Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Pathway-based analysis of genomic variation data

Nir Atias[1], Sorin Istrail[2] and Roded Sharan[1]

A holy grail of genetics is to decipher the mapping from genotype to phenotype. Recent advances in sequencing technologies allow the efficient genotyping of thousands of individuals carrying a particular phenotype in an effort to reveal its genetic determinants. However, the interpretation of these data entails tackling significant statistical and computational problems that stem from the complexity of human phenotypes and the huge genotypic search space. Recently, an alternative pathway-level analysis has been employed to combat these problems. In this review we discuss these developments, describe the challenges involved and outline possible solutions and future directions for improvement.

Addresses
[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
[2] Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI 02912, United States

Corresponding author: Sharan, Roded (roded@post.tau.ac.il)

Current Opinion in Genetics & Development 2013, **23**:622–626

This review comes from a themed issue on **Genetics of system biology**

Edited by **Shamil Sunyaev** and **Fritz Roth**

For a complete overview see the Issue and the Editorial

Available online 24th October 2013

http://dx.doi.org/10.1016/j.gde.2013.09.002

Introduction

Recent technological leaps in sequencing are generating genomic variation data at an ever growing scale. In a typical genome-wide association study (GWAS), the genomes of individuals carrying a certain phenotype are compared to genomes of individuals that do not carry this phenotype [1]. The observed differences include single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and more; their associations to the phenotypic data are then assessed. The sheer size of the data makes its interpretation a formidable task, challenged by statistical and computational obstacles. Some fundamental challenges include the significance scoring of SNPs while accounting for multiple hypothesis testing issues, the association of SNPs and CNVs to genes or other functional entities and the identification of subsets of SNPs/genes that underlie the phenotype of interest [2].

The failure of finding, for most complex diseases, robust statistical associations for individual SNPs, or individual genes, and the increased interest in understanding *genetic heterogeneity* (hundreds of rare and personalized mutations in many genes with small effect individually in explaining the disease phenotype, but all possibly affecting the 'pathway' of disease) [3], led to the quest for finding associations at a higher level of gene organization — namely, protein pathways. With 'pathway' having a number of definitions — in one extreme being just as a set of genes, and, at the other extreme, as a set of genes with a specific interaction pattern — there is significant interest in modeling, at some informative level, the concept of 'biological disease pathway', that is, a set of genes 'involved' in a disease, together with the overarching mechanism the set describes. The arising *pathway association* paradigm [4] allows pinpointing subsets of related SNPs while avoiding the multiple hypothesis testing problem that is associated with enumerating all possible subsets.

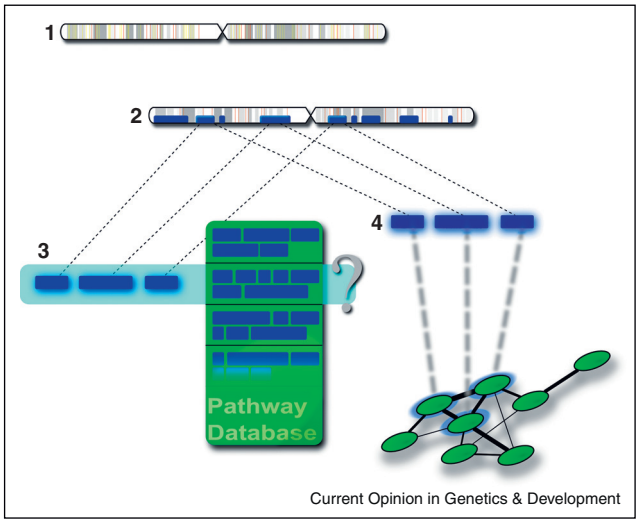
A key resource in the systematic identification of pathway associations is a protein–protein interaction network; by projecting associated SNPs onto this network one can zoom in on specific regions of the network that harbor significantly many associations. Such inference techniques were originally employed for interpreting gene expression data [5] and recently applied to analyze genomic variation data.

In this review we survey current approaches for pathway association, highlighting the technical difficulties of this domain and how they are being tackled: first, the SNP to gene mapping, while accounting for linkage disequilibrium (LD) patterns; second, the exponential number of subsets of genes at 'pathway' size that need to be considered; and third, the graph theoretic methods of constructing disease focused protein networks from associated SNPs that coherently capture the underlying molecular mechanisms. While the discussion below focuses on common genomic variation, the concepts and methods generalize to other types of variation such as somatic and rare variation.

Pathway association

Pathway association methods can be broadly categorized to: first, canonical pathway based analyses; and second, *de novo* pathway discovery methods (see Figure 1). Works in the first category exploit current knowledge in curated pathway databases such as KEGG or GO to directly associate known pathways with a disease of interest. Common methodologies for this association task start

Figure 1



An outline of pathway-based analyses for SNP data. (1) SNPs associated with a phenotype of interest (red) are identified among all SNPs (yellow). The statistical test for SNP association should correct for the linkage disequilibrium with neighboring SNPs and the large number of SNPs being tested. (2) Significantly associated genes (blue bars) are identified on the basis of the SNP associations. The mapping may account for SNPs that are in some genomic window containing each of the genes, taking into account gene length and the distance from neighboring genes. (3) Canonical pathway association is performed by, for example, testing sets of genes from known pathways for enrichment with associated genes. (4) *De novo* pathway discovery can be aided by projecting the identified genes onto a protein–protein interaction network, zooming in on regions that contain significantly many associated genes.

by mapping SNPs to genes and then employing standard enrichment tests to score pathways, viewed here as gene sets [6–9]. These associations often reveal significant signals that are missed by SNP-based or even gene-based methods. For example, Li *et al.* performed a meta-analysis of SNP data of pancreatic cancer against a list of curated pathways that are relevant to the disease [10•]. They identified a pancreatic development pathway as significantly associated with the disease even after discarding four genes that were previously known to be associated with it. In a comparison of several pathway association methods on lung cancer GWAS data sets [11], the authors point at a variant of SUMSTAT [12] as the best performing method. It is on the basis of computing a chi-square statistic for every gene and averaging these scores over the genes in a pathway.

As pathway knowledge is incomplete [13], works in the second category use the SNP or mutation data to zoom in on specific regions of a physical or functional interaction network. Most studies in this domain try to associate dense regions of a protein–protein interaction network with a phenotype of interest [14–15]. As an example, NETBAG is an approach that was used to identify rare *de novo* CNVs in autism [16••]. The approach is on the basis of identifying genes in CNV regions of the genome, projecting them onto a phenotypic-similarity network and finding connected clusters of such genes with dense interactions. A refinement of this method that integrates multiple types of variation data was applied to a schizophrenia dataset, pinpointing several cohesive gene networks related to axon guidance, neuronal cell mobility, synaptic function and chromosomal remodeling (Figure 2a) [17].

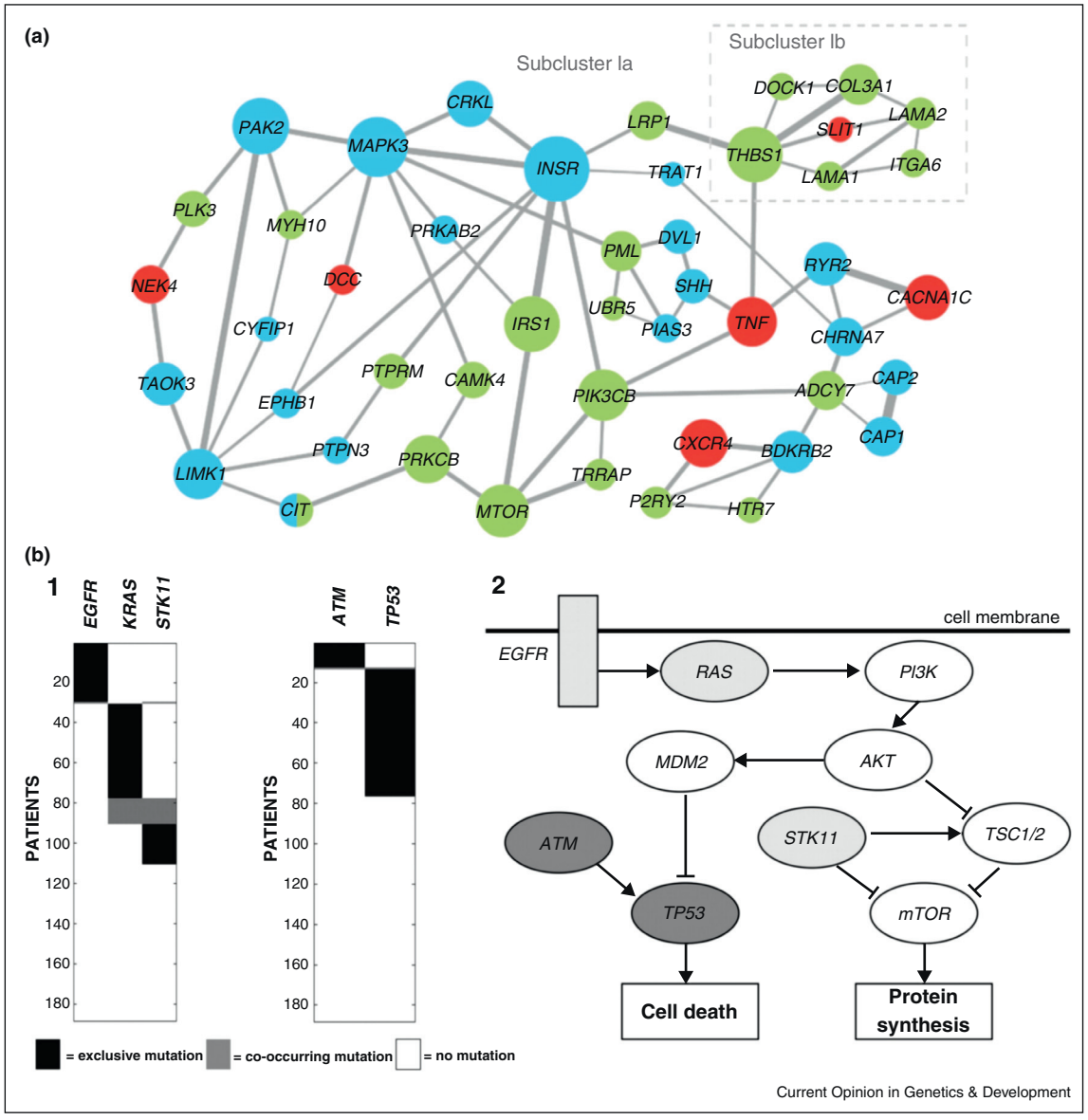
Other works in this category are tailored for specific diseases and are on the basis of particular hypothesized relations between the genomic and network data. For example, Vandin *et al.* [18••] exploit mutation data on cancer patients to reveal cancer-specific driver pathways. Their method is on the basis of the empirical observation that the disease targets specific gene sets, termed pathways, where each pathway is mutated in many patients (high coverage) but most patients have at most one mutation in a given pathway (high exclusivity). The identified gene sets included members of well-known cancer pathways including Rb, p53 and mTOR (Figure 2b).

Computational and statistical challenges

Pathway-based associations are emerging as a powerful alternative to SNP-based and gene-based associations. However, current pathway association methodologies suffer from several problems including SNP to gene mapping and the aggregation of SNP or gene scores to pathway p-values. The problem of associating genomic variation to particular genes is often handled in an ad-hoc manner such as using a cutoff on SNP significance in a predetermined window around the gene [19]. A more principled approach is to directly test for the association of a gene with a trait by considering all 'relevant' SNPs simultaneously, accounting for gene length and the linkage disequilibrium between adjacent markers [20]. The resulting association scores need to be aggregated to produce pathway-level p-values, appropriately calibrated to adjust for LD patterns and co-location of functionally related genes. A recent study handles these challenges by performing random circular permutations of the genome [21]. Such circular permutations offset the associations between SNPs and p-values while maintaining their relative genomic order, thereby preserving gene neighborhoods and LD patterns; they are used to estimate pathway score distribution under the null hypothesis of no association.

When aiming for *de novo* pathway discovery a further challenge has to be met — the combinatorial search over exponentially many pathways for high-scoring ones. A general technique that has proven useful for tackling this problem is the formulation of the search as an integer linear program (ILP, see Box 1). As an example, this

Figure 2



Example applications of *de novo* pathway inference methods. **(a)** The highest scoring cluster found by NETBAG+ (NETwork Based Analysis of Genetic associations) when analyzing genomic variations related to schizophrenia. The main cluster is composed of two subclusters exhibiting enrichment profiles and developmental gene-expression patterns that differ between the two subclusters and, additionally, from the patterns in the entire brain. *De novo* CNVs are denoted in blue, non-synonymous single nucleotide variants in light green; and GWAS implicated genes in red. Node size reflects the contribution to cluster score. Edge width is proportional to the prior likelihood that the two genes contribute to similar phenotypes. Adapted from [17]. **(b)** Analysis of Lung adenocarcinoma data. (1) Mutation in genes uncovered by the analysis (EGFR, RAS, STK11, ATM, TP53) exhibit mutual exclusion patterns in most patients. (2) The five mutated genes that were predicted to be relevant to the disease (out of 356 reported genes) are shown in the context of known pathways. EGFR, RAS and STK11 (light gray) are part of the mTOR pathway; ATM and TP53 (dark gray) are part of the cell-cycle pathway. Adapted from [18••].

approach was successfully used by Leiserson *et al.* [22] to discover multiple pathways that are mutated in cancer, optimizing the coverage and exclusivity of those pathways. Their method recovers sets of interacting genes that overlap known pathways, as well as gene sets containing subtype-specific mutations.

Finally, when searching for pathway associations over a network of protein–protein interactions, one has to statistically assess the network proximity of the associated genes or, more generally, identify sets of associated genes that are significantly close or connected in the network. The former problem can be tackled via standard

Box 1 Integer linear programming (ILP)

Linear programming is a powerful mathamtical framework for representing (and solving) optimization problems. The optimization problem is modeled using a set of real-valued variables. Linear combinations of these variables are used to express an optimization objective and a set of constraints that must be satisfied. Such programs can express many natural problems and solved in an efficient manner.

In integer linear programming, a subset of the variables are constrained to take only integral values. Introducing such integrality constraints makes the problem computationally hard to solve; nevertheless, many instances of these problems originating in real-world data are quickly solved to optimality using dedicated solvers such as CPLEX.

measures like average distance in the network [23]. The latter problem is more challenging and often solved via greedy approaches (e.g. NETBAG described above) or using network diffusion techniques (e.g. [24]). In case a connected pathway is sought, the problem can be formulated as a Steiner tree problem, where one seeks the lowest cost pathway that connects the associated genes. While this problem is known to be computationally hard, standard approximation or ILP-based techniques work well in practice (e.g. [25]).

The road ahead

Future developments in pathway association can benefit from integrating additional data types into the association process. In particular, gene expression data can inform the pathway discovery by: first, revealing differentially expressed genes that are correlated with the disease of interest and can be integrated with the SNP-implicated genes to improve pathway discovery [26]; and second, pinpointing SNPs that are associated with the expression of certain genes, thus overcoming the SNP to gene mapping problem [27].

Another venue for improvement concerns the protein network being used. Most current methods use a generic, static view of the network, ignoring its dynamics across different conditions or even individuals. Recent work has demonstrated the benefits in using context-specific networks, for example, for prioritizing disease genes in a tissue-specific manner [28]. Such networks can be inferred with the aid of gene expression data [29] or using direct experimentation [30].

A third venue for improvement concerns the refinement of network data with structure-based information on the interacting domains [31,32]. Specifically, Wang *et al.* [32] have applied three-dimensional docking algorithms to identify the interfaces between interacting proteins. Their findings suggest that disease-causing mutations are more likely to occur in binding domains so their ability to physically interact is impeded. Consequently, the association of genomic variants to binding domains may directly highlight interactions, rather than genes, as affected by the disease. This in turn, combined with growing experimental information on the structure of interacting proteins, can improve both the association of known pathways and the discovery of novel ones.

In this scientific renaissance of population genomics, computational and experimental methods play an essential role in interpreting the data being produced and deriving novel biological and medical insights. Key to this effort is the move from gene-level to pathway-level and network-level analyses. With increasing data on molecular pathways and condition-specific networks, pathway association is expected to become more powerful than ever before.

Acknowledgements

NA was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. SI was supported by an NSF grant 1321000. RS was supported by an I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant no. 757/12).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Visscher PM *et al.*: **Five years of GWAS discovery**. *Am J Hum Genet* 2012, **90**:7-24.
2. McClellan J, King MC: **Genetic heterogeneity in human disease**. *Cell* 2010, **141**:210-217.
3. Manolio TA *et al.*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**:747-753.
4. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies**. *Nat Rev Genet* 2010, **11**:843-854.
5. Ideker T *et al.*: **Discovering regulatory and signalling circuits in molecular interaction networks**. *Bioinformatics* 2002, **18(Suppl 1)**: S233-S240.
6. Holmans P *et al.*: **Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder**. *Am J Hum Genet* 2009, **85**:13-24.
7. Weng L *et al.*: **SNP-based pathway enrichment analysis for genome-wide association studies**. *BMC Bioinformatics* 2011, **12**:99.
8. Uzun A *et al.*: **Pathway-based genetic analysis of preterm birth**. *Genomics* 2013, **101**:163-170.
9. Donato M *et al.*: **Analysis and correction of crosstalk effects in pathway analysis**. *Genome Res* 2013 http://dx.doi.org/10.1101/gr.153551.112. (in press).
10. • Li D *et al.*: **Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer**. *Carcinogenesis* 2012, **33**:1384-1390.
The authors demonstrate the power of canonical pathway association by showing that the GWAS data is sufficient for the pathway association task, even when ignoring genes that were previously etablished as significantly associated with the disease.
11. Fehringer G *et al.*: **Comparison of pathway analysis approaches using lung cancer GWAS data sets**. *PLoS ONE* 2012, **7**: e31816.

12. Tintle NL *et al.*: **Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16**. *BMC Proc* 2009, **3(Suppl 7)**:pS96.

13. Califano A *et al.*: **Leveraging models of cell regulation and GWAS data in integrative network-based association studies**. *Nat Genet* 2012, **44**:841-847.

14. Vanunu O *et al.*: **Associating genes and protein complexes with disease via network propagation**. *PLoS Comput Biol* 2010, **6**:e1000641.

15. Vandin F *et al.*: **Discovery of mutated subnetworks associated with clinical data in cancer**. *Pac Symp Biocomput* 2012:55-66 http://dx.doi.org/10.1142/9789814366496_0006. Chapter number (6).

16. •• Gilman SR *et al.*: **Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses**. *Neuron* 2011, **70**:898-907.
The authors integrate heterogenous data to create a phenotypic-network model that complements the genome wide association data. Subsequent analysis uncovers dense regions in the network containing genes that are supported by both the phenotype and genotype information as relevant to the disease.

17. Gilman SR *et al.*: **Diverse types of genetic variation converge on functional gene networks involved in schizophrenia**. *Nat Neurosci* 2012, **15**:1723-1728.

18. •• Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer**. *Genome Res* 2012, **22**:375-385.
The authors use disease-specific insights to analyze GWAS data in cancer. Their model differentiates driver and somatic mutations and allows elucidating *de novo* pathways that underlie the disease.

19. Gudbjartsson DF *et al.*: **Many sequence variants affecting diversity of adult human height**. *Nat Genet* 2008, **40**:609-615.

20. Liu JZ *et al.*: **A versatile gene-based test for genome-wide association studies**. *Am J Hum Genet* 2010, **87**:139-145.

21. Cabrera CP *et al.*: **Uncovering networks from genome-wide association studies via circular genomic permutation**. *G3 (Bethesda)* 2012, **2**:1067-1075.

22. Leiserson MDM, Blokh D, Sharan R, Raphael BJ: **Simultaneous identification of multiple driver pathways in cancer**. *PLoS Comput Biol* 2013, **9(5)**:e1003054 http://dx.doi.org/10.1371/journal.pcbi.1003054.

23. Said MR *et al.*: **Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae***. *Proc Natl Acad Sci U S A* 2004, **101**:18006-18011.

24. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer**. *J Comput Biol* 2011, **18**:507-522.

25. Liu Y *et al.*: **Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data**. *BMC Syst Biol* 2012, **6(Suppl 3)**:pS15.

26. Jia P, Liu Y, Zhao Z: **Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer**. *BMC Syst Biol* 2012, **6(Suppl 3)**:pS13.

27. Zhong H *et al.*: **Integrating pathway analysis and genetics of gene expression for genome-wide association studies**. *Am J Hum Genet* 2010, **86**:581-591.

28. Magger O *et al.*: **Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks**. *PLoS Comput Biol* 2012, **8**:e1002690.

29. Barshir R *et al.*: **The TissueNet database of human tissue protein–protein interactions**. *Nucleic Acids Res* 2013, **41(Database issue)**:D841-D844.

30. Zhong Q *et al.*: **Edgetic perturbation models of human inherited disorders**. *Mol Syst Biol* 2009, **5**:321.

31. Kuzu G *et al.*: **Constructing structural networks of signaling pathways on the proteome scale**. *Curr Opin Struct Biol* 2012, **22**:367-377.

32. Wang X *et al.*: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease**. *Nat Biotechnol* 2012, **30**:159-164.