Bioinformatics Strategies for the Analysis and Integration of Large-Scale Multiomics Data

Tesi, Niccolo'; van der Lee, Sven; Hulsman, Marc; Holstege, Henne; Reinders, Marcel

DOI
10.1093/gerona/glad005

Publication date
2023

Document Version
Final published version

Published in
The journals of gerontology. Series A, Biological sciences and medical sciences

Citation (APA)
Tesi, N., van der Lee, S., Hulsman, M., Holstege, H., & Reinders, M. (2023). Bioinformatics Strategies for the Analysis and Integration of Large-Scale Multiomics Data. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *78*(4), 659-662. https://doi.org/10.1093/gerona/glad005

Journals of Gerontology: Medical Sciences
cite as: J Gerontol A Biol Sci Med Sci, 2023, Vol. 78, No. 4, 659–662
https://doi.org/10.1093/gerona/glad005

Editorial

Bioinformatics Strategies for the Analysis and Integration of Large-Scale Multiomics Data

Niccolo' Tesi, PhD,[1,2,*] Sven van der Lee, MD, PhD,[2,3] Marc Hulsman, PhD,[2,3] Henne Holstege, PhD,[2,3] and Marcel Reinders, PhD[1]

[1]Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands. [2]Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands. [3]Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands.

*Address correspondence to: Niccolo' Tesi, PhD, Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands. E-mail: n.tesi@amsterdamumc.nl

The authors of the manuscript *Identification of five potential predictive biomarkers for Alzheimer's disease by integrating the unified test for molecular signatures and weighted gene co-expression network analysis* in this issue of the Medical Sciences Section of the *Journals of Gerontology Series A* have exploited a number of different bioinformatic tools to analyze the vast amount of data that they were confronted with (1). This is increasingly happening as it is becoming easier and cheaper to generate comprehensive molecular and phenotypic data with which biological hypotheses can be sharpened. Here, we will give a basic understanding of the methods used by Zhou et al., which are becoming standard practices when analyzing high-throughput biological data (1).

Quantitative Trait Loci

A *quantitative trait locus* (*QTL*) analysis is a statistical method that associates phenotypic data that may be continuous (height, weight, blood pressure, gene expression) or binary (disease status) with genetic markers (2). A well-known example of this analysis is a genome-wide association study (GWAS). In GWAS, the frequencies of single-nucleotide polymorphisms (SNPs) across the genome are compared between a group of individuals that exhibit a given trait such as a disease (so-called cases), and a group of individuals in which the trait is absent (so-called controls). The statistical method then scores the strength of association between each SNP and the given trait, that is, it scores whether the SNP frequency is significantly different between cases and controls (3,4). It is important to realize that because millions of SNPs are tested in a GWAS, the chance of falsely denoting a significant SNP–trait association increases considerably. Therefore, stringent corrections on the scores need to be done, which are known as multiple testing corrections. A simple approach to do so is to divide the score by the number of tests done. In GWAS, as correlation patterns exist between SNPs, it is commonly assumed that 1 million different tests are executed, resulting in a corrected score (p value) of 5×10^{-8} (0.05/1,000,000) (3,4). Consequently, often a large number of individuals are necessary to reach enough statistical power to be able to detect significant associations. Importantly, even when significant SNP–trait associations are found, it is not always easy to decipher the functional effect of the relative SNPs.

Expression quantitative trait loci (*eQTLs*) is a specific quantitative trait locus analysis in which an association is tested between the presence of an SNP and the expression level of a gene, thus not a phenotypic trait of the individual, but a molecular trait (Figure 1A) (5). Finding these associations is especially interesting because it helps linking SNPs to genes, and consequently provides a way to functionally interpret the effect of a SNP. For this reason, eQTL analyses have become a standard procedure in GWAS downstream analyses.

Because eQTL analyses require, for the same sample, knowledge about the SNP genotypes as well as the gene expression levels, eQTL analyses are often based on public repositories. A well-known repository for this purpose is the *GTEx consortium* (*Genotype Transcription EXpression*), the largest publicly available archive of QTL interactions identified using 838 individuals and 49 different tissues (6).

Although eQTL analyses focus on the association with gene expressions, any other molecular trait can be used, such as protein expression (*protein QTL*, *pQTL*), gene-isoforms expression (*splicing QTL*, *sQTL*), DNA-methylation level (*meQTL*), or metabolic levels (*mQTL*). As a consequence, QTL analyses represent a powerful instrument to functionally annotate SNPs, and prioritize affected genes and biological pathways that are likely to play a role in a given trait.

An alternative QTL analysis that is gaining popularity is known as *TWASs* (*transcriptome-wide association studies*). TWAS aims to identify

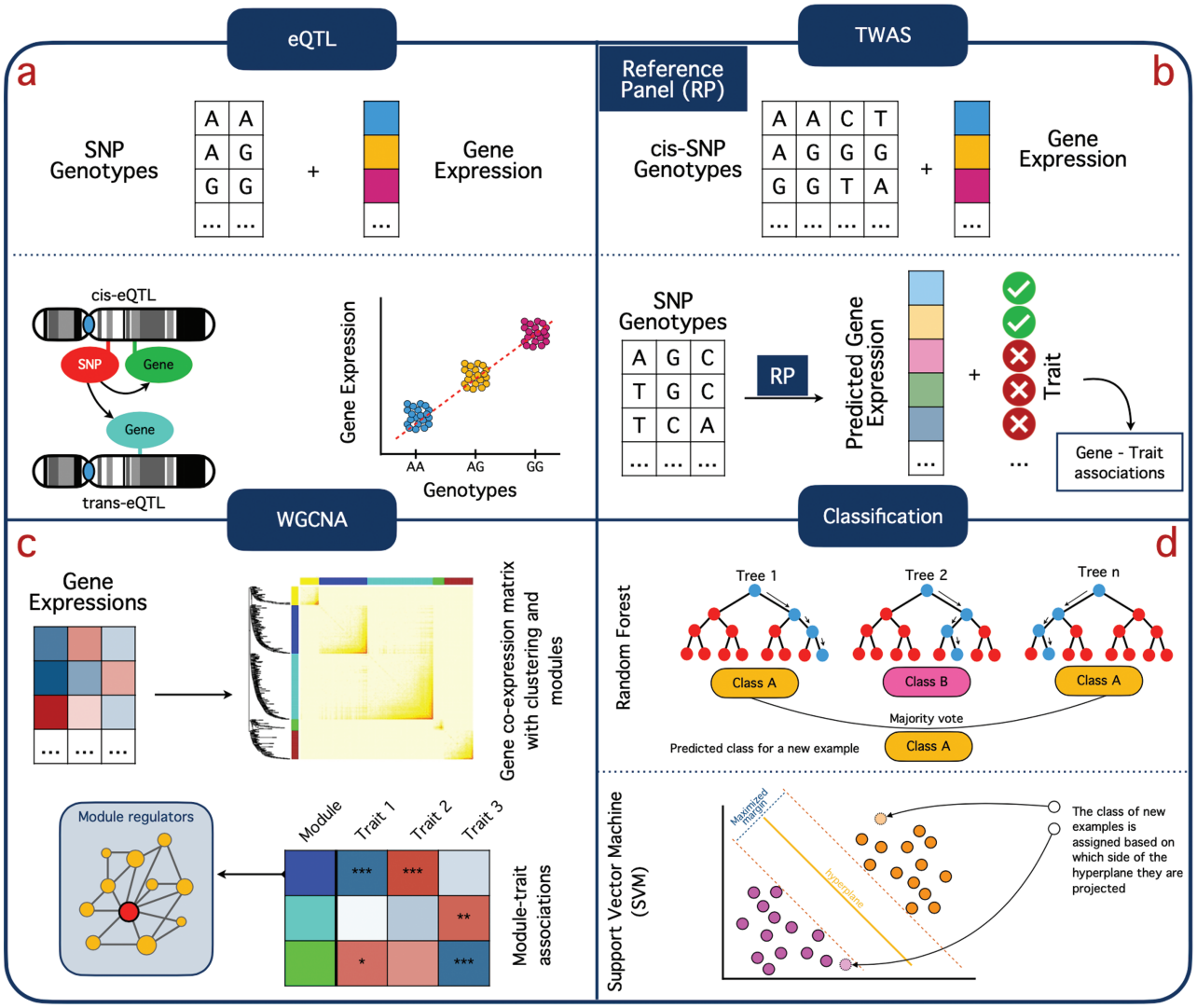

Figure 1. Summary of the main computational methods used by Zhou and colleagues. Panel A defines expression quantitative trait loci (eQTLs), associations between single-nucleotide polymorphism (SNP) genotypes and gene expressions. eQTLs in which the SNPs and their associated genes are located nearby are referred to as local eQTLs or cis-eQTLs. By contrast, those located distantly, often on different chromosomes, are referred to as distant eQTLs or trans-eQTLs. Panel B shows a summary representation of a transcriptome-wide association study: a model that predicts gene expression from nearby SNP genotypes is learned from a reference panel in which both SNP genotypes and gene expressions are known. This information is then used to predict the gene expressions for unobserved samples (using either SNP genotypes or genome-wide association study summary statistics). Finally, the predicted gene expressions are associated with the trait of interest, for example, through a differential expression analysis. Panel C shows the main idea behind weighted gene co-expression network analysis: gene expression profiles are used to construct a correlation-based matrix, which is further characterized using hierarchical clustering to find interacting genes. Sets of interacting genes are referred to as modules. These modules can be associated with the trait of interest or additional phenotypic data, or can be analyzed to find general regulators and biological pathways. Panel D shows, in a nutshell, the principles of Random Forest (RF) and Support Vector Machines (SVMs). Both methods learn from a set of examples with known outcomes, the training data. RF classifies new samples by building a multitude of decision trees, each predicting an outcome, which is then merged across all trees. By contrast, SVM maps data to a higher dimensional space and uses a hyperplane to separate samples belonging to different classes. New samples are mapped into the same space, and their class is predicted based on which side of the hyperplane they fall into.

genes whose expression is significantly associated with a given trait, and uses genetic information (SNP genotypes or GWAS summary statistics) as input, without requiring the measurement of gene expressions (as is necessary for eQTL) (7). A TWAS uses public repositories of eQTLs to *learn* models that predict gene expression based on genetic information in the surrounding of a gene (SNP genotypes or GWAS summary statistics). The predicted gene expressions are then subsequently associated with the trait, for example, to find out whether a gene is differentially expressed across the trait (Figure 1B) (7). As the number of tests equals the number of genes (instead of the number of SNPs), now fewer tests are being done, and therefore the power of a TWAS is generally higher compared to a GWAS (as the multiple testing correction is lower). Moreover, these TWAS analyses can even explore tissue-specific associations as the gene expression models can be tissue-specific. The latter is at the basis of *UTMOST* (*Unified Test for MOlecular SignaTures*), a computational framework that performs cross-tissue expression prediction and gene-level association analysis (8). This approach has been used by Zhou and colleagues in their article (1).

QTL analyses are popular in genomic studies. However, it is important to highlight that validation of the findings is crucial to avoid

Box 1: Decision Trees

A decision tree is a tree-like structure characterized by internal nodes, branches, and leaf nodes. Each internal node represents a *test* on a feature (eg, whether the expression of a gene is higher than a given threshold), each branch represents the outcome of the test, and each leaf node represents a class label (ie, the decision taken after testing on all features, for example, the disease status of a sample) (16). During the construction of the tree, the features to be used in the root and internal nodes (eg, *which* gene to use) as well as the relative thresholds (eg, *what* threshold of expression) need to be identified. One way to do so is to select the feature (and the relative threshold parameter) that maximizes the information content (Information Gain method), a measure of how much information a feature provides about a class (17). A feature and a given threshold will be chosen, for example, if they provide a perfect separation between the classes. Because of their implicit simplicity, decision trees are among the most popular machine learning algorithms. However, they tend to perfectly fit all samples in the training data set (overfitting) and may fail to fit additional data or predict future observations. To overcome this, the complexity of the final decision tree can be reduced by removing sections that are noncritical and redundant (pruning). Finally, to increase the robustness of the classifier, multiple decision trees can be constructed by repeatedly resampling from the training data with replacement. This approach is at the basis of Random Forest.

false-positive findings. A potential issue with QTL analyses is that they are often based on a set of individuals that may not be representative of the entire population or may be influenced by a selected set of environmental factors. As a result, detected QTLs cannot be replicated in other studies, also known as false-positive findings. To address this issue, many GWASs are now based on a multistage analysis approach, in which cohorts of individuals are split into a discovery set and a replication set. Because only SNPs exceeding a certain threshold of significance in the discovery phase are subject to replication, this procedure increases the robustness of the analysis and reduces the risk of false-positive findings (3).

Another common method that provides a more reliable estimate of an association is meta-analysis. Meta-analysis is a statistical technique that aggregates the statistics across multiple studies (eg, by a weighted averaging) without the need to have access to each individual sample in each study. It is a powerful technique that can identify patterns and trends that may not be apparent in any of the individual studies. Meta-analysis techniques help to increase the precision of an association and its generalizability, and reduce the risk of false-positive or false-negative findings. However, the power of meta-analysis depends on the quality and consistency of the studies being combined. If the studies being combined are small or have high levels of heterogeneity, the power of the meta-analysis may be limited (9).

Network-Based Methods

Biological systems are often represented as networks or graphs, that is, entities, such as genes, proteins, or metabolites, that interact with each other. For example, proteins that bind with each other, and thus form a complex, or genes that interact with each other within a biological pathway. A popular data-driven approach for exploring such interactions is through correlation-based networks (10). In such a network, entities are connected when they show a strong enough correlation across a set of measurements. For example, genes can be connected when their expressions in a particular tissue are correlated across many individuals. Hubs in these networks (nodes/genes that are connected to many other genes) might then indicate general regulators, such as transcription factors. Alternatively, many network-based methods aim to identify densely connected components (sets of genes that almost all are correlated to each other), which indicate clusters of genes that are likely to be functionally related, such as forming a complex or being part of a biological pathway (10).

The *weighted gene co-expression network analysis* (*WGCNA*) is a network-based framework built on co-expression in which the interaction between genes is not binary (ie, genes are connected or not), but rather based on a weight, for example, the correlation between the genes (10,11). The weights can be reinforced or downweighted using user-defined parameters, after which the resulting similarities are used in a hierarchical clustering scheme to find clusters of interacting entities (Figure 1C). These clusters of co-expressing entities are called modules, which can then be further characterized. For example, a detected module of co-expressed genes can be analyzed by calculating a functional enrichment to test whether the involved genes share a biological pathway. Alternatively, these modules can be related to phenotypic traits, such as patient survival, or a case–control status, with which novel biomarkers and regulators may be discovered.

Finally, different networks can also be contrasted with each other. For example, it is possible to identify changes in connectivity patterns or module structure between different conditions (differential network analysis). Likewise, the identification of shared modules across different networks (consensus network analysis) may highlight the structural building blocks of the networks (10,11).

Similar to QTL analyses, network-based methods are not free from limitations: they typically rely on data from a sample of the population, and they may be sensitive to the choice of modeling assumptions and parameters. Therefore, validation of these findings in an independent test set is always warranted.

Classification

Instead of finding associations between a gene and a trait, or clustering genes to correlate them to a trait, one can also try to learn a predictor for a certain trait based on the measured data. Such predictors can be useful during clinical decision support, for example, to predict the chance that a tumor metastasizes based on all measured protein expressions in a biological sample. These predictors, also known as classifiers, learn from a set of examples for which the final outcome is known (the training set). A simple classifier that does so is the 1-NN (1 Nearest Neighbor) classifier: given a measured gene or protein expression profile of a new patient (the input), it finds the most similar patient in the training set and then returns as a prediction the known outcome of that patient. While the 1-NN classifier only needs to memorize the training data, many more advanced techniques have been proposed with *deep learners* that mimic the neuronal architecture of brains as the most recent examples.

Although many classification methods have been developed, they can roughly be grouped into *discriminative* approaches, *statistical* approaches, and *distance-based* approaches. Discriminative approaches directly model the decision boundary between different classes. Examples of such methods include: logistic regression, which estimates a binary outcome based on a set of predictors (eg, gene expressions) so

that the decision boundary equals those inputs that result into an intermediate value (0.5 in case of a binary outcome). Other approaches in this category include decision trees and Support Vector Machines (SVMs) explained in more detail later (12). Statistical approaches estimate the probability of each of the classes based on a set of predictors (eg, gene expressions). This is realized by modeling the distribution of observations for each of the classes and using Bayes' theorem to combine that into an estimate of the probability for each class. These models are popular because they allow for the incorporation of prior knowledge or beliefs about the observed parameters (eg, known pathways). This can lead to more accurate predictions and better decision making in some cases. One simple example of a Bayesian approach for classification is the Naive Bayes classifier which assumes independence (so no structure) between observed parameters (12). Distance-based approaches rely on measures of distance or similarity to classify samples. One example of a distance-based method for classification is the already mentioned 1-NN classifier (12).

In their article Zhou and colleagues implemented Random Forest as well as SVM classifiers. A Random Forest is a predictor that builds many decision trees (hence the term forest), each predicting an outcome (Box 1). The outcomes of all trees are then merged, for example, by taking a majority vote over all trees (13). Each decision tree builds a hierarchical tree of single-feature decisions: for example, if a feature represents the expression of a gene, it tests whether a single gene measurement is larger than a selected threshold or not (Figure 1D and Box 1). All samples with lower expression follow the left branch of the tree, while those with higher expression go to the right branch. This continues until the samples in a branch are all of the same class. For a new sample to be predicted, it is possible to easily follow a path along the tree (based on its features) that in the end defines the predicted outcome (the class of all samples in that branch). The power of the Random Forest classifier is that it can handle numerical and categorical features, is not highly influenced by outliers, and is suitable for modeling both linear and nonlinear relationships.

SVM is a popular classifier in which the idea is to map each sample to a higher dimensional space defined by its features (eg, gene expressions). SVM then separates this input space with a hyperplane in a way such that the points on one side of the hyperplane are from one class (eg, samples with a disease), and points on the other side belong to the other class (eg, healthy controls; Figure 1D) (14). As there could be multiple hyperplanes separating the 2 classes, SVM chooses the plane that maximizes the distance of the points of both classes to the hyperplane (ie, the maximum margin criterium). This distance is called the margin, and the points that fall exactly on the margins are called the supporting vectors. Although SVM supposedly can separate classes linearly (as opposed to the Random Forest), it also uses a transformation of the original space such that the linear separation in this new space becomes a nonlinear separation in the original space (15). This is the so-called kernel-trick, which has popularized SVMs because in the transformed space only a hyperplane needs to be learned, so very few parameters need to be estimated, which makes the method robust, and it still has good performance with relatively small training sets.

Conclusion

The authors of the manuscript *Identification of five potential predictive biomarkers for Alzheimer's disease by integrating the unified test for molecular signatures and weighted gene co-expression network analysis* have illustrated the importance of bioinformatics and machine learning tools to analyze the ever-increasing availability of high-throughput (molecular) data (1). Luckily, many methods are available as part of common programming languages, such as Python or R, easing their usage. Yet, although an in-depth mathematical knowledge of these methods is often not necessary, it is important to understand their basic principles to know what they can be used for, and what their assumptions and limitations are. Nevertheless, we are strong advocates for more training on these methods and more collaboration with computational biologists and/or statisticians, especially in the early design and planning of a study. Together with new biological insights, this will eventually lead to many new breakthroughs.

Acknowledgments

Niccolò Tesi is appointed at ABOARD, which is a public-private partnership receiving funding from ZonMW Nationaal Dementiaprogramma (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). More than 30 partners participate in ABOARD (https://www.alzheimer-nederland.nl/onderzoek/projecten/aboard/partners). ABOARD also receives funding from Edwin Bouw Fonds and Gieskes-Strijbisfonds.

References

1. Zhou S, Ma G, Luo H, Shan S, Xiong J, Cheng G. Identification of five potential predictive biomarkers for Alzheimer's disease by integrating the unified test for molecular signatures and weighted gene co-expression network analysis. *J Gerontol A Biol Sci Med Sci*. 2022:glac179. doi:10.1093/gerona/glac179
2. Kearsey MJ. The principles of QTL analysis (a minimal mathematics approach). *J Exp Bot*. 1998;49(327):1619–1623. doi:10.1093/jxb/49.327.1619
3. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*. 2019;20(8):467–484. doi:10.1038/s41576-019-0127-1
4. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genom Inform*. 2012;10(2):117–122. doi:10.5808/GI.2012.10.2.117
5. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet*. 2006;7(11):862–872. doi:10.1038/nrg1964
6. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580–585. doi:10.1038/ng.2653
7. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245–252. doi:10.1038/ng.3506
8. Rodriguez-Fontenla C, Carracedo A. UTMOST, a single and cross-tissue TWAS (transcriptome wide association study), reveals new ASD (autism spectrum disorder) associated genes. *Transl Psychiatry*. 2021;11(1):256. doi:10.1038/s41398-021-01378-8
9. Winkler TW, Day FR, Croteau-Chonka DC, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc*. 2014;9(5):1192–1212. doi:10.1038/nprot.2014.071
10. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4(1):17. doi:10.2202/1544-6115.1128
11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf*. 2008;9(1):559. doi:10.1186/1471-2105-9-559
12. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.
13. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. doi:10.1023/A:1010933404324
14. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297. doi:10.1007/bf00994018
15. Theodoridis S, Koutroumbas K. *Pattern Recognition*. 4th ed. Burlington, MA: Academic Press; 2009.
16. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81–106. doi:10.1007/bf00116251
17. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Statist*. 1951;22(1):79–86. doi:10.1214/aoms/1177729694