, R.B. Altman, A.J. Butte, E.A. Ashley, M. Gerstein, K.C. Nadeau, H. Tang, M. Snyder, Personal omics profiling reveals dynamic molecular and medical phenotypes, Cell 148 (2012) 1293–1307.
[127] O. Stegle, S.A. Teichmann, J.C. Marioni, Computational and analytical challenges in single-cell transcriptomics, Nat. Rev. Genet. 16 (2015) 133–145.
[128] C. Trapnell, Defining cell types and states with single-cell genomics, Genome Res. 25 (2015) 1491–1498.
[129] N.E. Navin, The first five years of single-cell cancer genomics and beyond, Genome Res. 25 (2015) 1499–1507.
[130] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W.R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing, Nature 472 (2011) 90–94.
[131] K. Kyoda, Y. Tohsato, K.H. Ho, S. Onami, Biological Dynamics Markup Language (BDML): an open format for representing quantitative biological dynamics data, Bioinformatics 31 (2015) 1044–1052.
[132] P. Jia, S. Zheng, J. Long, W. Zheng, Z. Zhao, dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks, Bioinformatics 27 (2011) 95–102.
[133] K. Wang, H. Zhang, S. Kugathasan, V. Annese, J.P. Bradfield, R.K. Russell, P.M.A. Sleiman, M. Imielinski, J. Glessner, C. Hou, D.C. Wilson, T. Walters, C. Kim, E.C. Frackelton, P. Lionetti, A. Barabino, J. Van Limbergen, S. Guthery, L. Denson, D. Piccoli, M. Li, M. Dubinsky, M. Silverberg, A. Griffiths, S.F.A. Grant, J. Satsangi, R. Baldassano, H. Hakonarson, Diverse genome-wide association studies sssociate the IL12/IL23 pathway with Crohn disease, Am. J. Hum. Genet. 84 (2009) 399–405.
[134] S. Raychaudhuri, R.M. Plenge, E.J. Rossin, A.C. Ng, C. International Schizophrenia, S.M. Purcell, P. Sklar, E.M. Scolnick, R.J. Xavier, D. Altshuler, M.J. Daly, Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions, PLoS Genet. 5 (2009), e1000534. .
[135] D. Nam, J. Kim, S.Y. Kim, S. Kim, GSA-SNP: a general approach for gene set analysis of polymorphisms, Nucleic Acids Res. 38 (2010) W749–W754.
[136] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 15545–15550.
[137] M.X. Li, J.S.H. Kwan, P.C. Sham, HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis, Am. J. Hum. Genet. 91 (2012) 478–488.
[138] M.X. Li, H.S. Gui, J.S.H. Kwan, P.C. Sham, GATES: a rapid and powerful gene-based association test using extended Simes procedure, Am. J. Hum. Genet. 88 (2011) 283–293.
[139] K. Zhang, S. Chang, S. Cui, L. Guo, L. Zhang, J. Wang, ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework, Nucleic Acids Res. 39 (2011) W437–W443.
[140] C.A. de Leeuw, J.M. Mooij, T. Heskes, D. Posthuma, MAGMA: generalized gene-set analysis of GWAS data, PLoS Comput. Biol. 11 (2015), e1004219. .
[141] B.L. Yaspan, W.S. Bush, E.S. Torstenson, D. Ma, M.A. Pericak-Vance, M.D. Ritchie, J.S. Sutcliffe, J.L. Haines, Genetic analysis of biological pathway data through genomic randomization, Hum. Genet. 129 (2011) 563–571.
[142] Y.S. Park, M. Schmidt, E.R. Martin, M.A. Pericak-Vance, R.H. Chung, Pathway-PDT: a flexible pathway analysis tool for nuclear families, BMC Bioinformatics 14 (2013) 267.
[143] L. Wang, T. Matsushita, L. Madireddy, P. Mousavi, S.E. Baranzini, PINBPA: cytoscape app for network analysis of GWAS data, Bioinformatics 31 (2015) 262–264.
[144] V. Moskvina, C. O'Dushlaine, S. Purcell, N. Craddock, P. Holmans, M.C. O'Donovan, Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study, Genet. Epidemiol. 35 (2011) 861–866.
[145] B. Wang, J.M. Cunningham, X.H. Yang, Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data, Bioinformatics 31 (2015) 3043–3045.
[146] M. Kutmon, M.P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A.R. Pico, C.T. Evelo, PathVisio 3: an extendable pathway analysis toolbox, PLoS Comput. Biol. 11 (2015), e1004085. .
[147] B. Bokanizad, R. Tagett, S. Ansari, B.H. Helmi, S. Draghici, SPATIAL: a System-level PAThway Impact AnaLysis approach, Nucleic Acids Res. 44 (2016) 5034–5044.
[148] Q. Zhang, Q. Long, J. Ott, AprioriGWAS, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects, PLoS Comput. Biol. 10 (2014) e1003627.
[149] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N.L.S. Tang, W. Yu, BOOST: a fast approach to detecting gene-gene interactions in genome-wide case–control studies, Am. J. Hum. Genet. 87 (2010) 325–340.

[150] T. Kam-Thong, D. Czamara, K. Tsuda, K. Borgwardt, C.M. Lewis, A. Erhardt-Lehmann, B. Hemmer, P. Rieckmann, M. Daake, F. Weber, C. Wolf, A. Ziegler, B. Pütz, F. Holsboer, B. Schölkopf, B. Müller-Myhsok, EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units, Eur. J. Hum. Genet. 19 (2011) 465–471.
[151] R.K. Walters, C. Laurin, G.H. Lubke, Epi2Loc: an R package to investigate two-locus epistatic models, Twin Res. Hum. Genet. 17 (2014) 272–278.
[152] T. Schüpbach, I. Xenarios, S. Bergmann, K. Kapur, FastEpistasis: a high performance computing solution for quantitative trait epistasis, Bioinformatics 26 (2010) 1468–1469.
[153] C. Herold, M. Steffens, F.F. Brockschmidt, M.P. Baur, T. Becker, INTERSNP: genome-wide interaction analysis guided by a priori information, Bioinformatics 25 (2009) 3275–3281.
[154] C. Herold, M. Mattheisen, A. Lacour, T. Vaitsiakhovich, M. Angisch, D. Drichel, T. Becker, Integrated genome-wide pathway association analysis with INTERSNP, Hum. Hered. 73 (2012) 63–72.
[155] T. Vaitsiakhovich, D. Drichel, C. Herold, A. Lacour, T. Becker, METAINTER: meta-analysis of multiple regression models in genome-wide association studies, Bioinformatics 31 (2015) 151–157.
[156] S. Prabhu, I. Pe'er, Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease, Genome Res. 22 (2012) 2230–2240.
[157] G.D. Bader, M.P. Cary, C. Sander, Pathguide: a pathway resource list, Nucleic Acids Res. 34 (2006) D504–D506.
[158] D. Nishimura, A view from the web: biocarta, Biotech Softw. Internet Rep. 2 (2001) 117–120.
[159] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The gene ontology consortium, Nat. Genet. 25 (2000) 25–29.
[160] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C.A. Fulcher, T.A. Holland, I.M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L.A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D.S. Weaver, D. Weerasinghe, P. Zhang, P.D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, Nucleic Acids Res. 42 (2014) D459–D471.
[161] C.F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, K.H. Buetow, PID: the pathway interaction database, Nucleic Acids Res. 37 (2009) D674–D679.
[162] D. Croft, A.F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M.R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, P. D'Eustachio, The reactome pathway knowledgebase, Nucleic Acids Res. 42 (2014) D472–D477.
[163] Q. Jiang, S. Jin, Y. Jiang, M. Liao, R. Feng, L. Zhang, G. Liu, J. Hao, Alzheimer's disease variants with the genome-wide significance are significantly enriched in immune pathways and active in immune cells, Mol. Neurobiol. (2016) [Epub ahead of print].
[164] X. Yang, H. Zhu, Q. Qin, Y. Yang, Y. Yang, H. Cheng, X. Sun, Genetic variants and risk of esophageal squamous cell carcinoma: a GWAS-based pathway analysis, Gene 556 (2015) 149–152.
[165] S. Swarup, W. Huang, T.F.C. Mackay, R.R.H. Anholt, Analysis of natural variation reveals neurogenetic networks for *Drosophila* olfactory behavior, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 1017–1022.
[166] S.P. Kar, M.F. Seldin, W. Chen, E. Lu, G.M. Hirschfield, P. Invernizzi, J. Heathcote, D. Cusi, M.E. Gershwin, K.A. Siminovitch, C.I. Amos, Pathway-based analysis of primary biliary cirrhosis genome-wide association studies, Genes Immun. 14 (2013) 179–186.
[167] I. Menashe, J.D. Figueroa, M. Garcia-Closas, N. Chatterjee, N. Malats, A. Picornell, D. Maeder, Q. Yang, L. Prokunina-Olsson, Z. Wang, F.X. Real, K.B. Jacobs, D. Baris, M. Thun, D. Albanes, M.P. Purdue, M. Kogevinas, A. Hutchinson, Y.P. Fu, W. Tang, L. Burdette, A. Tardón, C. Serra, A. Carrato, R. García-Closas, J. Lloreta, A. Johnson, M. Schwenn, A. Schned, G. Andriole Jr., A. Black, E.J. Jacobs, R.W. Diver, S.M. Gapstur, S.J. Weinstein, J. Virtamo, N.E. Caporaso, M.T. Landi, J.F. Fraumeni Jr., S.J. Chanock, D.T. Silverman, N. Rothman, Large-scale pathway-based analysis of bladder cancer genome-wide association data from five studies of European background, PLoS One 7 (2012) e29396.
[168] J.I. Nurnberger, D.L. Koller, J. Jung, H.J. Edenberg, T. Foroud, I. Guella, M.P. Vawter, J.R. Kelsoe, Identification of pathways for bipolar disorder: a meta-analysis, JAMA Psychiatry 71 (2014) 657–664.
[169] P. Holmans, V. Moskvina, L. Jones, M. Sharma, A. Vedernikov, F. Buchel, M. Sadd, J.M. Bras, F. Bettella, N. Nicolaou, J. Simón-sánchez, F. Mittag, J.R. Gibbs, C. Schulte, A. Durr, R. Guerreiro, D. Hernandez, A. Brice, H. Stefánsson, K. Majamaa, T. Gasser, P. Heutink, N.W. Wood, M. Martinez, A.B. Singleton, M.A. Nalls, J. Hardy, H.R. Morris, N.M. Williams, A pathway-based analysis provides additional support for an immune-related genetic susceptibility to Parkinson's disease, Hum. Mol. Genet. 22 (2013) 1039–1049.
[170] Y.H. Lee, J.H. Kim, G.G. Song, Pathway analysis of a genome-wide association study in schizophrenia, Gene 525 (2013) 107–115.
[171] C.F. Kao, P. Jia, Z. Zhao, P.H. Kuo, Enriched pathways for major depressive disorder identified from a genome-wide association study, Int. J. Neuropsychopharmacol. 15 (2012) 1401–1411.
[172] L.E. Duncan, P.A. Holmans, P.H. Lee, C.T. O'Dushlaine, A.W. Kirby, J.W. Smoller, D. Öngür, B.M. Cohen, Pathway analyses implicate glial cells in schizophrenia, PLoS One 9 (2014) e89441.
[173] L. De Las Fuentes, W. Yang, V.G. Dávila-Román, C.C. Gu, Pathway-based genome-wide association analysis of coronary heart disease identifies biologically important gene sets, Eur. J. Hum. Genet. 20 (2012) 1168–1173.
[174] Y.W. Lv, J. Wang, L. Sun, J.M. Zhang, L. Cao, Y.Y. Ding, Y. Chen, J.J. Dou, J. Huang, Y.F. Tang, W.T. Wu, W.R. Cui, H.T. Lv, Understanding the pathogenesis of Kawasaki disease by network and pathway analysis, Comput. Math. Methods Med. 2013 (2013) 989307.
[175] L. Beltrame, E. Calura, R.R. Popovici, L. Rizzetto, D.R. Guedez, M. Donato, C. Romualdi, S. Draghici, D. Cavalieri, The biological connection markup language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways, Bioinformatics 27 (2011) 2127–2133.
[176] M.D. Leiserson, J.V. Eldridge, S. Ramachandran, B.J. Raphael, Network analysis of GWAS data, Curr. Opin. Genet. Dev. 23 (2013) 602–610.
[177] K. Lage, Protein-protein interactions and genetic diseases: the interactome, Biochim. Biophys. Acta 1842 (2014) 1971–1980.
[178] S. Fields, O. Song, A novel genetic system to detect protein-protein interactions, Nature 340 (1989) 245–246.
[179] P. Braun, Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays, Proteomics 12 (2012) 1499–1518.
[180] A.C. Gavin, K. Maeda, S. Kuhner, Recent advances in charting protein-protein interaction: mass spectrometry-based approaches, Curr. Opin. Biotechnol. 22 (2011) 42–49.
[181] K. Tarassov, V. Messier, C.R. Landry, S. Radinovic, M.M. Serna Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, S.W. Michnick, An in vivo map of the yeast protein interactome, Science 320 (2008) 1465–1470.
[182] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, M. Kuhn, P. Bork, L.J. Jensen, C. von Mering, STRING v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (2015) D447–D452.
[183] C. Gene Ontology, Gene ontology consortium: going forward, Nucleic Acids Res. 43 (2015) D1049–D1056.
[184] X. Chen, L. Wang, B. Hu, M. Guo, J. Barnard, X. Zhu, Pathway-based analysis for genome-wide association studies using supervised principal components, Genet. Epidemiol. 34 (2010) 716–724.
[185] A.J. Adewale, I. Dinu, J.D. Potter, Q. Liu, Y. Yasui, Pathway analysis of microarray data via regression, J. Comput. Biol. 15 (2008) 269–277.
[186] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron, (MIAME)-toward standards for microarray data, Nat. Genet. 29 (2001) 365–371.