## Title

Multidimensional Integrative Genomics Approaches to Dissecting Cardiovascular Disease



## Authors

Arneson, Douglas
Shu, Le
Tsai, Brandon
et al.

## Publication Date

2017

## DOI

10.3389/fcvm.2017.00008

Peer reviewed

frontiers in Cardiovascular Medicine

MINI REVIEW
published: 27 February 2017
doi: 10.3389/fcvm.2017.00008

# Multidimensional Integrative Genomics Approaches to Dissecting Cardiovascular Disease

*Douglas Arneson[1,2†], Le Shu[1,3†], Brandon Tsai[1], Rio Barrere-Cain[1], Christine Sun[1] and Xia Yang[1,2,3,4,5*]*

[1] Department of Integrative Biology and Physiology, University of California Los Angeles, Los Angeles, CA, USA, [2] Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, USA, [3] Molecular, Cellular, and Integrative Physiology Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, USA, [4] Institute for Quantitative and Computational Biosciences, University of California Los Angeles, Los Angeles, CA, USA, [5] Molecular Biology Institute, University of California Los Angeles, Los Angeles, CA, USA

OPEN ACCESS

**Edited by:**
Tanja Zeller,
University of Hamburg, Germany

**Reviewed by:**
Frank Kramer,
Universitätsmedizin Göttingen, Germany
Melanie Boerries,
German Cancer Research Center (HZ), Germany

***Correspondence:**
Xia Yang
xyang123@ucla.edu

†These authors have contributed equally to this work.

**Specialty section:**
This article was submitted to Cardiovascular Genetics and Systems Medicine, a section of the journal Frontiers in Cardiovascular Medicine

**Received:** 19 December 2016
**Accepted:** 09 February 2017
**Published:** 27 February 2017

**Citation:**
Arneson D, Shu L, Tsai B, Barrere-Cain R, Sun C and Yang X (2017) Multidimensional Integrative Genomics Approaches to Dissecting Cardiovascular Disease. Front. Cardiovasc. Med. 4:8. doi: 10.3389/fcvm.2017.00008

Elucidating the mechanisms of complex diseases such as cardiovascular disease (CVD) remains a significant challenge due to multidimensional alterations at molecular, cellular, tissue, and organ levels. To better understand CVD and offer insights into the underlying mechanisms and potential therapeutic strategies, data from multiple omics types (genomics, epigenomics, transcriptomics, metabolomics, proteomics, microbiomics) from both humans and model organisms have become available. However, individual omics data types capture only a fraction of the molecular mechanisms. To address this challenge, there have been numerous efforts to develop integrative genomics methods that can leverage multidimensional information from diverse data types to derive comprehensive molecular insights. In this review, we summarize recent methodological advances in multidimensional omics integration, exemplify their applications in cardiovascular research, and pinpoint challenges and future directions in this incipient field.

**Keywords: multidimensional omics integration, integrative genomics, cardiovascular disease, genomics, transcriptomics, epigenomics, metabolomics, proteomics**

## INTRODUCTION

Cardiovascular disease (CVD) is a highly prevalent complex disease involving multiple risk factors, pathological changes in diverse cell types, tissues, and organs, and multidimensional molecular perturbations. Common forms of CVD including coronary artery disease (CAD), myocardial infarction, and stroke are among the leading causes of death in the world and therefore demand a better understanding of the etiology. Thanks to the rapid advances of omics technology, we are experiencing an explosion of biomedical data that have the promise to improve our understanding of the molecular underpinnings of clinical phenotypes (1). Accompanying the growing data volume are bioinformatics methodologies and tools to analyze individual data types, as recently reviewed by us and others (2–4).

However, it is increasingly recognized that focusing on any particular type of data only offers limited insights into the mechanistic black box bridging molecular traits and disease phenotypes (5). This is due to the fact that biological processes do not operate through any isolated molecular data type but manifest collectively as molecular cascades and interactions across omics domains to

affect CVD etiology. Only comprehensive integration of multi-dimensional omics data can effectively capture a holistic view of pathogenic mechanisms.

Through recent efforts directly addressing this critical need, a number of integrative genomics approaches have been developed to model the interplays of data from multiple omics domains in a step-wise or meta-analytical fashion (6–8). The mathematical foundations of various integrative methods (9) and the principles and applications of such methods in cancer-related domains (10) have been previously reviewed. These methodological advances have significantly improved our ability to leverage the available rich data to recapitulate the flow of regulatory signals from the genetic background to the eventual disease outcome. Multidimensional analysis also has the built-in advantage of filtering away noise through the aggregation of biological information from independent and diverse sources. Pioneering efforts applying multidimensional data integration have led to numerous novel discoveries of biomarkers, disease pathways, and potential therapeutic targets for CVD (4, 11–15).

In this article, we focus primarily on multidimensional integrative methods applicable to CVD. We first provide an overview of the basic data types and principles of multidimensional data integration and then summarize methodologies and tools along with their representative applications in CVD. Lastly, we summarize the remaining challenges in the field and point to future research directions to improve the effectiveness of multidimensional data integration.

## OMICS DATA TYPES AND BIOLOGICAL RELATIONSHIPS BETWEEN DATA TYPES

The most common omics data types representing the various molecular domains are genomics, epigenomics, transcriptomics, metabolomics, proteomics, and microbiomics (**Figure 1**). We have recently thoroughly reviewed the basic principles, the commonly used bioinformatics methods to analyze each data type, and their applications in CVD research (16). Briefly, genomics assesses DNA sequence and structural variations including single-nucleotide polymorphisms, insertions and deletions, copy number variations, and inversions. Epigenomics is the measurement of DNA methylation, histone modifications (methylation, acetylation, phosphorylation, DP-ribosylation, and ubiquitination), and non-coding RNAs (microRNAs, long non-coding RNAs, small interfering RNAs) (17). Transcriptomics evaluates the transcriptional activities of all genes, including the expression levels of individual genes and transcripts, as well as alternative splicing. Metabolomics aims to profile the levels and flux of metabolites. Proteomics captures the protein levels as well as post-transcriptional modifications of proteins. Lastly, microbiomics measures the composition of bacterial communities as well as the genome and transcriptome of individual bacterial species. Between the omics dimensions, intrinsic biological relationships exist (**Figure 1**), as detailed in our previous reviews (16, 18). Briefly, genomic and epigenomic variations have the capacity to control or modulate the transcriptome and in turn affect the proteome. Metabolites are products of host proteome, or derived from the gut microbiota, and can modulate the epigenome to affect transcription and translation. Gut microbiota can affect the host immune system and metabolism, which are central to programming many aspects of host activities. These complex cascades and interactions are critical elements for consideration in multidimensional data integration.

## MULTIDIMENSIONAL DATA INTEGRATION METHODOLOGIES AND EXAMPLE APPLICATIONS IN CVD

### Principles of Multidimensional Data Integration

Multidimensional data integration aims to aggregate information from diverse molecular domains into predictive models that can inform on mechanisms underlying pathogenesis or help select composite biomarkers that have diagnostic or prognostic values. A critical and non-trivial consideration for multi-omics integration is data preprocessing, including quality control and data normalization (19, 20). Proper preprocessing is important for removing outliers and non-biological variation within a data type and increasing the biological comparability between data types. To date, a vast majority of the recently implemented multidimensional data integration tools fall into one of the following five broad categories: clustering/dimensionality reduction-based methodologies, predictive modeling approaches, pairwise integration, network-based methodologies, and composite approaches, as summarized in **Figure 1** and **Table 1**. The available methods are mostly designed for specific combinations of data types. The selection of proper methods requires consideration of data-driven statistical patterns and biological interpretability. However, depending on the specific applications, the weight for these two aspects may differ. Therefore, before choosing an appropriate method, it is imperative to first understand the biological question that is being addressed: biomarker discovery or mechanistic insight. For the discovery of diagnostic and prognostic biomarkers, data pattern is the key factor, whereas biological interpretation can be less important. Clustering/dimensionality reduction-based methodologies and predictive modeling methodologies are powerful for this task. For mechanistic studies, however, it is critical to couple intrinsic biological relationships among data types with data pattern searches to facilitate biological interpretation. Typically used methods here are pairwise integration and network-based approaches, although clustering/dimensionality reduction, predictive modeling, and composite methodologies can be used for both applications. In the sections below, we categorize and discuss the tools based on their general but not necessarily exclusive applications: biomarker or mechanism discovery.

### Omics Integration Methodologies for Biomarker Discovery

#### Clustering/Dimensionality Reduction-Based Approaches

Clustering/dimensionality reduction-based approaches have the capacity to transform different data types into a common data space, thus facilitating downstream integration. This can be achieved through graph or kernel-based methods followed by grouping data features into a smaller number

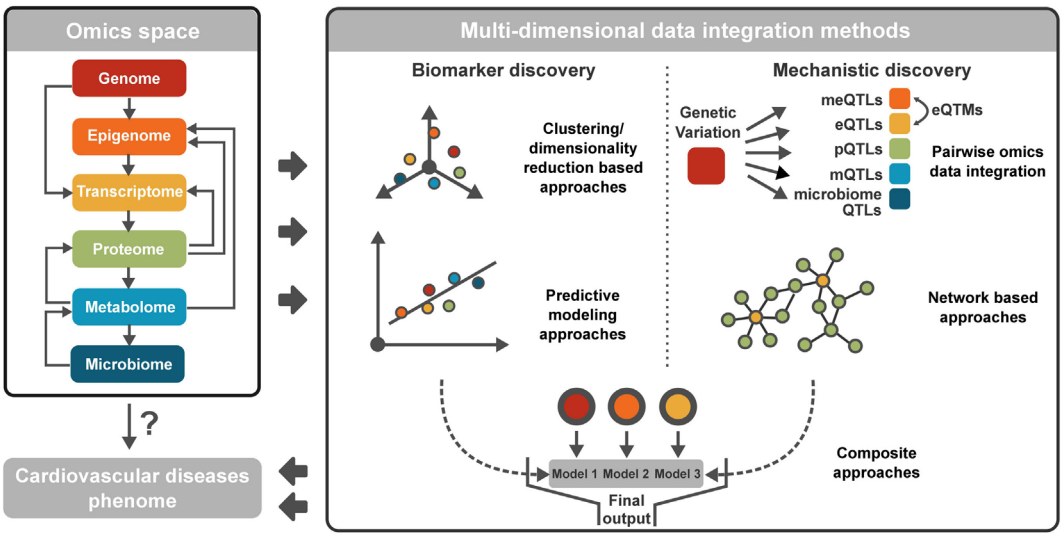


**FIGURE 1 | Summary of different omics data types and multidimensional data integration methods.** Cardiovascular disease (CVD) involves various omics spaces and complex inter-omics interactions. To discover accurate biomarkers and disentangle disease mechanisms of CVD, multidimensional data integration methods are available, broadly categorized into clustering/dimensionality reduction-based approaches, predictive modeling approaches, pairwise omics data integration, network-based approaches, and composite approaches integrating multiple modeling approaches.

**TABLE 1 | Comparison of multidimensional data integration methodologies discussed in the manuscript.**

| Method category | Brief description | Advantages | Limitations | Representative tools |
|---|---|---|---|---|
| Clustering/ dimensionality reduction-based approaches | Transform data into common space through graph or kernel-based methods | Easy to implement using common statistical techniques; retain within-data properties; robust to different units of measurements and different data sets from the public domain | Cross-data interaction may be altered; application limited to visual overview of data and detection of subpopulations | Clustering-based: iCluster (21); ICM (22); TMD (23); SNF (24)<br>Dimensionality reduction: Biofilter (25); CIA/MCIA (26); FALDA (27); GMDR (28) |
| Predictive modeling approaches | Machine learning based methodologies to predict prognosis or diagnosis and discover biomarkers | High predictive power; versatile methodologies; data-driven approach (does not require preexisting knowledge of omics interaction) | Overfitting issue; can require high computational power; does not integrate biological knowledge; higher accuracy requires larger data sets | Camelot (29); Kernel fusion (30); sMBPLS (31); MDI (32); PARADIGM (33); DIVIAN (34) |
| Pairwise omics data integration | Centered on interaction information between pairs of omics data | Easy to implement; reflects inter-omics interaction; causal implication | Available data dominated by expression quantitative trait loci (eQTLs); low robustness of *trans*-association signal | MERLIN (35); RAREMETAL (36); EMMA (37); GEMMA (38); PLINK (39); Matrix eQTL (40); SMR (41) |
| Network-based approaches | Reduce data complexity by converging multi-omics information onto networks | Networks can accommodate multiple layers of data; intuitive depiction and visualization of regulatory circuits | Computationally expensive; difficult to model feedback loops in multidimensional space | Weighted gene coexpression network analysis (42); MEGENA (43); Bayesian networks (44); TIGRESS (45); ARACNE (46); TIE* (47); GENIE3 (48); mixOmics (49) |
| Composite approaches | Flexible integration of multiple integration models | Flexibility and adaptability to diverse research needs | Few well-acknowledged frameworks available | Analysis Tool for Heritable and Environmental Network Associations (50, 51); Mergeomics (3, 52) |

of variables. These approaches are the most straightforward methods to define biomarkers of disease or disease subtypes, thereby facilitating diagnosis and prognosis. The advantages of clustering/dimension reduction include the ability to retain within-data type properties and the robustness to different units of measurement. The drawback, however, is that the transformation of different data types may alter the underlying interaction between data types, even if within-data properties are retained (6).

Clustering-based approaches typically include hierarchical clustering (53), biclustering (54), and k-means clustering (21), which are used to find disease subpopulations (21, 55), refine disease characteristics, and help identify markers (56). Various methods such as iCluster (21), ICM (22), TMD (23), and others have been developed to use clustering for multidimensional integration (**Table 1**). For example, iCluster models the associations between different data types and the structure within each data type to bring the data onto the same feature space allowing for

k-means clustering. This workflow has been applied on breast and lung cancer data sets to identify novel disease subtypes, which cannot be resolved using a single data type (21). We did not identify specific applications of multi-omics clustering in CVD research, although this type of approach has been applied based on individual data types (57–59). Future applications of such approach engaging multidimensional data will facilitate more accurate patient stratification based on multi-omics patterns and help identify unique biomarkers of CVD subtypes.

Dimensionality reduction can be achieved either intrinsically, which scales the dataset of interest using an analytical method, or extrinsically, which uses information outside of the dataset. Intrinsic approaches are the most widely used for dimensionality reduction of genomics data. Standard techniques include principle component analysis, factor analysis, multidimensional scaling, and others, which have been covered in a review of feature selection and extraction methods by Hira and Gillies (60). Tools utilizing dimensionality reduction techniques for multidimensional integration include CIA/MCIA (26), FALDA (27), and others (**Table 1**). Multifactorial dimensionality reduction has been applied by Badaruddoza et al. to identify environmental and genetic interactions in type 2 diabetes and CVD (61).

### Predictive Modeling Approaches

Predictive modeling is another powerful data-driven approach that is primarily utilized for the discovery of composite biomarkers in a multi-omics, big data landscape. In broad terms, it comprises a set of algorithms capable of learning from data to make predictions, which theoretically become more accurate with increasing amount of data. A series of machine learning techniques are commonly implemented, including logistic regression, support vector machines, random forest, neural nets, Bayesian models, and boosting (62) to select the most predictive features. This is typically done through weighting, where the most predictive features contribute more weight to the final model.

Among the various predictive modeling approaches used for multidimensional data integration (**Table 1**), an example is Causal Modelling with Expression Linkage for cOmplex Traits (Camelot) (29). Camelot implements elastic net regression to select the most significant features and uses bootstrapping to reduce the set of features to potential causal genes (29). There has also been widespread usage of machine learning methods in CVD-related fields to identify CVD risk variants (34) and estimate cardiometabolic risks (63). Specifically, Chen et al. trained an ensemble classifier to prioritize non-coding risk variants using multi-omics data and found that the variants associated with repressed chromatin were often the most informative (34). Kupusinac et al. leveraged artificial neural networks to predict cardiometabolic risk using easy to obtain, non-invasive primary risk factors and achieved comparable performance to predictions based on more invasive secondary risk factors (63).

## Omics Integration Methodologies for Mechanistic Discovery

### Pairwise Omics Data Integration

As discussed previously, there are intrinsic biological relationships between data dimensions that can inform on mechanisms, and quantitatively assessing the association between the omics domains can help capture such relationships in a data-driven manner. Pairwise omics data integration is therefore an intuitive and commonly used approach that characterizes interactions between two omics domains. This type of integration comes in two broad categories based on whether genetic information is under consideration (**Figure 1**). The first category is genetics of intermediate traits analysis, in which DNA variants are tested for association with downstream omics markers. The second category is correlation analysis between two non-genetic omics data types (e.g., between metabolites and microbiome).

For genetics of intermediate trait analysis, expression quantitative trait loci (eQTLs) are the most well-known pairwise integration where genetic variations are linked to transcriptomic alterations, achieved through an association test between variants and gene expression levels (64). There are numerous methods available to conduct eQTL analyses such as GEMMA (38) and Matrix eQTL (40), which have been discussed in detail elsewhere (65). Genetic loci can also be associated with omics data types other than transcriptomics, such as methylation quantitative trait loci (66), microRNA QTLs (miR-eQTLs) (67, 68), protein quantitative trait loci (69–71), metabolite quantitative trait loci (72–74), and microbiome quantitative trait loci (75). Correlations between downstream omics data are also informative, although it may be difficult to infer a causal relationship. For example, expression quantitative trait methylation has been defined as the correlation of CpG methylation levels to gene expression (66). This type of analysis can be extended to the other omics data types (e.g., between microbiome and metabolome).

The combination of genetics-based and non-genetic correlative analyses can help infer causality. This concept has been widely used in CVD research to infer candidate causal genes (12, 76–80). Schadt et al. (81) were among the first to develop a formal procedure to incorporate eQTLs, genetic disease association, and gene–trait correlation to infer disease causal genes. Yang et al. (76) applied this approach to identify tissue-specific causal genes for atherosclerotic lesions. Laurila et al. applied a combined approach using both eQTLs and pathway analysis to link genomics, adipose transcriptomics, and lipidomic profiling, highlighting a shift toward inflammatory HDLs in individuals with low HDL (82). Huan et al. (83) combined eQTLs, miRNA-eQTLs, correlative analysis between gene expression and microRNAs, and GWAS to identify microRNA–gene pairs that are putatively causal for CVD. In another effort toward this direction, Zhu et al. proposed a summary data-based Mendelian randomization method that integrates diverse types of QTLs with GWAS to infer candidate genes for complex traits (41).

### Network-Based Approaches

Network approaches have emerged as another powerful platform for multidimensional data integration. Networks depict omics markers as nodes and connections between markers as edges that reflect correlations, regulatory relations, or physical interactions. There are many types of network inference approaches, including regression, mutual information, correlation, and Bayesian networks (44) (**Table 1**). Among the widely used network methodologies, particularly in the CVD field, are correlation-based

methods such as the weighted gene coexpression network analysis (42). These approaches primarily focus on gene expression data and use correlation patterns to group functionally related genes into modules, which significantly reduce the complexity of overlaying other types of omics data onto transcriptomics. It is also feasible to apply these coexpression network approaches to other types of omics data (e.g., DNA methylation data).

In network-based applications, different data types are typically mapped to features (e.g., genes) that can be projected onto networks. For example, Huan et al. integrated coexpression networks with genetic variants to identify causal functional modules for coronary heart disease (78). Yao et al. built an eQTL coexpression network to reveal CVD-related modules (84). Shang et al. inferred a transcription factor regulatory network from blood macrophages transcriptomics profiles and identified a key driver, LIM domain binding 2, for atherogenesis (85). Public network depositories such as protein–protein interaction (86) and BioGRID (87) have also been used to identify novel candidate CVD genes from diverse datasets (88, 89). Recently, the Björkegren group integrated Bayesian networks with CAD genetics and transcriptomics data from CVD relevant tissue types and identified CVD-causal subnetworks and key drivers (80).

### Composite Approaches

Many of the available tools and methods applied to better understand the etiology of a complex disease like CVD utilize combinations of the various principles discussed above (**Table 1**). The integration of the various methods and data types is typically done in a sequential manner where a common overlapping feature (e.g., genes) is used to convert the output of one part of the analysis to be a compatible input for the next step. One example is the Analysis Tool for Heritable and Environmental Network Associations (50, 51), which utilizes neural nets and has been previously applied to predict HDL cholesterol (90). Specifically, this method generates a separate neural net model for each individual data type, and the features with the top predictive power from each model are combined in an integrative model, which possesses higher predictive power than any of the individual models (91). An alternative approach is to employ a majority voting scheme from each of the independent models from the individual omics types, thereby avoiding the additional step of merging multiple models but still leveraging information from multiple data types to predict a clinical outcome (92). As another example, Inouye et al. constructed metabolic networks where metabolites were identified to be associated with the genes identified in the eQTL analysis, thereby layering an additional data modality. The expression levels of the prioritized candidate genes were found to be associated with the phenotypes of the disease, demonstrating the effectiveness of this integrative method (15). Our lab has recently developed a highly generalizable analytical framework, named Mergeomics (3, 52), to more effectively incorporate multidimensional data and various integration strategies. Mergeomics can reveal pathogenic processes underlying diseases by interrogating enrichment patterns from diverse omics association data, and then leverage tissue-specific networks to identify key perturbation points of the significant processes. With this approach, we have prioritized novel regulatory genes and therapeutic targets for CAD and hypertension from diverse genomics, transcriptomics, and molecular network resources (12, 79, 93).

## CHALLENGES, GAPS, AND FUTURE DIRECTIONS

The explosion of omics data in recent years has shifted the bottleneck of scientific discovery from data generation to the need for efficient multidimensional integrative methods. As summarized in this review, there have been major progresses in the development of methodologies and tools that can accommodate and integrate multidimensional data, and the application of these integrative approaches have yielded significant insights into the complex etiology of CVD. However, this field is still in its infancy, and the flexibility, effectiveness, and robustness of data integration to extract biological insights is still restricted. The limitations are mainly due to the intrinsic complexity within individual datasets and between datasets, as well as technical difficulties in integrative modeling that accurately captures true biological complexity. Moreover, there is currently no optimal tool with broad applicability in varying analytical scenarios, as most tools are tailored to particular applications and are limited in data type coverage, thus restricting their generalizability. Further, the performance of the various methodologies has not been comprehensively compared, and there is a lack of general guidance in the field on best practices. To address these challenges, future efforts should focus on intimate collaborations between computational biologists, systems biologists, and experimental biologists in the following areas. First, there is a need for a comprehensive map of data types and data relations, application scenarios, and the desired outcomes. Such a map will facilitate the design of flexible and generalizable multidimensional integration methods. For example, clear differentiation of diagnostic, mechanistic, and therapeutic needs will help choose more appropriate algorithms. Second, comprehensive testing and evaluation of various statistical and mathematical models and computational algorithms are needed to document the performance. The recent effort on network method comparison *via* crowd sourcing is one of the first demonstrations of the value of this approach (94). Performance evaluation should also go beyond *in silico* studies to engage bench scientists to systematically test predictions from the modeling studies to help refine the computational methods. With growing interests and coordinated efforts, multidimensional omics integration will be the next wave of modern biology to help dissect major complex diseases like CVD by promising a holistic understanding of disease pathogenesis and more accurate and personalized diagnostic and prognostic markers.

## AUTHOR CONTRIBUTIONS

DA, LS, BT, RB-C, CS, and XY drafted and edited the manuscript.

## ACKNOWLEDGMENTS

The authors thank Dr. Yuqi Zhao and Dr. Zeyneb Kurt for their contribution in reviewing the manuscript.

## FUNDING

DA is funded by the NIH-NCI National Cancer Institute T32CA201160. LS is funded by UCLA Eureka Scholarship, Hyde Fellowship, and China Scholarship Council. XY is supported by NIH/NIDDK R01DK104363, AHA 13SDG17290032, AHA CVGPS Pathway Grant, and the Leducq Foundation Transatlantic Networks of Excellence Grant.

## REFERENCES

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* (2016) 17:333–51. doi:10.1038/nrg.2016.49
2. Meng Q, Mäkinen V-PP, Luk H, Yang X. Systems biology approaches and applications in obesity, diabetes, and cardiovascular diseases. *Curr Cardiovasc Risk Rep* (2013) 7:73–83. doi:10.1007/s12170-012-0280-y
3. Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics* (2016) 17:874. doi:10.1186/s12864-016-3198-9
4. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. *Nat Rev Genet* (2014) 15:34–48. doi:10.1038/nrg3575
5. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol* (2014) 8(Suppl 2):I1. doi:10.1186/1752-0509-8-S2-I1
6. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* (2015) 16:85–97. doi:10.1038/nrg3868
7. Sun YV, Hu Y-JJ. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv Genet* (2016) 93:147–90. doi:10.1016/bs.adgen.2015.11.004
8. Rotroff DM, Motsinger-Reif AA. Embracing integrative multiomics approaches. *Int J Genomics* (2016) 2016:1715985. doi:10.1155/2016/1715985
9. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* (2016) 17(Suppl 2):15. doi:10.1186/s12859-015-0857-9
10. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Børresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* (2014) 14:299–313. doi:10.1038/nrc3721
11. Krishnan A, Taroni JN, Greene CS. Integrative networks illuminate biological factors underlying gene–disease associations. *Curr Genet Med Rep* (2016) 4:155–62. doi:10.1007/s40142-016-0102-5
12. Zhao Y, Chen J, Freudenberg JM, Meng Q, Rajpal DK, Yang X. Network-based identification and prioritization of key regulators of coronary artery disease loci. *Arterioscler Thromb Vasc Biol* (2016) 36:928–41. doi:10.1161/ATVBAHA.115.306725
13. Talukdar HA, Foroughi Asl H, Jain RK, Ermel R, Ruusalepp A, Franzén O, et al. Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst* (2016) 2:196–208. doi:10.1016/j.cels.2016.02.002
14. Meng Q, Ying Z, Noble E, Zhao Y, Agrawal R, Mikhail A, et al. Systems nutrigenomics reveals brain gene networks linking metabolic and brain disorders. *EBioMedicine* (2016) 7:157–66. doi:10.1016/j.ebiom.2016.04.008
15. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, et al. Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet* (2012) 8:e1002907. doi:10.1371/journal.pgen.1002907
16. Shu L, Arneson D, Yang X. Bioinformatics principles for deciphering cardiovascular diseases. *Encycl Cardiovasc Res Med* (Forthcoming).
17. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. *Cell* (2007) 128:635–8. doi:10.1016/j.cell.2007.02.006
18. Zhao Y, Barrere-Cain RE, Yang X. Nutritional systems biology of type 2 diabetes. *Genes Nutr* (2015) 10:481. doi:10.1007/s12263-015-0481-3
19. Kohl M, Megger DA, Trippler M, Meckel H, Ahrens M, Bracht T, et al. A practical data processing workflow for multi-OMICS projects. *Biochim Biophys Acta* (2014) 1844:52–62. doi:10.1016/j.bbapap.2013.02.029
20. Chawade A, Alexandersson E, Levander F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res* (2014) 13:3114–20. doi:10.1021/pr401264n
21. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* (2009) 25:2906–12. doi:10.1093/bioinformatics/btp543
22. He S, He H, Xu W, Huang X, Jiang S, Li F, et al. ICM: a web server for integrated clustering of multi-dimensional biomedical data. *Nucleic Acids Res* (2016) 44:W154–9. doi:10.1093/nar/gkw378
23. Savage RS, Ghahramani Z, Griffin JE, de la Cruz BJ, Wild DL. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics* (2010) 26:i158–67. doi:10.1093/bioinformatics/btq210
24. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* (2014) 11:333–7. doi:10.1038/nmeth.2810
25. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput* (2009) 14:368–79. doi:10.1142/9789812836939_0035
26. Meng C, Culhane A. Integrative exploratory analysis of two or more genomic data-sets. *Methods Mol Biol* (2016) 1418:19–38. doi:10.1007/978-1-4939-3578-9_2
27. Liu Y, Devescovi V, Chen S, Nardini C. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst Biol* (2013) 7:14. doi:10.1186/1752-0509-7-14
28. Xu HM, Sun XW, Qi T, Lin WY, Liu N, Lou XY. Multivariate dimensionality reduction approaches to identify gene-gene and gene-environment interactions underlying multiple complex traits. *PLoS One* (2014) 9:e108103. doi:10.1371/journal.pone.0108103
29. Chen B-JJ, Causton HC, Mancenido D, Goddard NL, Perlstein EO, Pe'er D. Harnessing gene expression to identify the genetic basis of drug resistance. *Mol Syst Biol* (2009) 5:310. doi:10.1038/msb.2009.69
30. De Bie T, Tranchevent LC, van Oeffelen LM, Moreau Y. Kernel-based data fusion for gene prioritization. *Bioinformatics* (2007) 23:i125–32. doi:10.1093/bioinformatics/btm187
31. Li W, Zhang S, Liu C-CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* (2012) 28:2458–66. doi:10.1093/bioinformatics/bts476
32. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* (2012) 28:3290–7. doi:10.1093/bioinformatics/bts595
33. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* (2010) 26:i237–45. doi:10.1093/bioinformatics/btq182
34. Chen L, Jin P, Qin ZS. DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol* (2016) 17:252. doi:10.1186/s13059-016-1112-z
35. Abecasis GRR, Cherny SS, Cookson WO, Cardon LR. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* (2002) 30:97–101. doi:10.1038/ng786
36. Feng S, Liu D, Zhan X, Wing MK, Abecasis GRR. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* (2014) 30:2828–9. doi:10.1093/bioinformatics/btu367
37. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. *Genetics* (2008) 178:1709–23. doi:10.1534/genetics.107.080101
38. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* (2012) 44:821–4. doi:10.1038/ng.2310
39. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* (2007) 81:559–75. doi:10.1086/519795
40. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* (2012) 28:1353–8. doi:10.1093/bioinformatics/bts163

41. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* (2016) 48:481–7. doi:10.1038/ng.3538
42. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* (2008) 9:559. doi:10.1186/1471-2105-9-559
43. Song W-MM, Zhang B. Multiscale embedded gene co-expression network analysis. *PLoS Comput Biol* (2015) 11:e1004574. doi:10.1371/journal.pcbi.1004574
44. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. *Nat Methods* (2012) 9:796–804. doi:10.1038/nmeth.2016
45. Haury A-CC, Mordelet F, Vera-Licona P, Vert J-PP. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol* (2012) 6:145. doi:10.1186/1752-0509-6-145
46. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* (2006) 7(Suppl 1):S7. doi:10.1186/1471-2105-7-S1-S7
47. Statnikov A, Aliferis CF. Analysis and computational dissection of molecular signature multiplicity. *PLoS Comput Biol* (2010) 6:e1000790. doi:10.1371/journal.pcbi.1000790
48. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* (2010) 5:e12776. doi:10.1371/journal.pone.0012776
49. Lê Cao KA, González I, Déjean S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* (2009) 25:2855–6. doi:10.1093/bioinformatics/btp515
50. Turner SD, Dudek SM, Ritchie MD. ATHENA: a knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait loci. *BioData Min* (2010) 3:5. doi:10.1186/1756-0381-3-5
51. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics* (2014) 30:698–705. doi:10.1093/bioinformatics/btt572
52. Arneson D, Bhattacharya A, Shu L, Mäkinen V-PP, Yang X. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC Genomics* (2016) 17:722. doi:10.1186/s12864-016-3057-8
53. Qin LX. An integrative analysis of microRNA and mRNA expression – a case study. *Cancer Inform* (2008) 6:369–79.
54. Lee H, Kong SW, Park PJ. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics* (2008) 24:889–96. doi:10.1093/bioinformatics/btn034
55. Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, Cooper LA, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med* (2015) 372:2481–98. doi:10.1056/NEJMoa1402121
56. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform* (2012) 45:1191–8. doi:10.1016/j.jbi.2012.07.008
57. Joehanes R, Ying S, Huan T, Johnson AD, Raghavachari N, Wang R, et al. Gene expression signatures of coronary heart disease. *Arterioscler Thromb Vasc Biol* (2013) 33:1418–26. doi:10.1161/ATVBAHA.112.301169
58. Feng Q, Liu Z, Zhong S, Li R, Xia H, Jie Z, et al. Integrated metabolomics and metagenomics analysis of plasma and urine identified microbial metabolites associated with coronary heart disease. *Sci Rep* (2016) 6:22525. doi:10.1038/srep22525
59. Draisma HH, Reijmers TH, Meulman JJ, van der Greef J, Hankemeier T, Boomsma DI. Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families. *Eur J Hum Genet* (2013) 21:95–101. doi:10.1038/ejhg.2012.110
60. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics* (2015) 2015:198363. doi:10.1155/2015/198363
61. Badaruddoza N, Barna B, Matharoo K, Bhanwer A. A multifactorial dimensionality reduction model for gene polymorphisms and environmental interaction analysis for the detection of susceptibility for type 2 diabetic and cardiovascular diseases. *Mol Cytogenet* (2014) 7(Suppl 1):116. doi:10.1186/1755-8166-7-S1-P116
62. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. *Brief Bioinform* (2006) 7:86–112. doi:10.1093/bib/bbk007
63. Kupusinac A, Doroslovački R, Malbaški D, Srdić B, Stokić E. A primary estimation of the cardiometabolic risk by using artificial neural networks. *Comput Biol Med* (2013) 43:751–7. doi:10.1016/j.compbiomed.2013.04.001
64. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* (2015) 16:197–212. doi:10.1038/nrg3891
65. Duffy DL. Analysis of quantitative trait loci. *Methods Mol Biol* (2017) 1526:191–203. doi:10.1007/978-1-4939-6613-4_11
66. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* (2017) 49(1):131–8. doi:10.1038/ng.3721
67. Huan T, Rong J, Liu C, Zhang X, Tanriverdi K, Joehanes R, et al. Genome-wide identification of microRNA expression quantitative trait loci. *Nat Commun* (2015) 6:6601. doi:10.1038/ncomms7601
68. Gamazon ER, Innocenti F, Wei R, Wang L, Zhang M, Mirkov S, et al. A genome-wide integrative study of microRNAs in human liver. *BMC Genomics* (2013) 14:395. doi:10.1186/1471-2164-14-395
69. Melzer D, Perry JR, Hernandez D, Corsi A-MM, Stevens K, Rafferty I, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* (2008) 4:e1000072. doi:10.1371/journal.pgen.1000072
70. Stark AL, Hause RJ, Gorsic LK, Antao NN, Wong SS, Chung SH, et al. Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy. *PLoS Genet* (2014) 10:e1004192. doi:10.1371/journal.pgen.1004192
71. Cantu E, Suzuki Y, Diamond JM, Ellis J, Tiwari J, Beduhn B, et al. Protein quantitative trait loci analysis identifies genetic variation in the innate immune regulator TOLLIP in post-lung transplant primary graft dysfunction risk. *Am J Transplant* (2016) 16:833–40. doi:10.1111/ajt.13525
72. Kraus WE, Muoio DM, Stevens R, Craig D, Bain JR, Grass E, et al. Metabolomic quantitative trait loci (mQTL) mapping implicates the ubiquitin proteasome system in cardiovascular disease pathogenesis. *PLoS Genet* (2015) 11:e1005553. doi:10.1371/journal.pgen.1005553
73. Feng J, Long Y, Shi L, Shi J, Barker G, Meng J. Characterization of metabolite quantitative trait loci and metabolic networks that control glucosinolate concentration in the seeds and leaves of *Brassica napus*. *New Phytol* (2012) 193:96–108. doi:10.1111/j.1469-8137.2011.03890.x
74. Alseekh S, Tohge T, Wendenberg R, Scossa F, Omranian N, Li J, et al. Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *Plant Cell* (2015) 27:485–512. doi:10.1105/tpc.114.132266
75. Benson AK. Host genetic architecture and the landscape of microbiome composition: humans weigh in. *Genome Biol* (2015) 16:203. doi:10.1186/s13059-015-0775-1
76. Yang X, Peterson L, Thieringer R, Deignan JL, Wang X, Zhu J, et al. Identification and validation of genes affecting aortic lesions in mice. *J Clin Invest* (2010) 120:2414–22. doi:10.1172/JCI42742
77. Yang X. Use of functional genomics to identify candidate genes underlying human genetic association studies of vascular diseases. *Arterioscler Thromb Vasc Biol* (2012) 32:216–22. doi:10.1161/ATVBAHA.111.232702
78. Huan T, Zhang B, Wang Z, Joehanes R, Zhu J, Johnson AD, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arterioscler Thromb Vasc Biol* (2013) 33:1427–34. doi:10.1161/ATVBAHA.112.300112
79. Mäkinen V-PP, Civelek M, Meng Q, Zhang B, Zhu J, Levian C, et al. Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet* (2014) 10:e1004502. doi:10.1371/journal.pgen.1004502
80. Franzén O, Ermel R, Cohain A, Akers NK, Di Narzo A, Talukdar HA, et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* (2016) 353:827–30. doi:10.1126/science.aad6970
81. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* (2005) 37:710–7. doi:10.1038/ng1589
82. Laurila PP, Surakka I, Sarin A-PP, Yetukuri L, Hyötyläinen T, Söderlund S, et al. Genomic, transcriptomic, and lipidomic profiling highlights the

role of inflammation in individuals with low high-density lipoprotein cholesterol. *Arterioscler Thromb Vasc Biol* (2013) 33:847–57. doi:10.1161/ATVBAHA.112.300733

83. Huan T, Rong J, Tanriverdi K, Meng Q, Bhattacharya A, McManus DD, et al. Dissecting the roles of microRNAs in coronary heart disease via integrative genomic analyses. *Arterioscler Thromb Vasc Biol* (2015) 35:1011–21. doi:10.1161/ATVBAHA.114.305176
84. Yao C, Chen BH, Joehanes R, Otlu B, Zhang X, Liu C, et al. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation* (2015) 131:536–49. doi:10.1161/CIRCULATIONAHA.114.010696
85. Shang MM, Talukdar HA, Hofmann JJ, Niaudet C, Asl HF, Jain RK, et al. Lim domain binding 2: a key driver of transendothelial migration of leukocytes and atherosclerosis. *Arterioscler Thromb Vasc Biol* (2014) 34:2068–77. doi:10.1161/ATVBAHA.113.302709
86. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database – 2009 update. *Nucleic Acids Res* (2009) 37:D767–72. doi:10.1093/nar/gkn892
87. Chatr-Aryamontri A, Breitkreutz B-JJ, Heinicke S, Boucher L, Winter A, Stark C, et al. The BioGRID interaction database: 2013 update. *Nucleic Acids Res* (2013) 41:D816–23. doi:10.1093/nar/gks1158
88. Wang Z, Guo D, Yang B, Wang J, Wang R, Wang X, et al. Integrated analysis of microarray data of atherosclerotic plaques: modulation of the ubiquitin-proteasome system. *PLoS One* (2014) 9:e110288. doi:10.1371/journal.pone.0110288
89. Li H, Gordon SM, Zhu X, Deng J, Swertfeger DK, Davidson WS, et al. Network-based analysis on orthogonal separation of human plasma uncovers distinct high density lipoprotein complexes. *J Proteome Res* (2015) 14:3082–94. doi:10.1021/acs.jproteome.5b00419
90. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD. ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pac Symp Biocomput* (2013) 18:385–96. doi:10.1142/9789814447973_0038
91. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Min* (2013) 6:23. doi:10.1186/1756-0381-6-23
92. Drăghici S, Potter RB. Predicting HIV drug resistance with neural networks. *Bioinformatics* (2003) 19:98–107. doi:10.1093/bioinformatics/19.1.98
93. Huan T, Meng Q, Saleh MA, Norlander AE, Joehanes R, Zhu J, et al. Integrative network analysis reveals molecular mechanisms of blood pressure regulation. *Mol Syst Biol* (2015) 11:799. doi:10.15252/msb.20145399
94. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods* (2016) 13:310–8. doi:10.1038/nmeth.3773

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.