A FRAMEWORK FOR BIOLOGICAL DATA INTEGRATION AND FEATURE SELECTION IN LARGE DATA SETS

By

Agustin Gonzalez-Reymundez

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Genetics—Doctor of Philosophy

2021

ABSTRACT

A FRAMEWORK FOR BIOLOGICAL DATA INTEGRATION AND FEATURE SELECTION IN LARGE DATA SETS

By

Agustin Gonzalez-Reymundez

The increasing volume of high-dimensional biological data (*omics*) has intensified the discovery of thousands of biomarkers across the different fundamental components of the cell (e.g., genome, transcriptome, proteome, epigenome) and allowed the characterization of complex phenotypes (e.g., metabolome, imaginome, phenome). However, the ability to integrate omics into informative results is constantly challenged by a seemingly ever-increasing volume of data. Furthermore, huge data sizes impose a tradeoff between how complex an omic integration algorithm can be and how much data it can handle (e.g., how fast can the algorithm be scaled to integrate large data sizes). In this dissertation, we explore statistical frameworks to face the challenges of modern omic data, including the integration of high-dimensional data of large sample sizes. We have developed a novel framework of competitive analytical performance compared with existing methods but suitable for omic data reaching biobank scales (i.e., hundreds of thousands of samples and variables). We implemented this method as an R package and showed its application on two traits of a complex molecular basis: cancer and regulation of energy intake and expenditure. In chapter one, we review the technologies and methods used to generate and integrate omic data. Chapter two describes our novel method and software of omic integration, shows examples in synthetic data, and evaluates its computational and analytical performance. Chapter three presents an application of our method to reveal a novel pan-cancer classification of tumors beyond

the tissue of origin, regulated by distinct sets of molecular signatures. In chapter four, we present an application of our method to integrate phenomics data and identify patterns of energy balance regulated by genomic variation. Finally, in chapter five, we offer general conclusions to the entire thesis.

To Elise and Solaris

ACKNOWLEDGEMENTS

A vast number of people contributed to making this thesis possible. Above all, I would like to thank my research advisor, Dr. Ana Ines Vazquez, for her constant support, encouragement, and instrumental professional advice. I also would like to thank the guidance committee members, Dr. Eran Andrechek, Dr. Molly Bray, and Dr. Andrea Doseff, for all their invaluable feedback and contributions. I would also like to thank the members of the *QuantGen* group at Michigan State University, particularly Dr. Gustavo de los Campos, Alexa Lupi, Marco Lopez-Cruz, Fernando Aguate, and Alexander Grueneberg. They greatly contributed to the development of many of the computational and statistical elements presented here. Finally, I would like to thank my fellow students and friends from the Genetics and Genome Sciences Program and former director Dr. Cathy Ernst for their support throughout this process.

This research was funded by my adviser Dr. Vazquez and the Genetics and Genome Science Program, Research Alliance grants from Zoetis, and the NIH grant R01 DK 119836-01A1. Computational resources were provided by Michigan State University's High-Performance Computing Cluster.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1 INTRODUCTION 1

1.1 WHERE DO OMICS COME FROM? 1
1.2. HOW IS OMIC DATA INTEGRATED AND ANALYZED? 6
1.2.1 Matrix decomposition-based methods 6
1.2.2 Network-based methods 8
1.2.3 Correlation-based methods 9
1.2.4 Regression-based methods 10

CHAPTER 2 MULTI-OMIC INTEGRATION WITH SPARSE SINGULAR VALUE DECOMPOSITION 11

2.1. ABSTRACT 11
2.2. INTRODUCTION 12
2.3. MATERIAL AND METHODS 14
2.3.1. Statistical background and algorithms 14
2.3.2. Syntaxis of the main function moss 17
2.3.3. Performance evaluation 21
2.4. RESULTS 23
2.4.1. Example of unsupervised omic integration with MOSS 23
2.4.2. Example of supervised omic integration with MOSS 28
2.4.3. Evaluation of MOSS analytical performance 30
2.4.4. Benchmarking 31
2.5. DISCUSSION 34

CHAPTER 3 MULTI-OMIC SIGNATURES IDENTIFY PAN-CANCER CLASSES OF TUMORS BEYOND TISSUE OF ORIGIN 37

3.1. ABSTRACT 37
3.2. INTRODUCTION 39
3.3. MATERIAL AND METHODS 41
3.3.1. Pan-cancer data 41
3.3.2. Omic integration, clustering, and features selection. 42
3.4. RESULTS 46
3.4.1. Data description 48
3.4.2. Classification of pan-cancer tumors after removing tissue-specific signals 53
3.4.3. Clinical and demographical characterization of tumor clusters 54
3.4.4. Gene signatures characterizing tumor clusters 58

3.5. DISCUSSION 64

CHAPTER 4 PHENOMIC DATA INTEGRATION IN THE UK BIOBANK REVEALS GENETIC VARIANTS INVOLVED IN ENERGY BALANCE 71

4.1. ABSTRACT 71
4.2. INTRODUCTION 72
4.3. MATERIAL AND METHODS 74
4.3.1. Cohort 74
4.3.2. Statistical analysis 77
4.4. RESULTS 82
4.4.1. PEB variables were associated with specific groups of phenotypes. 82
4.4.2. PEB-induced aggrupation of phenotypically distinct participants. 87
4.4.3. Genomic variants associated with PEB. 89
4.5. DISCUSSION 96

CHAPTER 5 CONCLUSIONS 100

APPENDICES 105

APPENDIX A SUPPLEMENTARY MATERIAL FOR CHAPTER 2 106
APPENDIX B SUPPLEMENTARY MATERIAL FOR CHAPTER 3 109
APPENDIX C SUPPLEMENTARY MATERIAL FOR CHAPTER 4 135

BIBLIOGRAPHY 140

LIST OF TABLES

Table 3.1: Data description by cancer type after quality controls 50

Table 3.2: Clusters characterization after removing tissue effects 55

Table 4.1: Descriptive statistics of phenotypical variables..... 84

Table 4.2: Labels for PEB variables based on correlation with original phenotypes. 87

Table A.1: Code templates to replicate examples. 106

Table B.1: List of genes significantly deregulated in at least one pan-cancer cluster.. 109

LIST OF FIGURES

Figure 2.1: Scree plot, PEV trajectory, and partial derivatives 24

Figure 2.2: Omic contribution to selected features. 25

Figure 2.3: Cluster analysis 26

Figure 2.4: Association between clusters and SVD dimensions 27

Figure 2.5: Signature of features for two clusters. 28

Figure 2.6: PEV trajectory plot and partial derivatives for a PLS analysis. 29

Figure 2.7: Output of the function *moss_heatmap.* 30

Figure 2.8: Performance of feature detection with MOSS. 31

Figure 2.9:Benchmarking of omic integration methods 33

Figure 3.1: Omic integration and features selection method……...48

Figure 3:2: Pan-cancer clustering of tumor samples after adjusting for tissue effects…54

Figure 3.3: Gene signatures for Clusters 1 and 4…… 61

Figure 3.4: Gene signatures for Clusters 6, 7, and 8……63

Figure 4.1: Inclusion criteria and sample size 82

Figure 4.2: Relationship between PEB variables and original phenotypes 86

Figure 4.3: PEB-based cluster analysis 89

Figure 4.4: Summary of GWAS loci within genic regions 94

Figure A.1: Analytical performance of several omic integration methods……...108

Figure B.1: Clustering of tumor samples (no constraints)…… 130

Figure B.2: Re-classification of tumors and previously reported molecular subtypes.. 131

Figure B.3: Survival curves by pan-cancer tumor clusters…… 132

Figure B.4: Re-classification of tumors reveals differences in sample type……133

Figure B.5: Expression and copy numbers for transcription factors and targets……...134

Figure C.1: Plots of singular values trajectories 135

Figure C.2: Comparison between dense and sparse latent factors..... 136

Figure C.3: Annotated Manhattan plot for PEB 1 136

Figure C.4: Annotated Manhattan plot for PEB 2 137

Figure C.5: Annotated Manhattan plot for PEB 3 137

Figure C.6: Annotated Manhattan plot for PEB 4 138

Figure C.7: Annotated Manhattan plot for PEB 5 138

Figure C.8: Summary of previously reported information for genes in Figure 4.4……...139

CHAPTER 1

INTRODUCTION

The word “omic” is nowadays used as a generic term to represent collections of biological data obtained with high-throughput technologies (e.g., tandem mass spectrometry, array and sequencing technologies, agriculture imaging, metagenomics). These collections have emerged as a way of describing complex biological systems as a whole [1]. In section “1.1 Where do omics come from?” we review the types of systems that omics characterize and the technologies used to generate omic data. Once omics are measured, their information can be integrated to discover novel biomarkers or understand the interaction between multiple phenotypes [2]. A review of omic integration methods is presented in section “1.2 How is omic data integrated and analyzed?”.

1.1 WHERE DO OMICS COME FROM?

Perhaps the first application of the suffix *–ome* (a mass or totality of something) in genetics was in the word *genome*. The term is attributed to the German botanist H. Winkler, who chose it to represent an organism's haploid set of chromosomes and the genes it contains [3]. It has been suggested that the suffix *–ome* could have been chosen by Winkler as an analogy to broadly used botanical terms, such as microbiome, biome, and rhizome, each one representing an entire collection of biological entities of a certain kind [4].

The word genome gained broader popularity during the context of the apparent *C-value* paradox (the discrepancy between the amounts of DNA of an organism and the amount

needed to encode proteins). After discovering non-coding DNA in the 1940s and the solution of the paradox, the word genome started to be used as a synonym for the entire coding and non-coding DNA [5]. With the advent of Sanger's sequencing method in 1977 (based on the production of all possible DNA fragments from a template, relying on chain termination by modified nucleotides and separation by electrophoresis) [6], it soon became possible to obtain complete genome sequences of microbes. This achievement inspired the creation of *genomics* as a new scientific discipline. The *first generation* of sequencing methods, together with the development of the polymerase chain reaction (PCR) [7] and recombinant DNA technology [8] during the 1980s and early 1990s [9], became instrumental tools for carrying out the Human Genome Project.

During the early 2000s, the first draft of the reference human genome catalyzed the creation of the HapMap [10] and the Encyclopedia of DNA Elements (ENCODE) [11]. The formed aimed to study the common genetic variation between individuals while the latter aimed to annotate all functional elements within the human genome, such as genes and regulatory sequences. Both projects were fundamental for developing genotyping microarray technologies and the first genome-wide studies (GWAS – i.e., the inference of associations between thousands of genomic polymorphisms in a diverse population and a phenotype using linkage disequilibrium). Then, the word genome became a synonym for all the possible DNA sequences of a cell and their variants across populations.

In the late 2000s, massively parallel sequencing technologies (the so-called *second* or *next-generation* – e.g., 454 Life Sciences, Illumina Genome Analyzer, ABI Solid) were developed. These novel technologies allowed lower reactions costs and longer

sequencing reads, enabling the sequencing of whole genomes in large cohorts of participants [12]. Some of these cohorts, like the 1K Genomes in 2008 and the UK Biobank (sequencing 50 thousand individuals) in 2010, have contributed to detecting rare and very rare variants, usually associated with significant health and behavioral traits [13,14]. Thus, the genome was conceived as the entire collection of DNA sequences plus their common and rare variation across populations.

The close relationship between the evolving concept of the term genome, together with the development of high-throughput technologies, greatly influenced the creation of a plethora of analogous fields in biology. Each of this fields aimed to study the genome complement of a particular set of functional elements. The term *transcriptome*, for example, was first proposed by Victor Velculescu, who defined it in 1995 as the entire collection of RNA molecules of an organism [15]. The same year, Velculescu and collaborators introduced *Serial Analysis of Gene Expression* (SAGE), a revolutionary method to compare samples by taking a snapshot of the population of messenger RNA [16]. Previously, parallel gene expression analysis primarily relied on clonal DNA microarrays (i.e., cDNA samples hybridized against oligonucleotides matching known genes) [17]. By contrast, SAGE addressed transcript presence of known and unknown genes in a more accurate way than microarrays (which accuracy can suffer due to artifacts from background noise). With the advent of RNA-seq in 2008, the discovery of novel genes and the assessment of expression levels continued to improve by achieving higher coverage and depths than SAGE [18].

On the other hand, the term *proteome* is attributed to Marc Wilkins, who first used it in the early 1990s to represent the genomic complement of proteins [19]. The study of the

proteome was enhanced by more precise separation techniques (like capillary electrophoresis, liquid and gas electrophoresis), advances in mass spectrometry (for example, the creation of “soft” ionization approaches, such as electrospray ionization – ESI- and matrix-assisted laser desorption ionization –MALDI-) and the introduction of microarray technologies for protein analysis (e.g., immunoassays, functional microarrays, and reverse-phase array) [20]. These methods have vastly improved the ability to identify and quantify novel proteins and their interactions. Furthermore, due to its crucial role in phenotype expression, the proteome is composed of a much broader set of entities than the genome or transcriptome, including all possible peptides, context-dependent functions, and post-translational modifications.

Another example is the term *epigenome*, popularized in the 1990s as a merge between *epigenetics* (a term attributed to Conrad Waddington, who proposed it in the 1940s to describe inheritable traits in response to environmental stimuli) and the suffix *ome*. As in the previous omics, *epigenome* refers to the genomic complement of epigenetic marks. Therefore, the epigenome represents a broader category, composed of elements from all the previous molecular types discussed: DNA (e.g., patterns of DNA methylation), RNA (non-coding regulatory RNAs), and proteins (e.g., histone modification and chromatin remodelers). This diversity of elements imposed the need for a broader set of methods to characterize the epigenome [21]. To study the genome-wide patterns of DNA methylation, for example, methods like DNA restriction endonuclease assay (to compare the relative size of restriction fragments between individuals, depending on the sensibility of the restriction enzymes to methylated residues), and chromatin immunoprecipitation (Chip, using methylation-specific antibodies to isolate methylation fragments) followed by

microarray genotyping or sequencing, were proposed. Another method is the treatment of DNA with bisulfite reaction (i.e., applying sodium bisulfite to turn methylated cytosines into uracil) coupled with array hybridization or next-generation sequencing. Similar Chip methods have also been applied to study RNA-protein interaction (RIP, which removes DNA from samples and captures RNA with specific antibodies). Mass spectrometry has also been widely used to detect histone modification and isoforms.

The explosion of omic data in the last few years has also inspired the utilization of the *-ome* suffix to describe a much broader set of biological entities, derived from high-throughput methods, of high-dimensional nature, or alluding to the totality of items on a system. The term *phenome*, for example, was first proposed by Michael Soulé in 1967 as the collection of all possible phenotypes of an organism. However, with the advent of high-throughput phenotyping, the term is now applied to the set of all high-dimensional phenotypes acquired at an organism-wide scale [22]. Due to advances in phenotyping techniques, many projects have been able to produce extensive phenomic records for different organisms, including humans (e.g., UK Biobank [14]), mice (Euro Phenome [23]), and plants (International Plant Phenotyping for plants [24]). Technological advances that made extensive phenotyping possible include neuroimaging via structural MRI to study neuronal and cognitive functions [25], automated data loggers to record behavioral data [26], and spectroscopic imaging of crop plants to measure thousands of agroeconomic traits [27].

Soon it became clear that single-layer analyses could not truly capture the synergies between molecular factors across omics (e.g., how the expression of a gene is non-linearly affected by mutations and epigenetic alterations). Therefore, an integrative omic

approach has emerged [28]. The purpose of this approach is to represent multiple omics in a rational and unified way, highlighting the variability across subjects while minimizing the redundant signal from groups of related features (e.g., genes in the same ontology or pathway, linkage blocks) [29,30]. Several computational and statistical methods have been proposed to conduct this task. Next, we review some of the most popular algorithms and models available for omic integration.

1.2. HOW IS OMIC DATA INTEGRATED AND ANALYZED?

Omic integration refers to a vast group of techniques, all conceived to explore the combined effects and synergies across different high-throughput types of biological data. Here, we review some of the most popular methods currently available for omic integration, organizing them as 1) matrix decomposition based, 2) graph-based, 3) correlation-based, and 4) regression-based. The first group will include methods that distill an extended matrix of omics X (binding omics blocks by column) into a factor representing variability across subjects and a factor representing the contribution of each omic feature to that structure. The second group will include dimension reduction techniques, considered under the paradigm of graph embedding [31]. The third group comprises methods that explore associations among features within and across omics by explicitly exploiting the correlations between variables. The last group will include methods based on regression, where the response is an entire omic block.

1.2.1 Matrix decomposition-based methods

This group includes techniques that work on an extended omic matrix $X = [X_1 \quad \dots \quad X_L]$ (with $X_{l=1,\dots,L}$ being a matrix representing the l-th omic block attached by columns) and attempt a decomposition of it into two factors. The first factor collapses the

redundancies within and between omics -by creating orthogonal columns representing the independent signals across omic features (that we will call Z)-. The second factor represents the contribution of each omic feature to this combined effect (we will call this factor W). Many methods assume a linear relationship between the two factors of the form $X \approx ZW$. Standard Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and sparse versions fall under this category. Methods such as the Integrated Non-negative Matrix Factorization (iNMF [32]) assume a common Z among omics and minimizes $\|X_l - ZW_l\|_F^2$ (Where W_l are the weights of the l-th omic). Other methods such as Integrative Clustering (iCluster [33]) and Joint and Integrated Variation Explained (JIVE [34]) assume a latent variable model instead. iCluster defines a model $X = ZW + \varepsilon$ that imposes a LASSO penalty on the elements of Z and W. JIVE on the other hand, assumes a model $X = ZW + I + \varepsilon$, where I is an extra term that represents the individual variation within omics. In all these cases, Z and W (or W_l) are estimated by iterative procedures, such as expectation maximization algorithm (EM). Other methods use a Bayesian approach to find clusters of subjects while identifying regulatory modules across omics. These methods assume that X comes from a finite mixture. The components of the mixtures can be taken as clusters, inferred via Markov Chain Monte Carlo (MCMC) methods. The Joint Bayes Factor (JBF [35]) method, for example, assumes a model of the form $X_l = Z_l(W + W_l) + \varepsilon_l$, and imposes sparsity on the elements of W_l and W via a Bernoulli process. The method uses a Gibbs sampler to iteratively find W_l, W, and Z_l. Methods like Multiple Dataset Integration (MDI [36]) and Bayesian Consensus Clustering (BCC [37]), on the other hand, assume a Dirichlet process, where the clusters are represented by an indicator matrix, sampled from a multinomial

distribution. MDI assumes W_l is an indicator matrix of cluster membership for each omic. The contribution of each feature to each cluster is estimated by means of an "assignment probability", assumed to have a Beta prior distribution. BCC extends this idea, but by explicitly modeling adherence to a global cluster.

1.2.2 Network-based methods

Methods in this category can be conceived under the paradigm of graph embedding [31]. Under this framework, data points in a high dimensional space are represented by a graph. The graph's connections represent similarities between points. The goal is then to map points onto a space of smaller dimensions while preserving similarity. This mapping can be represented by $Z = f(X, A)$, where f is a function (which can be linear, non-linear, explicit, or implicit), X is the extended omics matrix, as before, and A is a matrix representing similarities between subjects. This framework includes methods such as Laplacian Eigenmap [38], ISOMAP [39], and Local Linear Embedding [40], together with linear counterparts, such as SVD. For instance, when f is linear, $Z'Z = I$, $A = XX'$, and the features are standardized, Z becomes the principal components of X. The graph embedding framework can also be extended to include non-linear mappings employing the "kernel trick". The intuition behind this is to map X onto a higher dimensional Hilbert space (we can think of $f(X)$ as a new matrix with row vectors in the new feature space), and then performing the linear algorithm in this new feature space (e.g., SVD on $f(X)$). Then, the problem is solved by taking Z as the principal components of $f(X)$. A large body of literature frame this problem under reproducible kernel Hilbert spaces [41]. For example, a kernel matrix K (e.g., $K = (XX' + c)^t$) can be used to define f implicitly, taking A to be equal to $K = f(X)f(X)'$. Then, Z is obtained by computing the eigenvectors of K

(the so called “Kernel PCA”) [42]. Depending on the dimensions of the problem, one can define a separated kernel by layers of information (K_l) and average them during the estimation of Z. Methods like Similarity Network Fusion (SNF [43]) define A as a scaled exponential kernel of the distance between samples. Others, such as regularized Multiple Kernel Learning for Dimension Reduction (rMKL-DR [44]), do so by adding constraints in the contribution of each kernel to Z. The theory also allows to reformulate some supervised methods, such as linear discriminant analysis and support vector machines. For these, A can be rendered to represent the similitudes within and between classes of subjects. Examples of these in omic integration are smooth t-statistics Support Vector Machines (stSVM [45]) and Features Selection Multiple Kernel Learning (FSMKL [46]).

1.2.3 Correlation-based methods

The methods we describe in this section directly exploit the correlations between features and can be considered variations of the traditional canonical correlation analysis (CCV) and partial least squares (PLS). In all cases, each omic block can be modeled as before: $X_l = Z_l W_l + \varepsilon_l$. The problem now is formulated by turning the omic blocks into vectors and multiplying them by the row vectors a_l and $a_{l'}$ $(l \neq l')$, so that $f(a_l Z_l W_l, a_{l'} Z_{l'} W_{l'}) = f(b_l W_l, b_{l'} W_{l'})$ is maximized. Here, f is the correlation function for CCV and a covariance function for PLS. In both cases, $b_\cdot$ and $W_\cdot$ are estimated. Variations of the problem include sparsity constraints during the estimation of $W_\cdot$ (sCCV [47]), and embedded structure of groups of features (ssCCV [48], sgCCV [49]). In PLS, the problem can also be generalized to multiple blocks (MBPLS [50]), with $f\left(b_l W_l, g(\sum_{l' \neq l} b_{l'} W_{l'}, \theta)\right)$, where g is a function of the linear combination of multiple sets of information, and θ is a vector of extra

parameters. Sparse version of this (sMBPLS [51]) add an iteration procedure and LASSO penalties on $b_{\cdot}$, $W_{\cdot}$, and θ.

1.2.4 Regression-based methods

Methods in this section treat the problem of omic integration as one of regression: $Y = f(X, \delta) + \varepsilon$, where the response is an omic block Y, X is one or more omic blocks, and δ is a matrix of coefficients. To exploit the true multivariate nature of the problem, methods such as the Reduced Rank Regression (RRR [52]) assume a linear f and impose restrictions in the rank of δ. This rank condition implies the existence of linear constraints due to dependencies within each omic block, and between them. Therefore, by estimating δ, we can address which features are associated within and between omic blocks. Variations of the method provide sparse solution with biologically more interpretable results (sRRR [53]). In the Bayesian context (BsRRR [54]), a different prior distribution can be assumed separately for each omic block and their effects: $Y = f(\sum_l X_l \, \delta_l)$; $X_l \sim p(X_l|\theta)$; $\delta_l \sim p(\delta_l|\omega)$ (where θ and ω are hyper-parameters estimated from the data or assumed to have a distribution themselves). This approach has the effect of handling different scales by omic (e.g., discrete for SNP, counts for RNA-seq data) and considers different penalties for dealing with high dimensionality.

The methods reviewed offer a general analytical framework to integrate different layers of data effectively. However, the computational performance of many of these methods suffers when data size becomes substantial [55,56]. The following chapters describe our method and R package for omic integration and present applications in two complex molecular basis traits: cancer and energy balance regulation.

CHAPTER 2

MULTI-OMIC INTEGRATION WITH SPARSE SINGULAR VALUE DECOMPOSITION

This chapter was prepared alongside Alexander Grueneberg, Guanqi Lu, Felipe Couto Alves, Gonzalo Rincon, and Ana I. Vazquez.

2.1. ABSTRACT

The availability of multi-layer omics data has drastically increased in the past years. Several methods have been developed to integrate these types of data effectively. However, our ability to integrate increasing volumes of omic data remains limited. This article presents m*ulti-omic integration with Sparse Singular Value Decomposition (MOSS),* a free and open-source R package to integrate multiple and large omics datasets. This package is computationally efficient and offers biological insight through cluster analysis and identification of biologically relevant omic features. Source code is freely available at CRAN.

2.2. INTRODUCTION

Omic data is characterized by many parameters per sample (usually a much larger number of parameters than the sample size, the so-called $p>>n$). Thus, traditional methods (e.g., ordinal least squares) are insufficient to obtain significant insights from this multi-layer, high-dimensional data. To effectively integrate high-dimensional sets of data, novel methods have been developed [33,34,57–61]. These methods typically combine some form of projection onto a lower-dimensional space (e.g., to reveal structure among samples and features) with some form of feature engineering (e.g., to determine what genes or other molecular entities are most informative at explaining the differences and similarities between omics). These methods have profoundly contributed to our understanding of variation in complex traits across diverse levels of regulation (e.g., mutations in coding genes and epigenetic regulation) [62,63].

Thanks to ongoing biobank efforts, omic data also increases the number of available samples, providing higher prediction ability and statistical power [64]. However, more extensive data sizes make computations progressively lengthier or impossible to perform [65]. Moreover, extensive data sizes also compromise parallelizing complex algorithms (e.g., convolutional neural networks) [66].

To handle these limitations, we developed *Multi-omic integration with Sparse Singular Value Decomposition (MOSS)*. *MOSS* is a free and open-source R package that performs data integration and feature selection on large data sets. It combines the flexibility of sparse singular value decomposition (sSVD) with parallel and in-disk computations to accommodate data sizes reaching biobank dimensions. In this article, we describe the package's main capabilities and its mathematical and computational foundations. We

evaluate *MOSS* analytical performance using a realistic simulation of multi-omic data and benchmark it against state-of-the-art methods of omic integration. Instructions on how to download and install MOSS can be found at CRAN, as well as the package's manual, vignette, and additional examples.

2.3. MATERIAL AND METHODS

2.3.1. Statistical background and algorithms

Omic integration models*:* *MOSS* fits a partial least squares (PLS) [67] model, $\mathbf{Q} = \mathbf{W\Sigma} + \boldsymbol{\varepsilon}$, to find elements maximizing the associations between orthogonal projections of an omic working as response (represented by the matrix $\mathbf{Y} = (y_{il})_{i=1,\dots,n}^{l=1,\dots,m}$) and omics working as predictors (represented by the matrix $\mathbf{Z} = \left(z_{ij}\right)_{i=1,\dots,n}^{j=1,\dots,p}$) omics. These projections are represented by matrices $\mathbf{Q} = \mathbf{YU}$ and $\mathbf{W} = \mathbf{ZV}$, where $\mathbf{U} = (u_{lk})_{l=1,\dots,m}^{k=1,\dots,q}$ and $\mathbf{V} = \left(v_{jk}\right)_{j=1,\dots,p}^{k=1,\dots,q}$ are orthonormal columns of loadings. The matrix $\boldsymbol{\varepsilon} = (\varepsilon_{il})_{i=1,\dots,n}^{l=1,\dots,m}$ represents uncorrelated residuals, with $\varepsilon_{il} \sim (0, \sigma_{\varepsilon}^2)$, not following any particular distribution. The PLS is iteratively solved by least squares to find $\mathbf{Q}$,$\mathbf{W}$, and $\mathbf{\Sigma}$. The rows of $\mathbf{Y}$ and $\mathbf{Z}$ are assumed to represent the same individuals or samples, while their columns are assumed to have zero means and unit variances. Data integration enters the model through $\mathbf{Z}$, as a set of normalized omic blocks appended column-wise, such as

$$\mathbf{Z} = \left[\frac{1}{||\mathbf{Z}_1||_2^2}\mathbf{Z}_1 \quad \dots \quad \frac{1}{||\mathbf{Z}_t||_2^2}\mathbf{Z}_t\right]$$

where t is an arbitrary integer representing the number of omic blocks, and $||.||_2^2$ is the square of the Frobenius norm of a matrix.

Models with covariates: To remove the effects of covariates, we use the model $\mathbf{Q} = \mathbf{X\delta} + \mathbf{W\Sigma} + \boldsymbol{\varepsilon}$, where the columns of matrix $\mathbf{X} = \left(x_{ig}\right)_{i=1,\dots,n}^{g=1,\dots,s}$ represent a set of s covariates and $\boldsymbol{\delta} = \left(\delta_{gk}\right)_{g=1,\dots,s}^{k=1,\dots,q}$ represent the effects of covariates on $\mathbf{Q}$. *MOSS* removes the effects of these covariates by pre-multiplying each term of the above equation by $\mathbf{A} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\boldsymbol{T}}\mathbf{X})^{-}\mathbf{X}^{\boldsymbol{T}}$, where $\mathbf{I}_n = \mathbf{diag}(\underbrace{1, \dots, 1}_{n-times})$.

Estimation of parameters: To estimate $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$, *MOSS* minimizes the following loss function:

$$L = \|\mathbf{YU} - \mathbf{ZV\Sigma}\|_2^2 + \lambda_U(\alpha_U\|\mathbf{U}\|_1 + (1 - \alpha_U)\|\mathbf{U}\|_2^2) + \lambda_V(\alpha_V\|\mathbf{V}\|_1 + (1 - \alpha_V)\|\mathbf{V}\|_2^2)$$

[Eq.2-1]

where $\mathbf{Y}$ and $\mathbf{Z}$ are either the original or covariates-adjusted omics. The first term in the above sum is the L2 norm of approximation of $\mathbf{Q}$ by $\mathbf{W\Sigma}$, subject to $\mathbf{\Sigma} = \mathbf{diag}(\sigma_1, \dots, \sigma_q)$, $\sigma_1 > \cdots > \sigma_q > 0$, and $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_q$. The second and third terms are Elastic Net (EN) [68] penalties on the elements of $\mathbf{U}$ and $\mathbf{V}$. The EN penalty balances well-established techniques of variable selection (zeroing out the noise and redundant signal between omic features) and shrinkage (to account for the high number of omic features that often exceed the number of samples). The expression $||.||_1$ corresponds to the L1 norm. Here $\lambda_. = q(\nu_.)$, is considered as a monotonically decreasing function of $\nu_.$ (the number of desired elements different from zero) onto positive real numbers $\lambda_.$, and $\alpha_.$ is any number between zero and one. The value of $\alpha_.$ balances shrinking and variable selection.

When sparsity is not imposed (i.e., $\lambda_. = 0$), solutions for [Eq.2-1] are obtained by taking partial derivatives on $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$, and setting them to zero. Considering $\mathbf{B} = \mathbf{Z}^T\mathbf{Y}$, solutions for $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$ can be obtained from the partial singular value decomposition of $\mathbf{B}$ of rank q ($\{\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}, \widetilde{\mathbf{\Sigma}}\} = SVD(\mathbf{B}, q)$). When $\lambda_. > 0$, solutions are obtained iteratively from the following set of equations:

$$\begin{cases} \mathbf{U}^* = \frac{\widetilde{\mathbf{U}} - \frac{1}{2}\lambda_U\alpha_U\mathbf{sign}(\mathbf{U}^*))}{1+\lambda_U(1-\alpha_U)} \\ \mathbf{V}^* = \frac{\widetilde{\mathbf{V}} - \frac{1}{2}\lambda_V\alpha_V\mathbf{sign}(\mathbf{V}^*))}{1+\lambda_V(1-\alpha_V)} \end{cases} \qquad \textbf{[Eq.2-2]}$$

Equations in [Eq.2-2] are solved using the algorithm in [69] extended to include the EN parameter $\alpha_.$, where $\mathbf{U}^* = \mathbf{U}\mathbf{\Sigma}$ and $\mathbf{V}^* = \mathbf{\Sigma}\mathbf{V}$. To ensure that $\mathbf{U}$ and $\mathbf{V}$ are orthonormal in each iteration, the QR decomposition of $\mathbf{U}^*$ and $\mathbf{V}^*$ ($\mathbf{U}^* = \mathbf{L}^{(U)}\mathbf{R}^{(U)}$ and $\mathbf{V}^* = \mathbf{L}^{(V)}\mathbf{R}^{(V)}$) is used. In each iteration, $\mathbf{U}$ is recovered as $\mathbf{U} = \mathbf{L}^{(U)}\mathbf{D}^{(U)}$, where $\mathbf{D}^{(U)} = \mathbf{diag}(d_1, d_2, \ldots, d_q)$, and $d_k = 1/||\mathbf{L}_k^{(U)}||_2$ (the inverse of the norm of each column of $\mathbf{L}^{(U)}$). The same steps are used to recover $\mathbf{V}$. After a fixed number of iterations, or at convergence, the final values of $\mathbf{U}$ and $\mathbf{V}$ are used to recover $\mathbf{\Sigma}$ ($\mathbf{\Sigma} = \mathbf{U}^T\mathbf{B}\mathbf{V}^T$).

Tuning hyperparameters: The value $\lambda_.$ is tuned following [69], modified to tune λ_U, λ_V, or both. Briefly, $\lambda_.$ is chosen as the $\nu_.$-th order statistic of $\mathbf{U}$, or $\mathbf{V}$, where ν_U and ν_V are fixed numbers representing the desired number of samples and features loadings different from zero (i.e., the degrees of sparsity), respectively. Then, the proportion of variance explained (PEV) by each one of a grid of values of ν_U and ν_V is calculated. The trajectory of PEV across values of ν_U and ν_V is then used to select an "optimal" $\lambda_.$ value to solve [Eq.2-2]. This selection is made automatically via two alternative methods. The first method uses the first empirical partial derivative of PEV $\left(\frac{\partial PEV}{\partial \nu_.}\right)$ to choose the value of $\nu_.$ at which the change in PEV is maximum ("liberal" method). The second method choses the value of $\nu_.$ at which the change in PEV stabilizes ("conservative" method). Similarly, *MOSS* displays a classic plot of $\sigma_1, \ldots, \sigma_q$ (Scree plot), to visualize the change in variance explained by each latent dimension. The number of latent dimensions q is not tuned by *MOSS* internally. However, automatic suggestions are provided based on the above tuning methods, where the trajectory of $\sigma_1, \ldots, \sigma_q$ is used instead of the PEV one.

Cluster analysis: *MOSS* can use the columns $\mathbf{Q}$ to detect clusters of samples via Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [70]. DBSCAN is one of

the most potent clustering techniques to delimit clusters of irregular shapes. Essentially, DBSCAN identifies groups of densely packed points without specifying the number of clusters a priori. In *MOSS*, neighborhoods of nearby points can then be tuned by evaluating different cluster partitions over a grid of values of *e*, a hyperparameter controlling the neighborhood. *MOSS* chooses the number of "optimal" clusters that maximizes the Silhouette score [71] over a grid of possible *e* values, as in [72].

Visualization of clusters: Additionally, *MOSS* can use *t* - Stochastic Neighbor Embedding (tSNE) to project a group of columns of $\mathbf{Q}$ onto a two-dimensional display [72]. Essentially, tSNE projects multiple dimensions onto a lower-dimensional display while conserving local neighborhoods (eventually representing data clusters) [73]. tSNE is an effective technique to reveal clusters [74]. The algorithm has two fundamental parameters: *perplexity* (which accounts for the adequate number of local neighbors) and *cost* (related to the difference between the neighborhood's distribution in the higher and lower dimensional spaces). Since low costs are more likely to reveal clusters, *MOSS* tunes the tSNE projection by choosing the map of minimum cost among multiple random starts of the algorithm.

2.3.2. Syntaxis of the main function moss

The package's primary function is called ***moss***. This function works along with other auxiliary functions to integrate data sets, pre-process, integrate them, and generate plots. The details of each function can be obtained from the package help pages. Following, we describe the inputs and outputs of *moss*.

Input: The input data must be passed through *data.blocks* as a list of omic blocks. Each row must represent a subject or sample, and each column an omic feature or variable.

MOSS allows each block to be of the class matrix, array, or filed-backed big matrix (FBM) [75]. Objects of class FBM can be passed to *moss* whenever data sets are too big to be handled in RAM. For this, the only requirement is for package *bigstatsr* [75] to be installed. Alternatively, omic blocks passed to *moss* as 'matrix' or 'array' can be internally turned into FBM by setting the argument *use.fbm* = TRUE. In our experience, this can speed up computations when matrices are small enough to fit in memory but still too large to be handled in a reasonable time. If covariates adjust omic blocks, these can be passed as a matrix, vector, or data frame, through argument *covs*.

Standardization, normalization, and imputation: Arguments *scale.arg* and *norm.arg* control, respectively, the standardization of each column within an omic block (i.e., centering to zero mean and scaling to unit variance), and normalization. Omics within *data.blocks* are expected to have named rows. A warning message is displayed if at least one omic is missing row names, or the row names are inconsistent across blocks. In the presence of missing data, a simple imputation by the mean of each column is provided. Nevertheless, before calling *moss*, the user is recommended to run standard quality controls (such as calculating the proportion of missing data across rows and columns, zero variance features, and minor allele frequency).

Methods: Without additional constraints, *MOSS* treats [Eq.2-1] as a partial least squares model (PLS, *method="pls"*). To specify which omic will be used as responses $\mathbf{Y}$, a number from 1 to *t* (the number of omic blocks) must be passed to *resp.block*. By imposing additional constraints, more multivariate techniques can be performed. For example, when $\mathbf{Y}$ is assumed to be the identity matrix, *MOSS* treats [Eq.1-1] as a principal components analysis (PCA) (*method="pca"*). Alternatively, if $\mathbf{Y}$ is a column matrix with

values representing different categories, [Eq.2-1] is treated as a linear discriminant analysis (LDA; [76]) (*method="lda"*).

Cluster analysis and visualization: If argument *cluster*=TRUE, package *dbscan* [77] is used to find clusters on **Q**. By default, this is done on all **Q** columns. However, a different set of columns can be passed as a vector of indexes through argument *axes.pos*. By default, the number of clusters is tuned on a grid of 100 values of *e (eps_res)*, evenly spaced between 0 and 4 (*eps_range*). With this setting, clusters with less than two samples are discarded. Same options are obtained by setting *cluster=list(eps_res=100, eps_range=c(0,4), min_clus_size=2).* This option allows for alternative values of *eps_res*, *eps_range*, and *min_clus_size*.

To obtain the two-dimensional embedding of **Q** (or a subset of columns indicated by *axes.pos*), users can do *tSNE=TRUE*, or *tSNE=list(perp=50, n.iter=1e3, n.samples=1)*, where *perp*, *n.iter*, and *n.samples* are the perplexity parameter of tSNE, number of iterations, and number of random initial conditions, respectively.

Sparsity constraints: Within *MOSS*, $\nu_\cdot$ is specified by arguments *nu.u* (vector of integers between one and the total number of samples) and *nu.v* (vector of integers between one and the total number of features). If the values of *nu.u* and *nu.v* is not specified, only a standard (i.e., dense) SVD is computed. The values of $\alpha_\cdot$ are specified through arguments *alpha.u* and *alpha.v*. Argument *exact.dg* tells *moss* to chose $f: \nu_\cdot \rightarrow \lambda_\cdot$, such as the number of elements different from zero in each column of **U** and **V** is exactly $\nu_\cdot$. Argument *lib.thresh=TRUE* (default) tells *moss* to select the value of $\nu_\cdot$ at which the change in PEV is maximum. If *lib.thresh=FALSE*, the value of $\nu_\cdot$ at which PEV reaches a plateau is chosen.

Parallel computing: By default, the process of tuning the degree of sparsity is done in series. However, this can be changed by setting argument *nu.parallel=TRUE*. This option uses package *future.apply*, to allow for simple parallel distribution of tasks on a local machine or computer cluster [78].

Plots: If argument *plot=TRUE*, several high-level plots will be produced (see Results). This argument requires package *ggplot2* [79] to be installed.

Outputs: Function *moss* returns a named list with the results of the data integration plus additional analyses. The list includes matrices **B** and **Q**, along with two lists containing the results of the dense and sparse SVDs. The output list also has the plots with embedding, cluster analysis, selected items by omic, and signatures of features by a cluster of samples.

Data: Analytical performance was evaluated on data generated with the R package *MOSim* [80]. *MOSim* uses existing omic data to sample pairs of genes and regulators from which differential expression is simulated for a given experimental design. We used *MOSim's* accompanying mouse omic data from the STATegra project [81] to seed all omics. In all scenarios, three omics representing gene expression (RNA-seq count data), micro-RNA seq (miRNA-seq), and ATAC-seq data of DNAase I activity (DNase-seq) were simulated. Signal effects were imposed by assigning different proportions of miRNA-seq and DNase-seq features (5% and 20% of total features) to regulate the expression of 15% of total genes across three clusters of samples. A first simulation scenario with a small number of samples (100) and features (1,000) illustrated *MOSS* main capabilities. Then, simulations for an increasing number of samples and features were used to evaluate *MOSS'* performance at recovering signals (i.e., groups of differentially

expressed genes and corresponding regulatory features) in more realistic scenarios. Due to a restriction of *MOSim* to simulate massive data sizes (hundreds of thousands of rows and millions of features), we benchmarked *MOSS* on representing rank-one bi-clusters, embedded in three synthetic omic blocks with Gaussian noise, and for increasing numbers of samples and features (see the help page for function *simulate_data()*).

2.3.3. Performance evaluation

We evaluated the performance of *MOSS* in terms of its ability to detect differentially expressed genes and their regulatory features. To do that, we used *MOSim* with ten different random starts. We simulated the three omics described above for all combinations of 100, 1,000, and 10,000 samples across three clusters for 1,000 and 10,000 features in each random start. We used 5% and 20% of miRNA-seq and DNAse-seq as regulatory elements for all data sizes. In all scenarios, 15% of total genes were set as differentially expressed (DE). In each scenario, we calculated true positives (TP) as the number of signal features (DE genes and regulatory features) detected by *moss*, true negatives (TN), as noisy features (not DE genes or regulatory elements) not detected by *moss*, false positives (TP) as noisy features detected by *moss*, and false negatives, as signal features not detected by *moss*. These quantities were used to calculate the accuracy $\left(\frac{TP + TN}{TP + TN + FP + FN}\right)$, sensitivity $\left(\frac{TP}{TP + FN}\right)$, specificity $\left(\frac{TN}{TN + FP}\right)$, and precision $\left(\frac{TP}{TP + FP}\right)$.

To evaluate *MOSS* computational time in the context of other methods of omic integration, we used *iCluster* [33], *NMF* [82], *SNFtool* [43], *mixOmics* [58], and *OmicsPLS* [83] R packages to run in the same scenarios. The scenarios consisted of simulations for different combinations of 100, 1,000, 10,000 and 100,000 samples, and 1,000, 10,000,

100,000, and 1,000,000 features. We allowed all methods to have enough available memory (100 Gb) to produce results in hours rather than days.

2.4. RESULTS

2.4.1. Example of unsupervised omic integration with MOSS

The following example shows how to perform omic integration with *MOSS* on a simulated data set with three omics and three clusters of samples. The method was a sparse PCA assuming ten latent dimensions. EN was used to tune the degree of sparsity of features. tSNE was used to embed the first three columns of **Q** onto two dimensions. DBSCAN was used to delimit clusters.

Figure 2.1-A shows the PCA's scree plot (*out_moss$scree_plot*). A clear jump in the singular values occurs between the first and the second dimension. Nevertheless, the scree plot trajectory reaches a plateau after the fourth dimension. Figure 2.2-A shows the trajectory for PEV and derivatives for varying degrees of sparsity for features. (*out_moss$tun_dgSpar_plot*). A degree of sparsity of 300 is suggested according to the conservative tuning method.

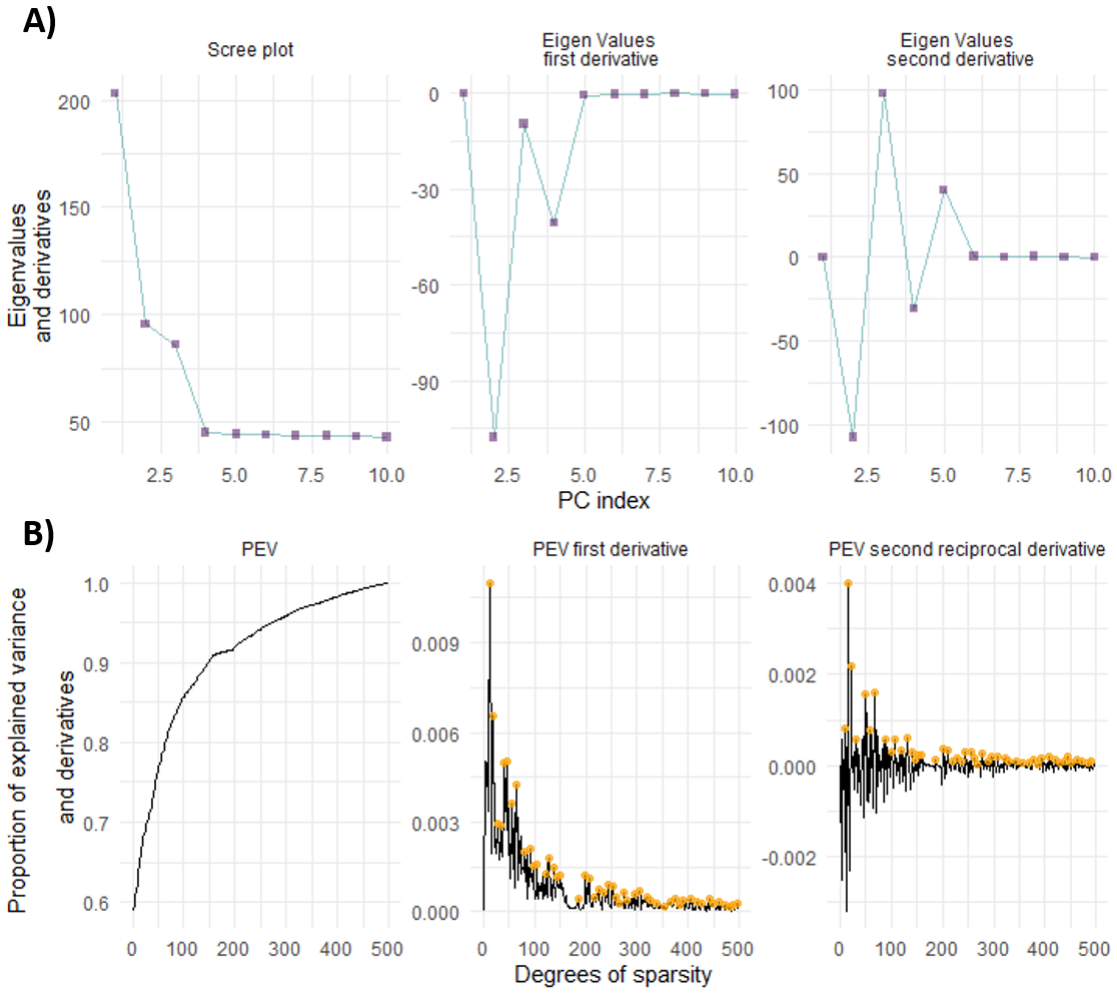

Figure 2.1: Scree plot, PEV trajectory, and partial derivatives. A) The left-most panel shows the singular values corresponding to the first 10 SVD dimensions (scree plot). The following two panels correspond to the first and second empirical partial derivatives of the scree plot. B) Similarly, the left-most panel shows the PEV trajectory on a grid of degrees of sparsity, with the center and right-most panels representing its first and second empirical partial derivatives, respectively.

Figure 2.2 shows the results of the feature selection across omics. The first two latent dimensions had the largest number of features selected. Most of these features were DNA-seq, followed by gene expression.

Conversely, the features selected in the third dimension were mainly miRNA-seq, followed by DNA-seq features. The highest absolute loading values were obtained for the first and third latent dimensions (*out_moss$selected_items*).

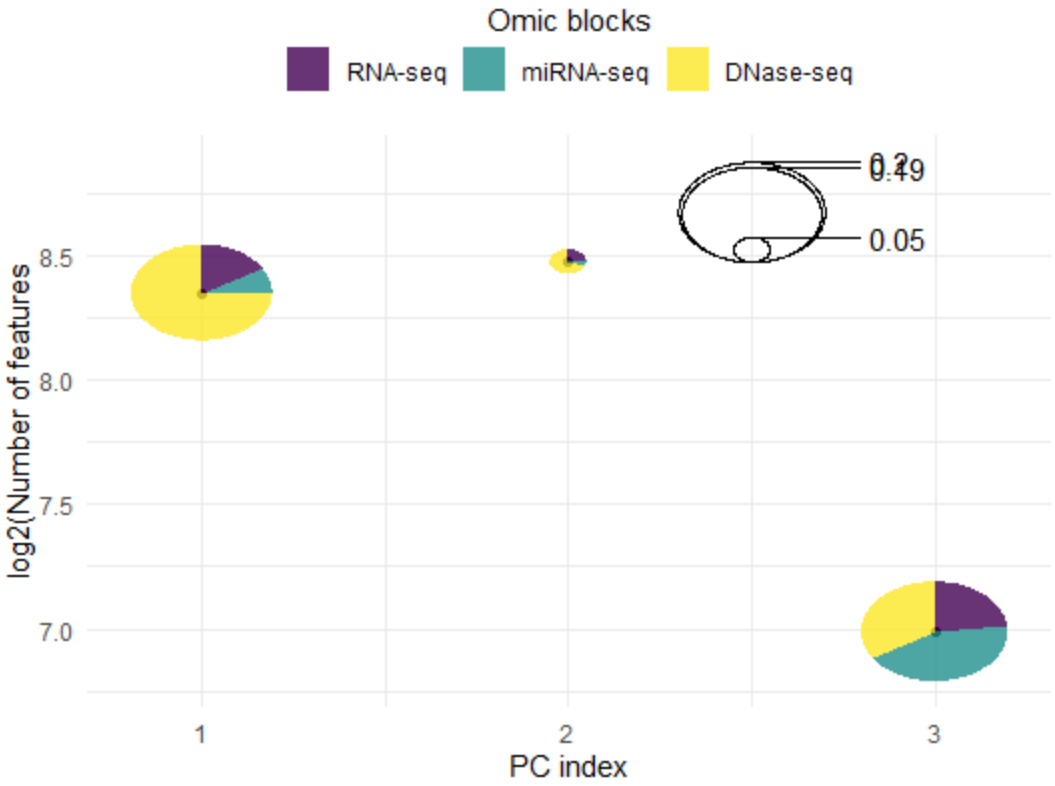

Figure 2.2: Omic contribution to selected features. The numbers of features selected by latent factors are shown in a log2 scale. The pie charts slides represent the relative contribution of each omic (RNA-seq, miRNA-seq, and DNase-seq) to the features selected by dimension (PC index). The pie charts ratio represents the quotient between the squared loadings of the selected features by dimension and their standard deviation.

The tSNE map and cluster analysis are presented in Figure 2.3 (*out_moss$clus_plot*).

The three simulated clusters are detected. From Figure 2.4, we can see that the first and

third dimensions are the more relevant for cluster formation.

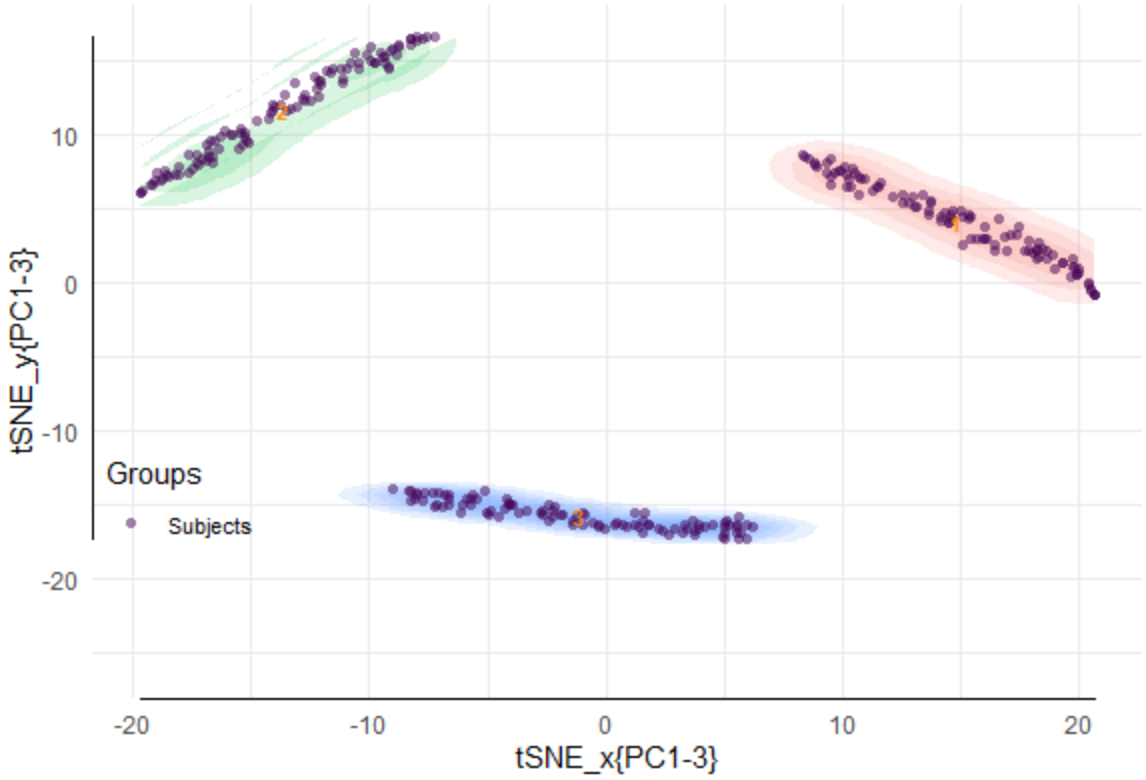

Figure 2.3: Cluster analysis. The three first SVD latent factors of the integrated omics were embedded onto two dimensions via tSNE. Clusters (labeled **1**, **2,** and **3**) were delimited via DBSCAN.

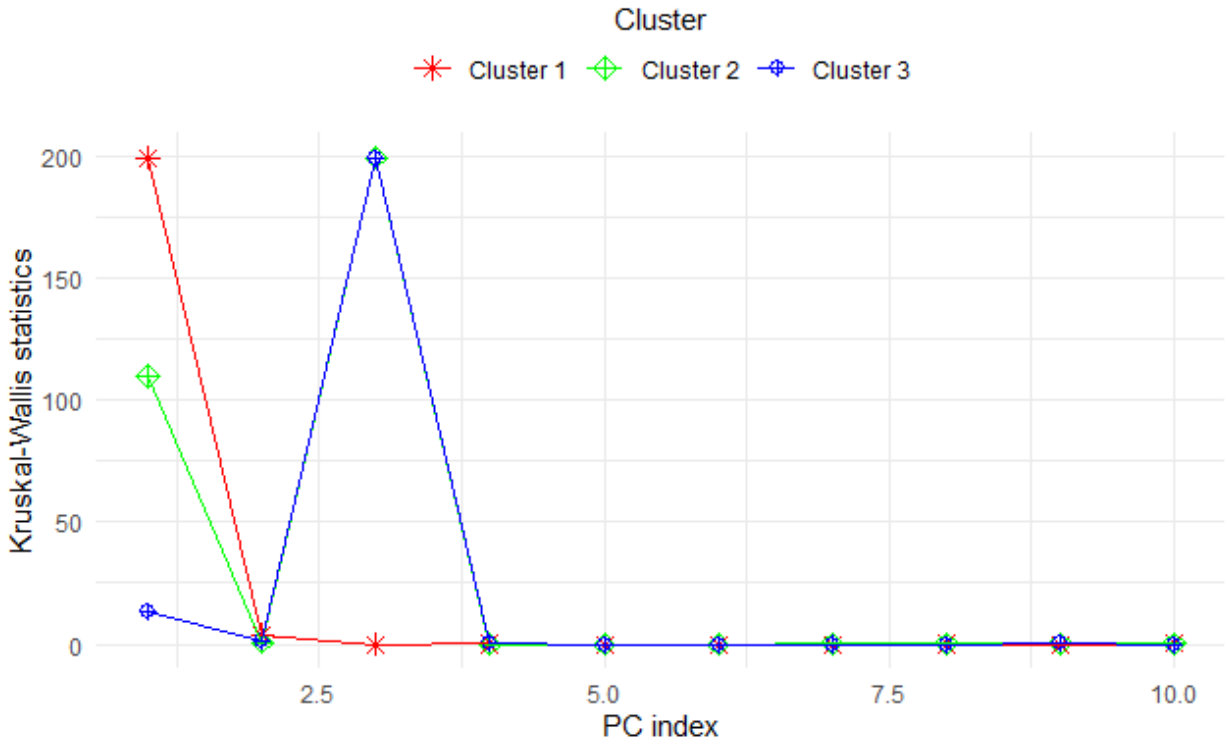

Figure 2.4: Association between clusters and SVD dimensions. The plot shows the results of Kruskal-Wallis association tests between clusters and the first ten SVD dimensions (PC index). A different point shape and color represent each cluster.

Lastly, to determine what selected features dominated each cluster, the user can look at signatures within *out_moss$feat_signatures*. If the number of features is too large to visualize correctly, the function *moss_signatures* can be called (Table A.1).

Figure 2.5 shows the top 5% of candidate features. The percentage is based on the squared means of features values within the respective omic. Candidates are defined as features with standard error intervals excluding zero.

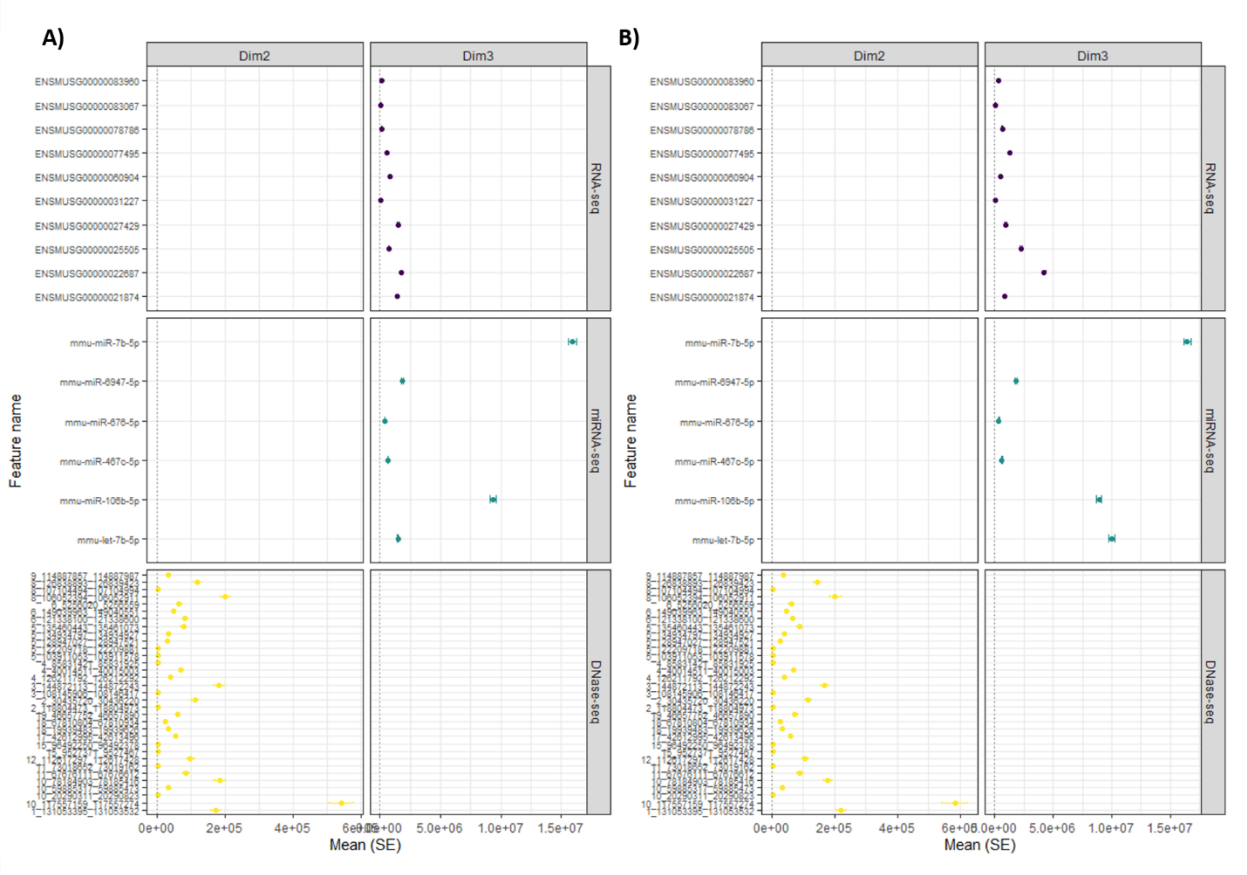

Figure 2.5: Signature of features for two clusters. The plot shows the top features (y-axis) selected by latent factor and omic block. Points correspond to the average feature values plus and minus one standard error. For clarity, only the top 1% of selected features and the first two clusters are shown.

2.4.2. Example of supervised omic integration with MOSS

Setting *method = "pls"* and *resp.block = 1* tells *moss* to run a PLS using gene expression as a multivariate response and the remaining omic blocks as multivariate predictors. The following example shows how to run a sparse PLS, where the degree of sparsity is separately tuned for both responses and predictors (code provided in Table A.1).

The trajectory of PEV and derivatives for both responses and predictors are displayed in Figure 2.6 (*out_moss$tun_dgSpar_plot*). In this example, we have set *lib.thresh = TRUE* to use the liberal method of features selection. Figure 2.6 suggests that the liberal method would select less than ten genes and approximately 50 regulatory features. As before,

the contribution of each omic to the selected items, clusters, and features signatures can be obtained from *out_moss*. In addition, the accompanying function moss_heatmap can retrieve a heatmap of the covariance matrix between selected responses, and predictors can be retrieved using the accompanying function *moss_heatmap*. This function uses the *ComplexHeatmaps* R package [84] (Table A.1).

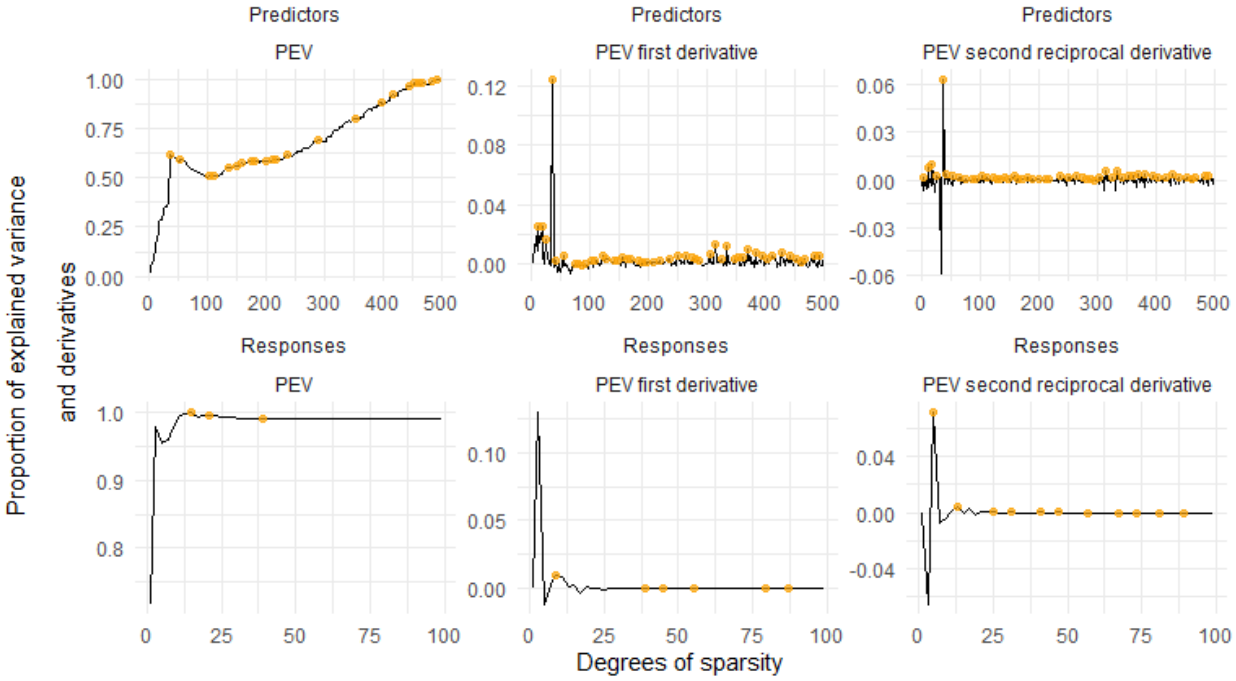

Figure 2.6: PEV trajectory plot and partial derivatives for a PLS analysis. The top three panels represent the PEV trajectory on a grid of degrees of sparsity for predictor features and its first and second empirical partial derivatives, respectively. The bottom three panels show this information for responses. Here, responses and predictors were gene expression and regulatory omics, respectively.

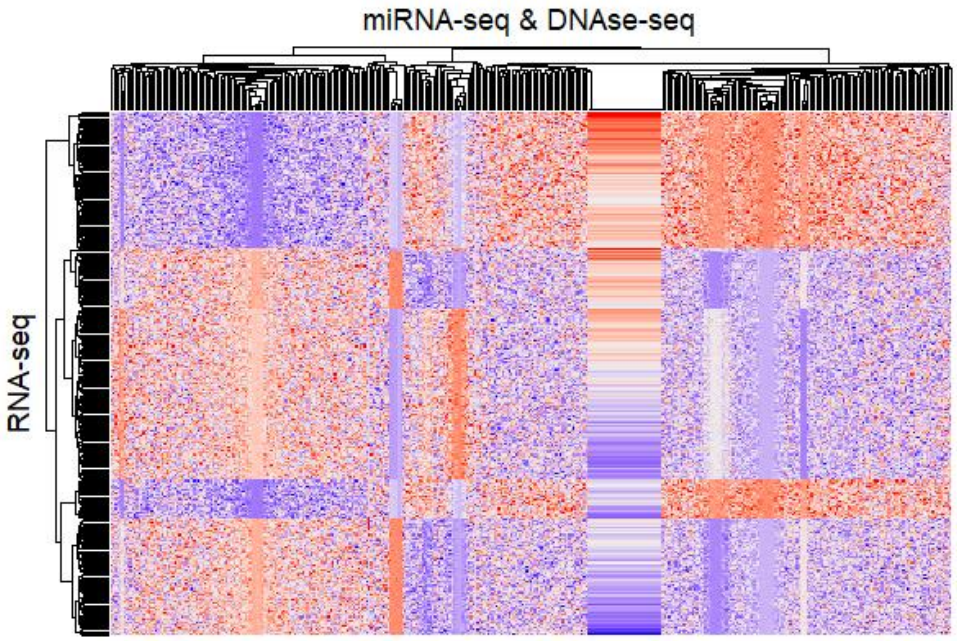

Figure 2.7: Output of the function *moss_heatmap.* Rows and columns represent genes and regulatory features selected by a given combination of latent factors. In the example, the first and third dimensions of the sparse SVD of the covariance between gene expression and remaining omics were used. Features names are omitted for clarity.

2.4.3. Evaluation of MOSS analytical performance

Results of performance evaluation are shown in Figure 2.8. The accuracy of *MOSS* to detect signal features was in the order of ~0.9 for all scenarios, except for the largest number of features and lowest signal intensity. Sensitivity was high for all scenarios of high signal intensity, dropping to ~0.6 for more than 100 features. When signal intensity was low, sensitivity dropped in all scenarios, especially for the largest number of features. However, increasing sample sizes corresponding with higher sensitivities. Specificity was highest for scenarios of low signal intensity and moderated to a high number of features. Higher signal intensities corresponded to lower specificities, but this difference was almost negligible for the number of features. Precision was high in all scenarios. However, it dropped for the lowest signal intensity values and the number of features.

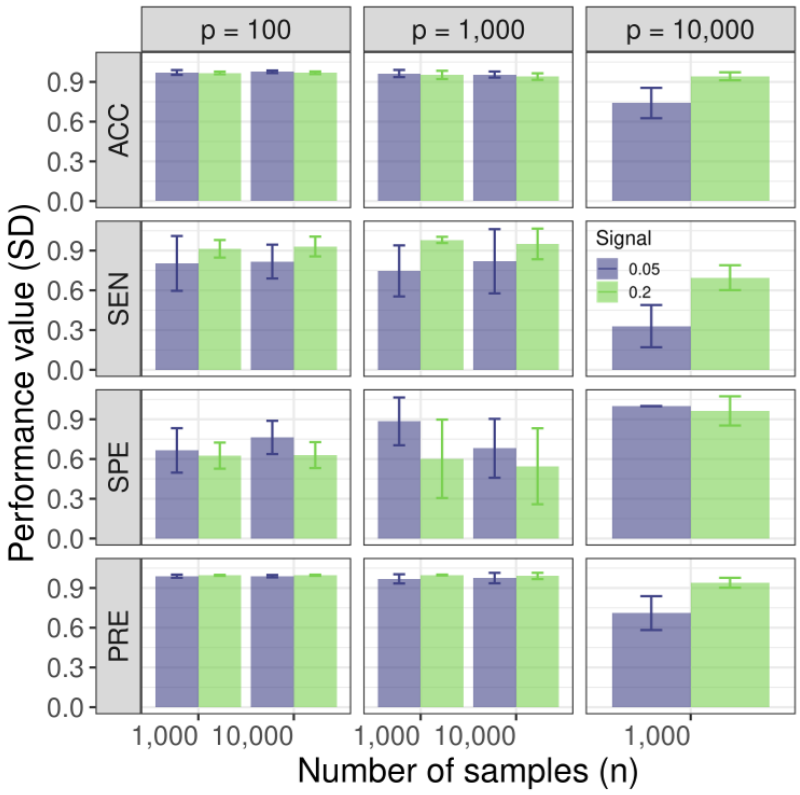

Figure 2.8: Performance of feature detection with MOSS. Performance at detecting features was evaluated on simulated data with different combinations of numbers of samples (**n**) and features (**p**) for two alternative signal intensities (**0.05** and **0.2**), defined as the proportion of features with regulatory effect. The performance metrics were accuracy (**ACC**), sensitivity (**SEN**), specificity (**SPE**), and precision (**PRE**). Rectangles and bars represent metric values averages and average plus-minus standard errors across the simulations' random starts, respectively. Adjacent columns by **n** tick mark represent alternative signal intensities.

2.4.4. Benchmarking

Figure 2.9 shows the computational time taken by *MOSS* and other omic integration methods for different combinations of samples (*n*) and the number of features (*p*). In all cases, three omic blocks were simulated. We allowed all methods to have enough available memory (100 Gb) to produce results in hours rather than days. Nevertheless, for scenarios approaching bio-bank dimensions (i.e., hundreds of thousands of samples and omic features), all methods either crashed or could not finish before 24 hours. The exception to this pattern was the use of moss *MOSS* with *use_fbm* =TRUE. For a

relatively low number of samples (e.g., n <$1x10^3$), all methods, except *iCluster*, *NMF*, and *mixOmics*, produced results in less than one hour. In *iCluster*, p =$1x10^5$ causes the program to run for more than a day, and with p=$1x10^6$, to crash. *NMF* and *mixOmics* took approximately 3 and 10 hours, respectively, when n=1,000 and p=1e6. Finally, the method with the most substantial influence of sample size was *SNFtool*, with better performances for smaller sample sizes, with the opposite being true for larger sample sizes.

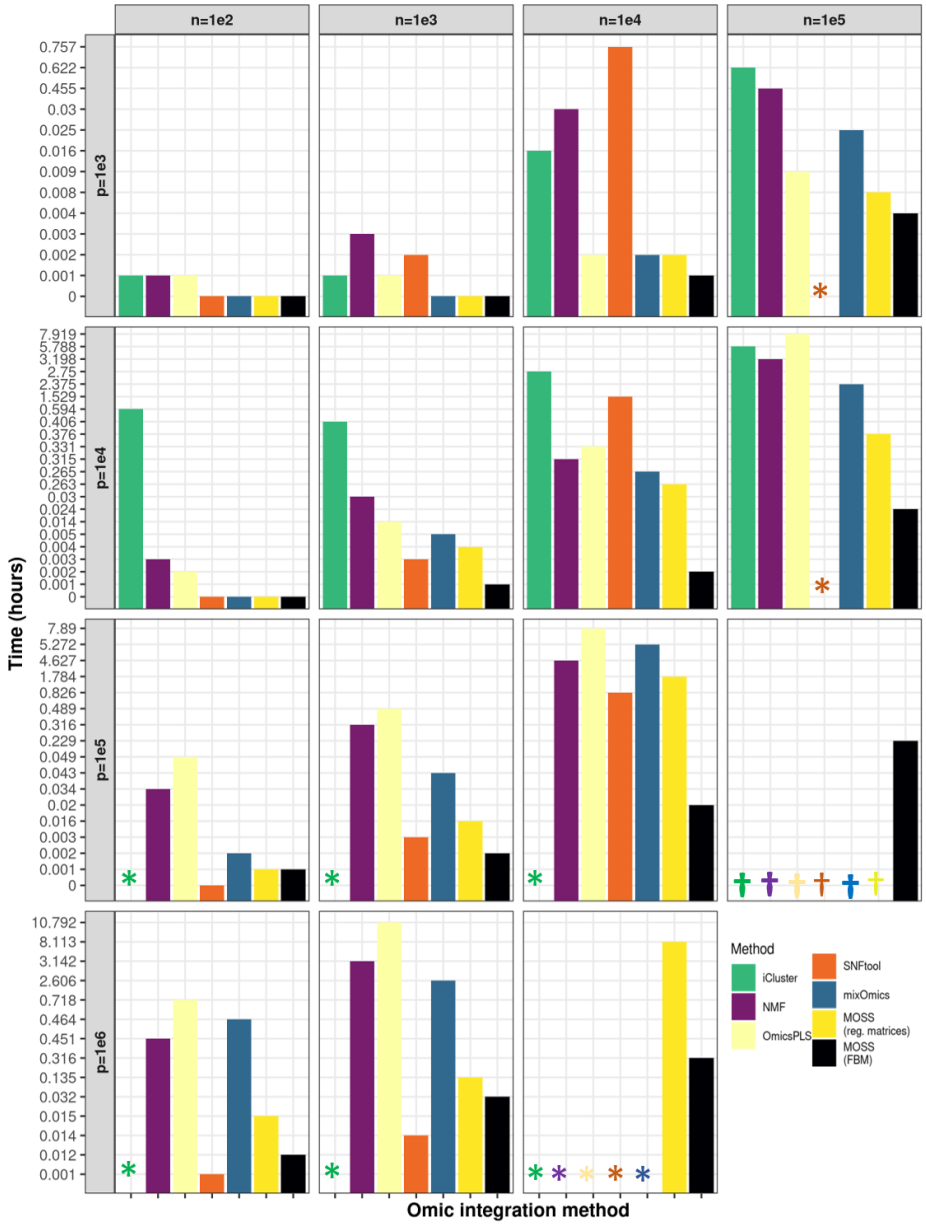

Figure 2.9:Benchmarking of omic integration methods. The plot shows the computation time taken by *MOSS* and five other omic integration methods (*iCluster*, *NMF, SNFtool*, *mixOmics*, and *OmicsPLS*) to integrate three omic blocks and perform feature selection in different simulated scenarios. Scenarios corresponded to a different combination of numbers of samples and features in simulated data. Column panels (**n**) represent the number of samples, and row panels (**p**) represent the number of features. Each rectangle corresponds to a different omic integration method. The rectangle's height represents computing time in hours. The symbols "*****" and "**†**" represent a method running for more than a day or crashing, respectively. MOSS was used with regular matrices or filed-backed big matrices (FBM).

2.5. DISCUSSION

Omic integration emerged as a group of techniques for analyzing multiple omic data layers collectively and retrieving helpful information of shared processes within the cell [63]. However, the computational and statistical tools used to carry out these tasks are constantly challenged by the vast amount of generated data [29,85]. As an essential step in understanding the biology of complex traits, omic integration methods should retrieve informative results in a reasonable amount of time. For that purpose, we have developed *MOSS*, a free and user-friendly tool that rapidly retrieves information about the principal axes of variation across omic data, identifies features of possible biological roles, detects clusters of individuals, and represents them in terms of features of potentially functional role.

We evaluated *MOSS* in terms of several metrics representing the performance of feature selection. In terms of both accuracy (i.e., ability to collectively detect signal and miss noise) and precision (ability to detect true from the false signal), *MOSS* best performance occurred in scenarios where the proportion of the number of features (*p*) to the number of samples (*n*) was low to moderate. The lower accuracy and precision in the scenarios with large *p/n* and low proportion of regulatory features could have been a consequence of the heuristic used to select features. Like related least squares-based algorithms, this performance could improve if the value of *n* is increased [86]. However, lower performance in high *p/n and* low signal-high noise scenarios is an unsolved challenge among omic integration methods [56]. Knowing if MOSS's decrease in performance for this setting is smaller than for competitive methods would require further simulation studies.

The evaluation of the analytical performance of competitive methods was out of the scope of this paper and can be found elsewhere (e.g., [56,65]). Nevertheless, to support the standalone performance of MOSS, we have included the results of a small simulation in Figure A.1.

Since the proportion of differentially expressed genes remained the same across scenarios, increasing *p/n* could also increase the number of false positives. Hence, lower signals could imply that some regulatory features are missed, reducing true positives. Similarly, increasing false positives can explain the drop in sensitivity (i.e., the ability to detect signal correctly) in the high *p/n* scenario. Likewise, specificity (i.e., ability to correctly miss noisy features) increased with fewer regulatory features, increasing true negatives, regardless of the higher false positives expected in the high *p/n* scenario.

A thorough evaluation of performance across alternative tuning methods for feature selection is outside the scope of this paper. Nevertheless, we acknowledge that performance on the feature selection approach, and an alternative tuning method could improve *MOSS* performance. However, heuristics based on training sets can improve computational times compared to traditional cross-validation [87].

We have shown the ability of *MOSS* to retrieve biologically meaningful results in different simulated scenarios, ranging from a few numbers of samples and features to data volumes approaching biobank scales. Although trying to create synthetic but realistic data via the *MOSim* package, as in any other simulation approach, we acknowledge that only a finite combination of scenarios was explored. Nevertheless, in an earlier publication, we have shown that *MOSS* can also retrieve biologically meaningful results on real data. For example, in [88], we have used *MOSS* to effectively integrate information from ~60,000

features from gene expression, DNA methylation, and copy numbers across ~5,000 tumors from different cancer diagnoses. In that work, we showed *MOSS*'s ability to detect clusters of tumors beyond original diagnoses, which shared molecular features of potential therapeutic use.

One of *MOSS's* essential capabilities is the handling of data sizes reaching biobank dimensions. However, even when regular R matrices are used, *MOSS* can perform in a short amount of time compared to other methods of omic integration and feature selection. In addition, package *bigstatsr* allows *MOSS* to perform a dense-partial SVD in data sets as big as the UK Biobank [75]. In addition, MOSS includes a convenient parallel computing scheme, as provided by the *future.apply* R package. Although this implies that the user loses some control on how parallel jobs are administered, since *future.apply* works in multiple platforms, this option reduces the guesswork and the dependency of the parallel computing strategy on operating systems used.

MOSS is a flexible, fast, and robust tool to perform data integration. It shares capabilities with popular methods, including estimation of latent data dimensions, feature selection, and convenient graphical displays. Nevertheless, unlike these methods, *MOSS* integrates datasets too large to handle in RAM, requiring shorter amounts of time.

CHAPTER 3

MULTI-OMIC SIGNATURES IDENTIFY PAN-CANCER CLASSES OF TUMORS BEYOND TISSUE OF ORIGIN

This chapter has been adapted from the article published in the open-access journal, *Scientific Reports* (DOI: 10.1038/s41598-020-65119-5).

3.1. ABSTRACT

Despite recent advances in treatment, cancer continues to be one of the most lethal human maladies. One of the challenges of cancer treatment is the diversity among similar tumors that exhibit different clinical outcomes. Most of this variability comes from widespread molecular alterations that can be summarized by omic integration. We have identified eight novel tumor groups (C1-8) via omic integration, characterized by unique cancer signatures and clinical characteristics. C3 had the best clinical outcomes, while C2 and C5 had the poorest outcomes. C1, C7, and C8 were upregulated for cellular and mitochondrial translation and low proliferation. C6 and C4 were also downregulated for cellular and mitochondrial translation and had high proliferation rates. C4 was represented by copy losses on chromosome 6 and had the highest number of metastatic samples. Copy losses on chromosome 11 characterized C8, also having the lowest lymphocytic infiltration rate. C6 had the lowest natural killer infiltration rate and was represented by copy gains of genes in chromosome 11. C7 was represented by copy gains on chromosome 6 and had the highest upregulation in mitochondrial translation.

We believe that, since molecularly alike tumors could respond similarly to treatment, our results could inform therapeutic action.

3.2. INTRODUCTION

Despite recent advances that have improved cancer treatment, it reigns as one of the most lethal human diseases. Cancer can be considered a highly heterogeneous set of diseases: while some tumors may have a good prognosis and are treatable, others are quite aggressive, lethal, or may not have a standard of care [89–91]. Cancer can also defy standard classification: a well-classified tumor may not respond to standard therapy, as expected, and may behave as a different cancer type [92–94]. Fortunately, with the advances of sequencing technologies, data has become available for research as never before. The Cancer Genome Atlas (TCGA), for instance, offers clinical and omic (e.g., genomic, transcriptomic, and epigenomic data) information from thousands of tumors across 33 different cancer types [95]. Much of this omic data can enable us to classify tumors and explain the striking variation observed in clinical phenotypes [96–99].

Omic integration has been successfully applied in previous classification efforts [72,100–102]. These classifications have highlighted how molecular groups of tumors highly agree with human cell types. Alternatively, we hypothesize internal subtypes hidden by cell type and tissue characteristics influencing cell behavior. These subtypes could be distinguished by molecular alterations unlocking cancerous cell-transformation events. To test this hypothesis, we have developed a statistical framework that summarizes omic patterns in main axes of variation, describing the molecular variability among tumors. Key features characterizing each axis (i.e., features contributing the most to inter-tumor variability) are retained, while irrelevant ones are filtered. Retained features are then used to cluster tumors by molecular similarities and find specific molecular features representing each group.

Here we show that, after removing all tissue-specific effects, the cancer signal immediately emerges. The new molecular aggrupation, emphasizing shared tumor biology, can supply new insights into cancer phenotypes. We expect this novel classification to aid in developing therapeutic alternatives for tumors without a current standard of care.

3.3. MATERIAL AND METHODS

3.3.1. Pan-cancer data

The TCGA offers a demographically diverse sample with comprehensive and modern multi-omic data. We retrieved data from 5,408 from 33 cancer types made available by the Genome Data Commons (GDC) repository [103] via the TCG-Assembler R package [104]. Omic data consisted of curated level-three data of genome-wide gene expression (GE), DNA methylation (METH), and copy number variants (CNV) profiles by tumor sample. GE profiles by sample corresponded with the logarithm of RNA-Seq counts by gene (Illumina HiSeq RNA V2 platform). METH profiles corresponded with CpG sites B-values from the Illumina HM450 platform, summarized at the CpG island level, using the maximum connectivity approach from the *WGCNA* R package [105], and further transformed into M-values (M=β/(1-β);[106]). CNV profiles corresponded to gene-level copy number intensity derived from Affymetrix SNP Array 6.0 platform, using human genome V19 as reference. The quality-control filtering process included excluding features with all zeros or coefficient of variation less than 1%. Samples or features disproportionally missing data (>20%) and single-sample batches were also excluded. Within the remaining samples, missing values were imputed by k-near neighbors, with k = 3. Finally, each omic block was adjusted by batch effects using *ComBat* [107]. The final sample size after retaining subjects with information for all three omics was n=5,408.

Demographic information included gender, self-reported race and ethnicity, and patient's age at diagnosis (Table 3.1). Clinical information consisted of overall survival time and vital status at the final follow-up, sample type (from the primary tumor, metastases, or normal tissue), tumor-free fraction. We also used previous information from “The Immune

Landscape of Cancer" [108] and calculated significant differences between clusters using the Kruskal-Wallis tests [109]. These covariates included: intra-tumor heterogeneity fraction (as sub-clonal genome fraction), and rates of non-silent mutations, aneuploidy, homologous recombination defects (all three derived as deviations from the normal genome), proliferation (normalized difference between the number of dividing and non-dividing cells), and information from immune infiltrations (including scores for CD4+ cells, macrophages, lymphocytes, and natural killers) (See supplementary material in [108] for a detailed description of the scores' calculation). Briefly, immune infiltration fractions were derived by CIBERSORT [110], assigned to different cell classes, and multiplied by the leukocyte fraction derived from methylation data [108].

3.3.2. Omic integration, clustering, and features selection.

The following four steps can conceptually describe our method.

Step 1) Identification of major axes of variation and features selection. Integrative methods should capture combined effects across omic sites that could either span across omic layers (e.g., epigenetics, gene expression) or extend genome-wide (e.g., considering concomitantly contiguous CpG sites or even separated away sites). Let,

$$\mathbf{X} = [\mathbf{X_1}, \quad \ldots, \quad \mathbf{X_L}]$$

where $\mathbf{X}_l$ l*:{1,…, L}* is a matrix representing the l-th omic, which row i^{th} contains information representing a sample on one subject, and column j^{th} represents an omic feature (e.g., a feature could be the expression of a specific gene or the methylation level for a given CpG site). Each group of features coming from a different omic block is centered, standardized, and divided by $\sqrt{p_l}$, where p_l is the number of features from the l-th omic block. Normalization is done so larger groups of features do not dominate the data

integration step. Next, we conduct a sparse Singular Value Decomposition (**sSVD**) of $\mathbf{X}$ to generate one factor that collapses the redundancies in the omics (by creating independent columns representing the independent signals across features) and one factor that collapses redundancies across samples, grouping subjects with similar signaling. This linear factorization can be represented as $\mathbf{X} = \mathbf{ZW}$, where $\mathbf{Z}$ represents (linearly) independent axes of variability across subjects (i.e., a lower rank approximation), while $\mathbf{W}$ represents loadings representing the contribution of each omic feature to this variability. This representation is familiar to many unsupervised omic integration methods but is independent of distributional assumptions on each element. In this formulation, $\mathbf{Z}$ and $\mathbf{W}$ can be obtained by minimizing:

$$\|\mathbf{X} - \mathbf{ZW}\|_2^2 + P_{\lambda,\alpha}(\mathbf{W}) \qquad \textbf{[Eq.3-1]}$$

To the left of the plus sign is the Frobenius norm (a matrix analogous of Euclidean distance) of the difference between $\mathbf{X}$ and the product of $\mathbf{Z}$ and $\mathbf{W}$. To the right of the plus sign is a penalty on the elements of **W** to impose sparsity. The purpose of this penalty is to zero-out those features with minor contributions to the columns of **Z**. To remove the effect of tissues, or other covariates that can influence the selection of features, we pre-multiplied **X** by $\mathbf{I} - \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T$, where **I** is a diagonal matrix of ones, and **Q** is an indicator matrix to represent the membership to a given organ or tissue.

Step 2) Identify omic features (expression of genes, methylation intensities, copy gains/losses) influencing the axes. The linear decomposition achieved by SVD is an intuitive and straightforward way of integrating omics. However, the variability across omics can be governed by just a few features (i.e., highly *sparse* data) or by groups of interdependent features (i.e., very *redundant* data). To handle these limitations, we chose

$P_{\lambda,\alpha}(\mathbf{W})$ to be the Elastic Net penalty [68], $\lambda(\alpha\|\mathbf{W}\|_1 + (1-\alpha)\|\mathbf{W}\|_2^2)$, where α balances the regularization between LASSO and ridge regression types of regularization, and λ is associated with the degree of sparsity (i.e., how many features enter in the model?). Unlike LASSO, EN can select groups of correlated features, while zeroing out the irrelevant ones [111]. Equation Eq.3-1 is solved by obtaining $\boldsymbol{z}_1\boldsymbol{w}_1$ (where $\boldsymbol{z}_1$ is the first column of **Z** and $\boldsymbol{w}_1$ is the first row of **W**) with coordinate descent for given values of λ and α, following the algorithm of [69], as implemented in [112], but with the following thresholding operator: sign($\boldsymbol{w}_1$)| $|\boldsymbol{w}_1|$ - $\lambda\alpha$ $|_+$ / λ $(1-\alpha)$ (where $|\boldsymbol{x}|_+$ represents the positive part $\boldsymbol{x}$). Consecutive layers are then obtained by subtracting the previous ones from **X** and repeating the same procedure, as many times as the number of desired axes of variation. The optimal value for λ was empirically determined, as suggested by [69]. We start by 1) calculating **W** over a dense grid of values for λ (lower λ yields less sparsity), 2) calculating the proportion of variance of **X** explained by **ZW** (*PVX*) for each λ, and 3) choosing the λ at which *PVX* has its minimum second derivative. Since *PVX* decreases monotonically with λ, this point represents a drastic drop on *PVX*, suggesting that the most relevant features accounting for the data variability are already incorporated [69]. The value α was fixed to 0.5 to have an equal contribution of LASSO and Ridge penalties. Once a subset of features was selected, we mapped them onto genes using annotation data of genomic position downloaded from the USCE web browser tool (GRCh38 [113]). The enrichment of functional classes (ontologies, pathways, complexes) among these genes was tested using the *Enrichr* package[114].

Step 3) Mapping major axes of variation via tSNE and cluster definition by DBSCAN. Additionally, SVD can be coupled with non-linear embedding methods to deal with highly

heterogeneous data. Here, we applied *t* - Stochastic Neighbor Embedding (tSNE) on $\mathbf{Z}$ [72]. tSNE is a technique that efficiently takes on local neighborhoods present in high dimensions (eventually representing clusters of data) and conserves them while projecting onto a lower-dimensional display [73]. Hence, tSNE becomes a powerful technique to reveal clusters, even in very heterogeneous and convoluted data settings [74]. The algorithm has two fundamental parameters: perplexity (which accounts for the effective number of local neighbors) and cost (related to the difference between the neighborhood's distribution in the higher and lower dimensional spaces). Since low cost indicates displays more likely to reveal clusters, we selected the maps corresponding with the lowest costs among perplexities of 50 and 100, using 100 thousand iterations to ensure convergence. We applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN [77]) to identify clusters. DBSCAN is one of the most powerful clustering techniques to delimit clusters of irregular shape, such as the ones tSNE produces [74]. Essentially, DBSCAN identifies densely packed groups without specifying the number of clusters prior [77]. Neighborhoods of nearby points can then be tuned by evaluating different cluster partitions over a grid of possible neighborhood sizes. We tuned this parameter by maximizing the Silhouette score, as in [72].

Step 4) Molecular and clinical characterization of clusters. The association between clusters and scores representing genes and the selected functional classes was studied to define each cluster's signatures. Scores were calculated by tacking the columns of **X** mapping onto a gene, or functional class, and post-multiplying it by the corresponding elements of **W'**. Using the scores of each gene and functional class as a response and the clusters as explanatory variables, we conducted a series of ANOVA tests to determine

what genes or functional classes were significant in at least one cluster. All pairwise comparisons between significant genes and functional classes were studied via Tukey tests. Gene signatures were defined based on those genes significantly deregulated in a single cluster. For both types of tests, we used a Bonferroni multiple-test correction with P(type I –error) = 0.05 / {#selected genes and functional classes}.

We used the STRING database of protein-protein interactions to discuss the possibility of physical or functional relationships between the genes in each signature [115]. We considered an interaction biologically meaningful when backed up by empirical data, such as immune precipitation, microarrays, and curated databases. Interactions suggested by text-mining (two genes reported in the same scientific publication) were not considered, except in the cases when a publication's results gave evidence of interaction (e.g., genes co-expressing, co-locating).

The association between clusters and phenotypes (e.g., clinical, demographic, and immunologic covariates) was evaluated via the Kruskal-Wallis test [109] (non-parametric analogous of ANOVA). The Dunn test further evaluated all significant results [116] for pairwise differences (non-parametric analogous of Tukey tests). All steps of our method were implemented in the R programming language [117], using *irlba* [112], *dbscan* [77], and *Rtsne* [118] packages.

3.4. RESULTS

Signals coming from tissue and cell type strongly influence a naïve initial classification of tumors across cancer types. We performed omic integration based on sparse singular value decomposition, removed tissue effects, and sought to re-classify tumors based on subtler omic patterns. Our method can be illustrated in four steps (Figure 3.1, Materials

and Methods). *Step 1* applies sparse Singular Value Decomposition (sSVD) to an extended omic matrix ***X***, obtained from concatenating a series of scaled and normalized omic blocks for the same subjects. Briefly, the principal axes of variation across tumors (i.e., left principal components or scores) and **X**'s matching features 'activities' (i.e., the right principal components or loadings) are found. Sparsity is then imposed on the activity values, so features with minor influence over the tumors' variability are removed. *Step 2* consists of identifying what features (expression of genes, methylation intensities, copy gains/losses) influence these axes the most (i.e., features not removed by sSVD) and mapping them onto genes and functional classes (e.g., pathways, ontologies, targets of micro-RNA). *Step 3* involves the identification of local clusters of tumors, following [72]. *Step 4* involves characterizing clusters in molecular (e.g., genes, pathways, complexes) and clinical (e.g., survival probability, immune infiltration) information, distinguishing each cluster from the rest.

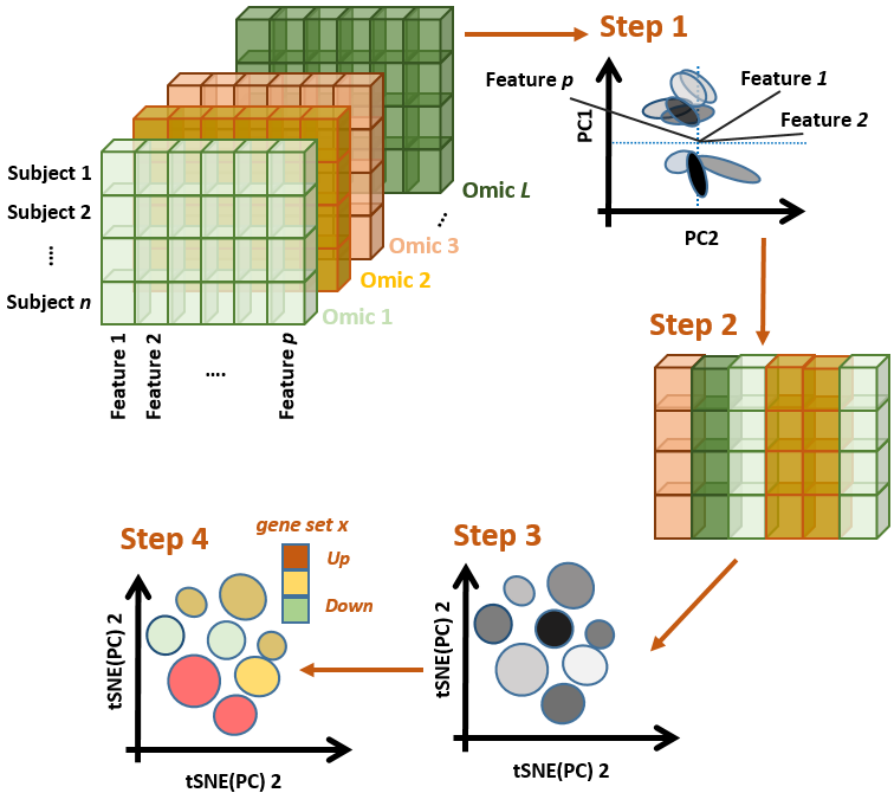

Figure 3.1: Omic integration and features selection method. Step 1) Singular value decomposition of a concatenated list of omic blocks and identification of major axes of variation. Step 2) Identification of omic features (expression of genes, methylation intensities, copy gains/losses) influencing the axes and mapping them onto genes and functional classes (e.g., pathways, ontologies, targets of micro-RNA). Step 3) Mapping major axes of variation via tSNE and cluster definition by DBSCAN. Step 4) Phenotypic characterization of each cluster of subjects.

Using samples from 33 different cancer types provided by The Cancer Genome Atlas (TCGA) and accompanying information from whole-genome profiles of gene expression (GE), DNA methylation (METH), and copy number variant alterations (CNV), we re-classified tumors based on molecular similarities between the three omics.

3.4.1. Data description

The data, including information of sample size and type of sample (i.e., from normal, metastatic, or primary tissue), demographics (age, sex, and ethnicity) and survival

information (overall survival status and times), are summarized in Table 3.1. In addition, omic data included information for gene expression (**GE**, as the standardized log of RNAseq data for 20,319 genes), methylation (**METH**, as standardized M-values summarized at the level of 28,241 CpG islands), and copy number variants (**CNV**, as the standardized log of copy/gain intensity, summarized at the level of 11,552 genes).

Table 3.1: Data description by cancer type after quality controls. Samples are described by cancer type (TCGA code and cancer name), in terms of relative sample size (**n**), percent of females (**F%**), ethnicities (percent of non-Hispanic Whites, Afro-descendants, and Asians), **Age** (at the moment of diagnosis, in years), type of sample (**TS%**, as a percent of normal –**N**- and metastatic –**M**- samples), and survival (**Surv**, as expected time to 50% survival, in years). **Age** and **Surv** are represented by median values, with first and third quartiles as measurements of dispersion.

Code	Type	n	F%	AD	W	A	Age	N	M	Surv
ACC	*Adrenocortical*	23	61	0	100	0	48 (35-57)	0	0	6.6 (2.5-6.6)
BLCA	*Bladder urothelial*	271	99	13	80	7	58 (49-66)	1	0	3.0 (1.2-3.0)
BRCA	*Breast invasive*	639	69	18	75	7	58 (46-71)	7	0	10.2 (6.5-10.2)
CESC	*Cervical squamous cell*	234	25	8	78	14	60 (53-69)	1	1	11.2 (3.1-11.2)
CHOL	*Cholangiocarcinoma*	12	36	0	100	0	55 (46-67)	75	0	1.7 (0.7-5.3)
COAD	*Colon adenocarcinoma*	264	36	12	79	9	58 (41-66)	7	0	8.3 (3.6-8.3)
DLBC	*Lymphoma*	26	54	19	81	0	60 (54-63)	0	0	17.6 (17.6-
ESCA	*Esophageal*	134	60	12	88	0	68 (59-73)	2	0	2.3 (1.1-4.4)
GBM	*Glioblastoma multiforme*	49	23	12	78	10	66 (60-73)	0	0	0.9 (0.4-1.2)
HNSC	*Head and Neck squamous*	89	48	8	91	1	61 (59-71)	1	0	5.9 (1.2-5.9)
KICH	*Kidney chromophobe*	2	0	0	100	0	52 (50-54)	0	0	--**
KIRC	*Kidney renal clear cell*	43	51	2	91	7	67 (62-75)	0	0	7.5 (7.5-7.5)

Table 3.1 (cont'd)

KIRP	*Kidney renal papillary cell*	37	62	20	80	0	65 (59-72)	0	0	--
LAML	*Acute myeloid leukemia*	28	0	0	94	6	60 (57-67)	0	0	--
LGG	*Brain lower grade glioma*	93	42	11	88	1	70 (62-75)	0	0	9.5(3.1-12.2)
LIHC	*Liver hepatocellular*	62	25	8	92	0	69 (61-74)	13	0	4.6 (1.6-8.6)
LUAD	*Lung adenocarcinoma*	381	29	6	90	5	66 (59-72)	4	0	4.2 (2.1-9.2)
LUSC	*Lung squamous cell*	289	28	9	89	2	57 (46-64)	0	0	4.7 (1.8-10.5)
MESO	*Mesotheli-oma*	68	0	7	93	0	60 (53-66)	0	0	1.6 (0.9-2.4)
OV	*Ovarian serous*	5	0	0	100	0	60 (55-61)	0	0	2.9 (2.9-2.9)
PAAD	*Pancreatic adenocarcinoma*	151	24	4	76	20	67 (60-74)	3	0	1.6 (1.0-4.1)
PCPG	*Pheochro-mocytoma and paraganglioma*	144	0	0	100	0	61 (56-65)	0	1	--
PRAD	*Prostate adenocar-cinoma*	490	36	5	94	1	62 (54-70)	6	0	9.6 (9.6-9.6)
READ	*Rectum adenocar-cinoma*	83	42	0	85	15	63 (54-73)	2	0	3.9 (3.9-3.9)
SARC	*Sarcoma*	181	41	0	100	0	58 (46-69)	0	1	6.7 (3.1-6.7)

Table 3.1 (cont'd)

SKCM	*Skin melanoma*	378	85	15	83	2	61 (50-70)	0	75	7.4 (2.6-20.1)
STAD	*Stomach adenocar-cinoma*	263	37	4	70	25	67 (58-73)	0	0	4.6 (1.3-4.6)
TGCT	*Testicular germ*	134	0	4	92	4	31 (26-37)	0	0	--
THCA	*Thyroid*	501	73	6	80	13	46 (35-58)	8	1	--
THYM	*Thymoma*	106	45	6	85	9	58 (48-68)	1	0	9.6 (9.6-9.6)
UCEC	*Uterine corpus*	146	100	43	57	0	65 (57-72)	14	0	9.2 (3.6-9.2)
UCS	*Uterine carcinosa-rcoma*	4	100	0	75	25	63 (54-74)	0	0	1.4 (0.3-2.2)
UVM	*Uveal melanoma*	78	45	0	100	0	62 (51-74)	0	0	3.8 (2.4-3.8)

*: Only the three most abundant ethnicities in the data set were considered to calculate
**: Survival quantiles for cancer types with less than five death events were not

The first 50 main axes of variations of the extended omics matrix were selected using an apparent bend in the scree plot of Eigen-values (Material and Methods). The projection of the 50 axes onto two dimensions is shown in Figure B.1. As expected, cell-of-origin effects dominate the clustering of tumors at a pan-cancer level, with clusters enriched by previously reported pan-cancer clusters (e.g., collection of gastric cancer, gliomas, kidney, and squamous tumors), types, and subtypes (e.g., Luminal and Basal breast tumors), and single cancer types (e.g., Thyroid carcinoma, Prostate adenocarcinoma).

3.4.2. Classification of pan-cancer tumors after removing tissue-specific signals

Once the tissue signal was identified, it was removed from the extended omic matrix. Next, sparsity constraints were imposed on the omic features to zero out those with irrelevant contributions to axes of variation and cluster formation. The selected features (i.e., with non-zero effects) across the three omics corresponded with the 18th, 25th, 33rd, and 38th axes (sorted from more to minor variance explained) and mapped onto a total of 1200 genes. The cluster identification and projection onto two dimensions revealed eight classes (Figure 3.2). Because of removing the effects of tissue localization, all clusters were formed by samples from multiple cancer types. Some clusters differed statistically from their cancer types composition (Table 3.2). However, all cancer types overlapped with more than one cluster (Figure 3.2; Table 3.2, bottom). Furthermore, this overlap was not influenced by previously reported subtypes (Figure B.2).

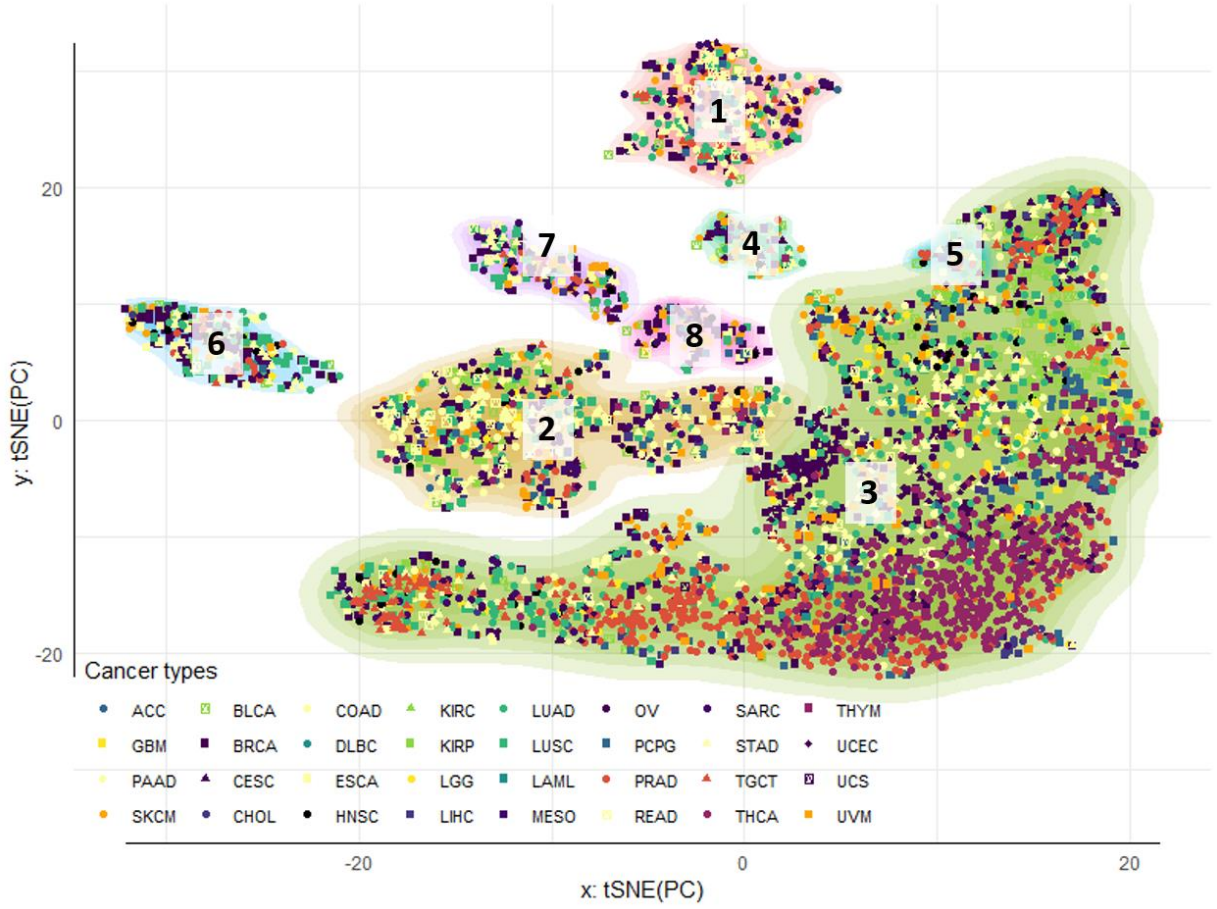

Figure 3:2: Pan-cancer clustering of tumor samples after adjusting for tissue effects. Tumor clusters were obtained by sequential application of tSNE and DBSCAN algorithm for 5,408 samples across 33 cancer types. The contours reflect cluster membership, and the points' colors and shapes represent similar anatomical sites and cancer types, respectively. The two-dimensional tSNE projection was obtained from the four deep principal axes of the extended omic matrix projected outside the tissue-specific effects after sSVD and removing the first two axes. After re-classifying tumors, the few samples from Kidney chromophobe tumors (KICH) did not map in any of the eight clusters obtained.

3.4.3. Clinical and demographical characterization of tumor clusters

Clusters differed statistically in terms of patient age (with Cluster 3 and 8 containing samples from slightly younger patients) and sex (with Clusters 2 and 7 having significantly more females than Cluster 8, due to their slightly higher composition of gynecological cancers) (Table 3.2). However, none of the clusters were significantly associated with ethnicity (Table 3.2).

Table 3.2: Clusters characterization after removing tissue effects. The clusters produced by the integration of whole-genome profiles of gene expression (GE), copy number variants (CNV), and DNA methylation (METH) were characterized in terms of clinical, demographic, immune, and molecular information. The table shows those variables with significant differences in at least one cluster. For each variable, different letters represent significant differences between clusters.

	Clusters	1	2	3	4	5	6	7	8
Clinical information	*Cancer type*[#]	bc	c	d	ab	ab	ab	bc	a
	Metastasis (%)	5c	4de	3e	17ab	5de	7cd	12bc	21a
	*Survival time (years)**	2a	2a	3b	2a	2ab	2ab	2ab	2a
	Stage [119]	IVab	IVbc	IIIc	IVab	IIIabc	IIIab	IIIabc	IVab
	Tumor-free fraction (%)	60a	70a	80b	60a	60a	60a	60a	60a
	Intratumor heterogenity (%)	13ab	14ab	4d	10c	15a	12abc	14ab	9bc
	Proliferation	0.4a	0.3a	-0.4b	0.3a	0.3a	0.4a	0.4a	0.5a
Demographic information	*Age (years)*	61a	62a	57b	60ab	60ab	61ab	62a	57b
	Sex (% of females)	52ab	54a	50ab	50ab	53ab	46b	58a	41b
Genome instability	*Non-silent mutation*	2bc	2bc	1d	3a	2abc	2c	3ab	2bc
	Aneuploidy	12a	12a	3b	10a	14a	11a	12a	10a
	Homologous recomb defects	22ab	16c	8d	23ab	22abc	25a	27a	19bc
Immune infiltration	*Th1 CD4+ cells (x10^{2})*	-6b	-6b	-3a	-7b	-8b	-7b	-6b	-6b
	Th2 CD4+ cells (x10^{2})	3c	2c	2c	4ab	5abc	5ab	5ab	6a
	Th17 CD4+ cells (x10^{2})	-8b	-8b	6a	-15c	-5b	-5b	-9b	-9b
	Activated natural killer cells (x10^{-1})	2bc	2bc	3a	3ab	2bc	1c	2bc	2bc
	Lymphocytes (x10^{-2})	5bc	6b	4a	4bc	5bc	3bc	5bc	3c
	Tumor-infiltrating lymphocytes	1.7b	1.7b	1.9a	1.7b	1.8ab	1.6b	1.8b	1.6b

Table 3.2 (cont'd)

Functional classes**	*DNA replication*[&¶,(1)] *(x10^{-1})*	-6d	6a	-1bc	6a	4ab	7a	-3c	-2bc
	Mitochondrial translation[&¶,(2)]	0.4d	-0.3b	0.0c	-0.9a	0.3cd	-1.1a	1.9e	0.5d
	mir-has-615b targets[I ,(3)]	-1.1c	0.7a	-0.1b	0.7a	-0.2b	0.8a	-1.1c	-0.1b
	S phase and DNA synthesis[¶,(4)]	-1.5f	1.0b	-0.1d	0.5c	0.3c	1.3a	-0.4e	-0.4e

[#]Cancer types by cluster (%).

C1	COAD (14.2), LUAD (11.7), BRCA (10.7), SKCM (8.1), SARC (7.1), READ (6.4), PRAD (4.8), ESCA (4.6), CESC (4.1), LUSC (4.1), STAD (4.1), BLCA (3.8), PAAD (3.6), TGCT (2.5), ACC (2.3), MESO (2), LIHC (1.5), UCEC (1.5), PCPG (1), HNSC (0.8), KIRC (0.3), LGG (0.3), OV (0.3), and UVM (0.3).
C2	BRCA (11.1), COAD (11.1), STAD (9.6), LUSC (7.4), LUAD (7.1), SKCM (6.1), CESC (5.6), BLCA (5.4), SARC (5.4), READ (4), ESCA (3.1), KIRP (2.5), PAAD (2.5), PRAD (2.5), PCPG (2.2), HNSC (1.7), LIHC (1.5), UVM (1.5), MESO (1.4), UCEC (1.4), ACC (1.3), KIRC (1.1), GBM (1), THYM (1), LGG (0.8), THCA (0.7), TGCT (0.6), DLBC (0.1), and LAML (0.1).
C3	THCA (16.1), PRAD (13.2), BRCA (9.3), LUAD (6.3), SKCM (4.4), BLCA (4.3), LUSC (3.9), STAD (3.8), COAD (3.4), TGCT (3.4), UCEC (3.4), PAAD (3.3), CESC (3.2), THYM (3.2), PCPG (3.1), LGG (2.5), SARC (1.7), UVM (1.6), HNSC (1.3), LIHC (1.2), KIRC (1.1), MESO (1.1), ESCA (1), GBM (1), LAML (0.9), DLBC (0.7), READ (0.5), KIRP (0.4), CHOL (0.4), UCS (0.1), ACC (0.1), and OV (0.1).
C4	SKCM (21.7), BLCA (13), CESC (9.6), LUAD (9.6), LUSC (8.7), BRCA (7.8), ESCA (4.3), UVM (4.3), MESO (3.5), HNSC (2.6), SARC (2.6), GBM (1.7), LIHC (1.7), STAD (1.7), UCEC (1.7), COAD (0.9), KIRP (0.9), PRAD (0.9), READ (0.9), TGCT (0.9), and THYM (0.9).
C5	BLCA (18.4), LUAD (15.8), CESC (10.5), SKCM (10.5), PRAD (7.9), BRCA (5.3), ESCA (5.3), STAD (5.3), COAD (2.6), GBM (2.6), HNSC (2.6), LIHC (2.6), LUSC (2.6), PAAD (2.6), PCPG (2.6), and TGCT (2.6).
C6	BRCA (31.5), LUSC (9.7), ESCA (8.6), SKCM (8.6), BLCA (8.2), STAD (6.5), LUAD (5.7), PRAD (5.7), HNSC (3.9), CESC (2.5), SARC (2.2), PAAD (1.8), GBM (0.7), LGG (0.7), UCEC (0.7), UVM (0.7), CHOL (0.4), DLBC (0.4), MESO (0.4), PCPG (0.4), READ (0.4), and TGCT (0.4).
C7	SKCM (14.7), BRCA (11.5), LUSC (11), ESCA (8.4), STAD (7.3), SARC (6.8), CESC (5.8), LUAD (5.8), UVM (4.7), BLCA (4.2), PAAD (3.1), HNSC (2.6), COAD (2.1), PRAD (2.1), LIHC (1.6), MESO (1.6), READ (1.6), UCEC (1.6), TGCT (1), DLBC (0.5), GBM (0.5), LGG (0.5), OV (0.5), and THCA (0.5).

Table 3.2 (cont'd)

C8	SKCM (24.8), BRCA (23.9), CESC (12.8), PCPG (6.8), BLCA (5.1), SARC (5.1), LUSC (4.3), HNSC (3.4), UCEC (2.6), COAD (1.7), ESCA (1.7), MESO (1.7), READ (1.7), TGCT (1.7), LUAD (0.9), OV (0.9), and UVM (0.9).

*Values represent median survival times by cluster. Letters represent significant differences under the log-rank test to compare the entire survival curves of each cluster.

**Databases: GO Biological process ([&]), miRTabrBase ([ǀ]), Reactome ([¶]). Functional classes significant at FDR adj. p-value < 0.05.

Overlap between a selected group of genes and databases:

[(1)]*: GINS1, POLD3, PRIM2, POLD4, PCNA, MCM8,* and *MCM3.*

[(2)]*: MRPS26, MRPL2, MRPL51, MRPS35, MRPL16, MRPS18A, MRPS10, MRPL14, MRPL48, MRPL21* and *MRPL11.*

[(3)]: *PANK2, SF3B2, PCNA, HSP90AB1, NOP2, ATN1, CHD4, HOXC13, PRICKLE4, DPP3, C12ORF57, LDHB, CCND3, CCND2, STK35, RAB23, PPP6R3, IDH3B, RPS3, SIRPA, PSMF1, DNM1L, NKX2-5, PRNP, UVRAG, PPIL1, TPI1, DST, CSNK2A1, SMOX, YIPF3, DDX11, ENTPD6, MAD2L1BP, PPP2R5D, MUT, FBXL14, MRPL21, KLHL42, WNK1, RPL7L1, NCAPD2, FKBP4* and *GAPDH.*

[(4)]: *GINS1, POLD3, PRIM2, POLD4, PCNA, CDKN1B, CCND1, MCM8, MCM3, PSMF1* and *CDC25B.*

The most notorious distinctions between clusters were their differences in prognosis and severity traits (Figure B.3). Cluster 3 (the largest cluster in Figure 3.2) was distinguished by better prognosis/less severity cancers than the remaining clusters, followed by Clusters 2, 5, 6, and 7. In general, clusters 4 and 8 had the worst prognosis and more aggressive tumors (Table 3.2). Cluster 3 was also the one with the fewest metastatic samples (Figure B.4), higher survival rates, highest tumor-free fraction, lowest stage, lowest intra-tumor heterogeneity (ITH, that estimates the fraction of sub-clonal and clonal genomes in each sample[108]), and lowest proliferation (Table 3.2, Figure B.3). By comparison, Clusters 4 and 8 had significantly more metastatic samples than Cluster 3. Cluster 8 also had higher ITH rates than Cluster 3. The highest ITH rates were found in Cluster 5.

Cluster 3 also had the lowest rates of non-silent mutations, aneuploidies, and homologous recombination dysfunction (HRD). The remaining clusters were very similar in terms of genome instability indicators, except for Cluster 2. This cluster had significantly higher rates of HRD than Cluster 3 but significantly lower rates than every other cluster (Table 3.2). Cluster 3 was characterized by the highest rates of tumor-suppressive immune cells and tumor-infiltrating lymphocytes (Table 3.2). In addition, Cluster 6 had the lowest infiltration of activated natural killer (ANK) cells. Cluster 8 also had the lowest lymphocytic and highest Th2 CD4+ infiltrations, respectively (Table 3.2).

3.4.4. Gene signatures characterizing tumor clusters.

The clusters were also characterized by distinct sets of omic features, significantly enriched for functions involved in the cell cycle (DNA replication, DNA synthesis, and targets of hsa-mir-615-b, a micro-RNA involved in cell proliferation) and mitochondrial translation (initiation, elongation, and termination) (Table 3.2). To study the pairwise differences across clusters, these gene sets were projected onto scores for each gene, as linear combinations between the features' values mapping onto the gene (i.e., its expression, methylation, and copy number values) and their related activities (i.e., the feature's effects arising from the sparsity constraints) (see Materials and Methods section). In general, Cluster 3 was characterized by intermediate values of these scores, while the remaining clusters were characterized by higher (i.e., gene set with higher expression than Cluster 3) or lower (gene sets with lower expression than in Cluster 3) gene set scores. Clusters 2, 4, and 6 had significantly higher scores for cell proliferation and significantly lower for mitochondrial translation. On the other hand, clusters 1, 7, and 8 had significantly lower scores of proliferation and higher for mitochondrial translation.

Sparse factorization of the extended omic matrix resulted in the selection of features mapping onto 1200 genes. From this list, 441 genes were significantly different in at least one cluster. These results were obtained by a series of analyses of variance (ANOVAs), using the scores of each gene as response variables and clusters as explanatory variables. This list included 34 validated cancer genes, including oncogenes (*ERC1*, *HSP90AB1*, *NUMA1*, *PPFIBP1*, *ZNF384*, *CHD4*, *KRAS*, *HIST1H3B*, *CCND1*, *CCND2*, *PIM1*, *CCND3*, *HMGA1*, *HOXC11*, *HOXC13*, *KDM5A*, *SRSF3*, *TFEB*), tumor suppressors (*FANCE*, *CDKN1B*, *ASXL1*, *ETNK1*) and fusion-proteins (*ERC1*, *HSP90AB1*, *NUMA1*, *PPFIBP1*, *ZNF384*). Many genes also mapped onto known transcription factors (including *KDM5A*, *RELA*, *SRF*, *CTBP2*, *FOXA2, NONOG, FOLSL1, TEAD4,* and *FOXM1*) and some of their targets (Figure B.5). However, the expressions of TFs and their targets were not significantly correlated within or between clusters (Figure B.5), suggesting mechanisms of control of the gene expression other than TFs regulation.

We then interrogated all pair-wise comparisons between the scores of the 441 significant genes using Tukey tests (Table B.1). We identified a subgroup of 123 significant genes that distinguished each cluster from the rest (for example, *POLH* had significantly higher scores in Cluster 4 than in every other cluster). The genes characterizing each cluster were then used to define signatures. With this criterion, only Clusters 1, 4, 6, 7, and 8 were characterized by distinct signatures of 57, 4, 23, 24, and 15 genes each, respectively. Since the gene scores are combinations of omic features, we looked at the gene expression in each signature and the potential role of copy numbers and methylation in regulating it (Figures 3.3-4).

Cluster 1's signature was composed of genes mapped on chromosome 20. A group of 56 of the 57 genes exhibited significant copy losses in Cluster 1. Of this group, 50 genes (*ATRN, AP5S1, TMEM230, MGME1, NDUFAF5, CENPB, CRLS1, CRNKL1, CSNK2A1, DDRGK1, DSTN, DTD1, ESF1, FAM110A, FASTKD5, FKBP1A, IDH3B, ITPA, SMIM26, MAVS, MCM8, MKKS, MRPS26, NAA20, NOP56, NRSN2, NSFL1C, PANK2, PCNA, POLR3F, PSMF1, PTPRA, RBBP9, RBCK1, RRBP1, SIRPA, SMOX, SNPH, SNRPB2, SNRPB, SNX5, SOX12, STK35, TBC1D20, TRMT6, UBOX5, VPS16, ZCCHC3, ZNF133* and *ZNF343*) were also downregulated. The genes with significant copy-losses and basal expression values (*TGM6, SOX13, PROKR2, PRND, OXT, LRRN4,* and *FERMT1*), *LRRN4,* and *FERMT1* were also significantly hyper- and hypo-methylated, respectively (Figure 3.3).

Cluster 4's signature was composed of four genes mapping onto chromosome 6: *TDRD6*, *POLH*, *PAQR8,* and *GUCA1A*. All these genes exhibited significant copy losses in Cluster 4, and all of them except *GUCA1A* were also downregulated. Additionally, *POLH* was hypo-methylated, while *PAQR8* was hyper-methylated (Figure 3.3).

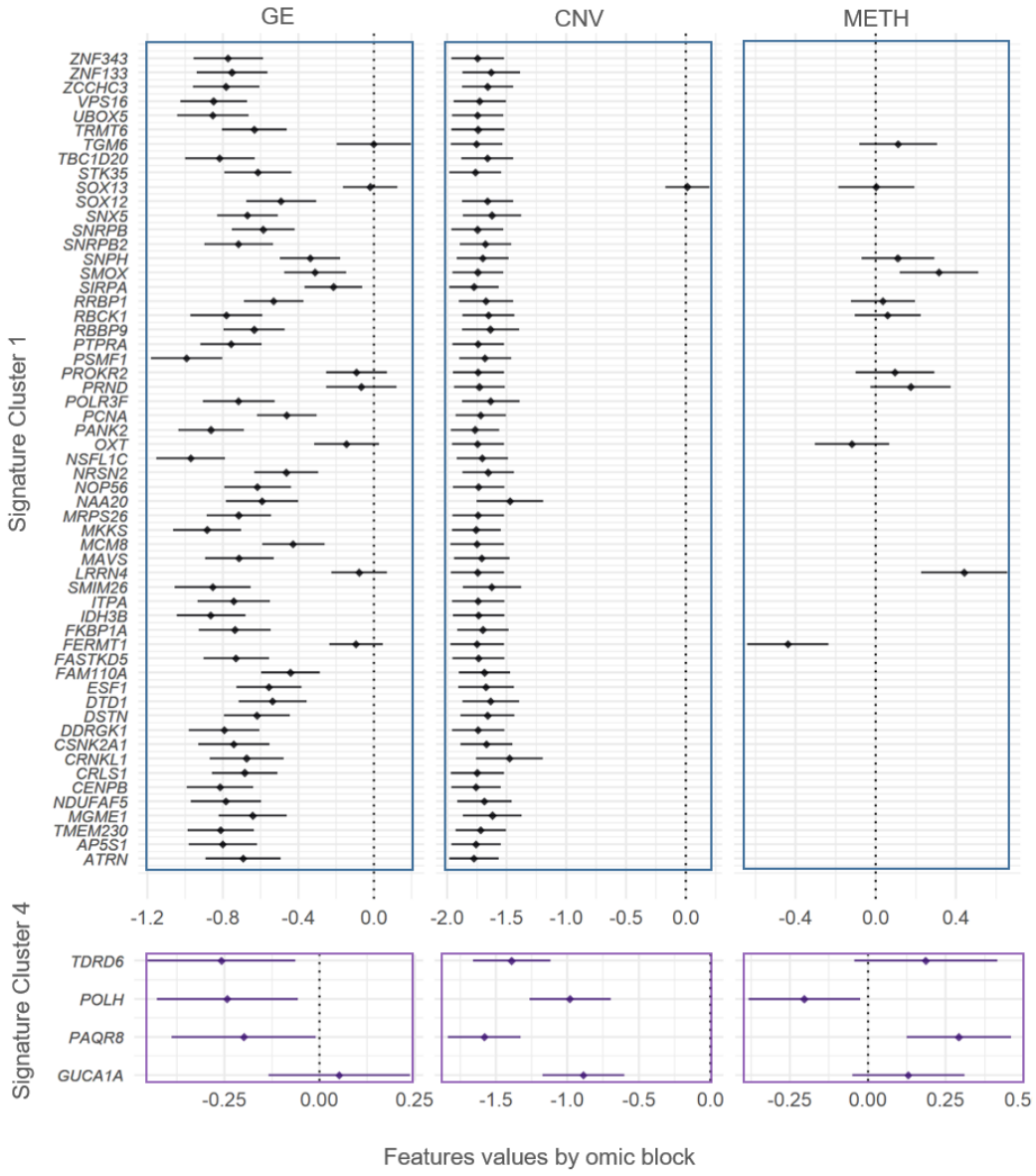

Figure 3.3: Gene signatures for Clusters 1 and 4. The genes significantly de-regulated exclusive of Clusters 1 and 4 were used to define signatures (y-axis). The features values (x-axis) of each gene are separated in gene expression (GE, first column of panels), copy number variants (CNV, second column of panels), and DNA methylation (METH, third column of panels), and summarized by Bonferroni confidence intervals (adjusting for all the 441 significant genes in at least one cluster). Dots represent the average of features values across samples.

Cluster 6's signature was composed of 23 genes mapping onto chromosome 11:

ALDH3B1, ANKRD13D, ANO1, AQP11, ARRB1, EMSY, CCND1, CTTN, KRTAP5-10,

LRP5, LRRC32, TESMIN, MYO7A, NUMA1, PAK1, PPFIA1, RBM4, RPS6KB2, RSF1,

SHANK2, TMEM134, TPCN2, and *USP35*. These genes exhibited significant copy gains, and all of them were also significantly upregulated, except for three genes with basal expression in Cluster 6: *MYO7A*, *LRRC32*, and *ALDH3B1*. In addition, genes *USP35, SHANK2, MYO7A, LRRC32, CTTN, CCND1, ARRB1*, and *ALDH3B1* were additionally hypo-methylated, while genes *RSF1* and *PPFIA1* were hyper-methylated (Figure 3.4).

Cluster 7's signature was composed of 24 genes mapping onto chromosome 6. All of these genes (*BTBD9, RRP36, CCND3, CNPY3, CUL7, FRS3, GUCA1A, BICRAL, KLC4, KLHDC3, LRFN2, MEA1, MED20, MRPL2, MRPS10, PEX6, PPP2R5D, RPL7L1, SRF, TAF8, TBCC, TOMM6, TRERF1,* and *UBR2*) exhibited significant copy gains. In addition, all of them were significantly up-regulated, except by *LRFN2, GUCA1A, BTBD9*, which had basal levels in Cluster 7. Genes *TRERF1, LRFN2,* and *FRS3* were additionally hypo-methylated, while *GUCA1A* was hyper-methylated (Figure 3.4).

Cluster 8's signature was composed of 15 genes mapping onto chromosome 11. These genes (*ALDH3B1, ANO1, CCND1, CPT1A, CTTN, LRP5, MRPL21, NADSYN1, PPFIA1, RNF121, RSF1, SHANK2, TPCN2, UNC93B1,* and *USP35*) exhibited significant copy losses. All of them except *ANO1* (with basal levels in cluster 7) were significantly downregulated. In addition, genes USP35 and *NADSYN1* were significantly hypermethylated, while *UNC93B1, RSF1, MRPL21,* and *ANO1* were hypo-methylated (Figure 3.4).

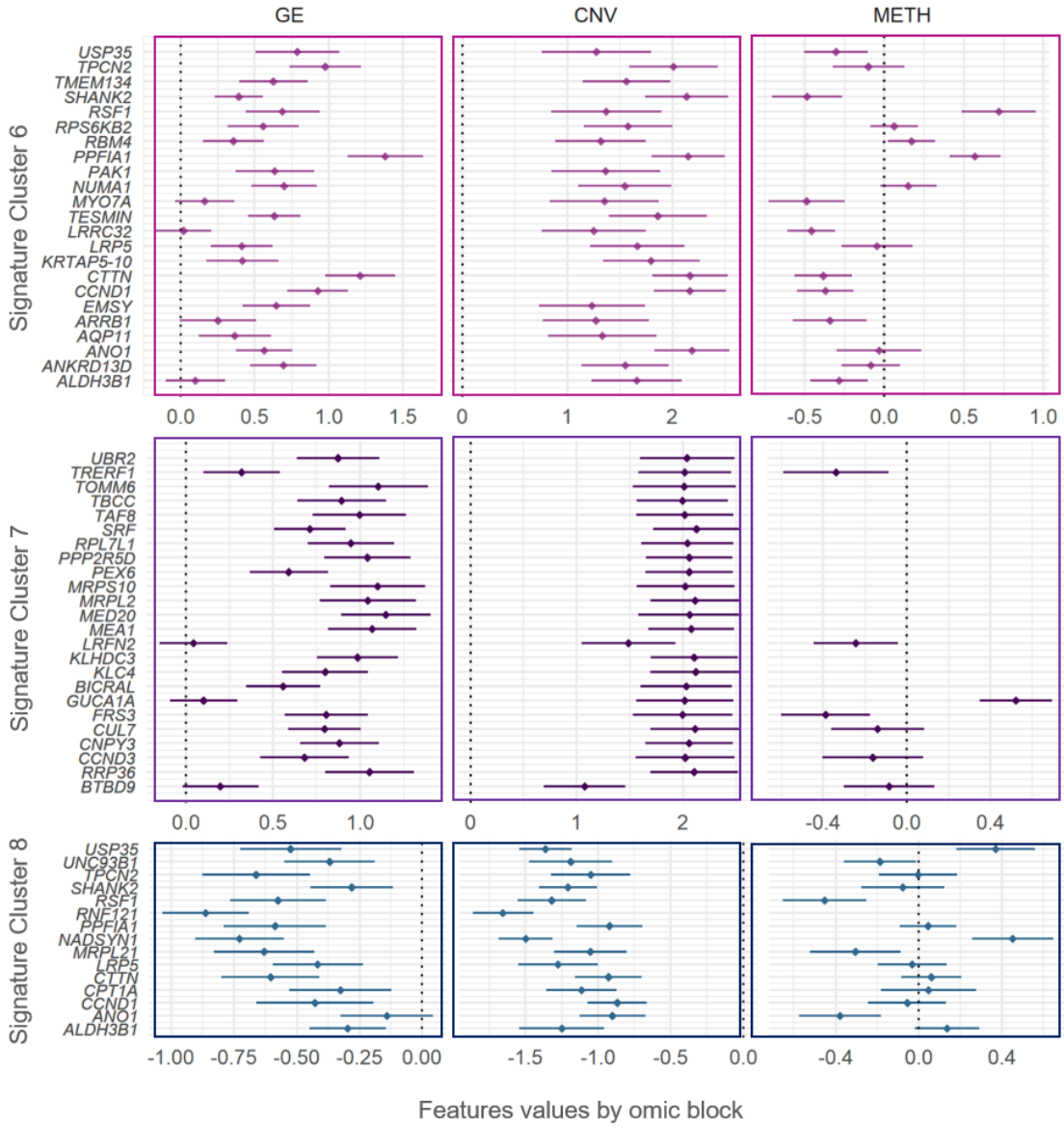

Figure 3.4: Gene signatures for Clusters 6, 7, and 8. The genes significantly de-regulated exclusively in Clusters 6, 7, and 8 were used to define signatures (y-axis). The features values (x-axis) of each gene are separated in gene expression (GE, first column of panels), copy number variants (CNV, second column of panels), and DNA methylation (METH, third column of panels), and summarized by Bonferroni confidence intervals (adjusting for all the 441 significant genes in at least one cluster). Dots represent the average of features values across samples.

3.5. DISCUSSION

Most pan-cancer classifications rely on molecular alterations that discriminate between the tissue of origin [72,100–102,120]. However, as soon as tissue effects were removed, we have found that the cancer signal immediately emerged. Distinct cancer classes were formed, containing tumors from different cancer types. Particular functional groups of omic features also characterized these classes. An SVD of the extended omics matrix can result in a multitude of axes of variation. Such axes have the potential to explain different patterns of variability across subjects. In this study, we preceded our cluster analysis by selecting axes of variation (i.e., basis vectors spanning the features space of the concatenated omics) having features loadings different from zero (each axis of variation has an accompanying vector of loadings representing features activities). We have obtained the cluster display in Figure 3.2 as a result of this selection criterion.

Furthermore, most of the variability between clusters of tumors associates with the canonical relationship between gene expression and copy number. According to this, the primary source of co-variability among features seemed to be dominated by positive covariation of expression and copy number (i.e., copy losses match with lower expression levels, and vice versa, Figure B.5). The expression of regulatory elements within the group of selected features (including transcription factors and the micro-RNA hsa-mir-615b) was, on the other hand, not associated with the expression of their predicted targets. These observations support the role of copy numbers as a significant force affecting tumor progression [121–123]. Experimental evidence has shown large effects of methylation at characterizing both normal and tumor tissues [124–127]. Contrarily, epigenetics has an essential role during tissue differentiation, as well as in cancer.

However, our analysis might suggest a minor role in leading the cancer cluster differences. We believe that this minor role could be the result of an intense correction for tissue-specific effects. Other possible explanations include artifacts of data processing, such as summarizing methylation at the CpG island level. Although the map at the CGI level covered both genic and non-genic regions and facilitated computations, this summary could have come at the cost of washing out CpG site-specific effects on cancer. A third possibility is that the abnormal methylation patterns are essential but shared by two or more cancer clusters. Our features highlighted are the ones that differentiate between clusters. Regardless, we observed abnormal methylation patterns that might suggest a role in the expression of some genes characterizing tumor classes (e.g., expression of *LRN4* and *GUCA1A* negatively correlated with average methylation of promoters' CpG islands).

The tumor clusters C1, C4, C6, C7, and C8, had exclusive signatures (i.e., different from every other cluster). Interestingly, the clusters without distinct individual signatures had more favorable outcomes (C3, C2, and C5). One possible explanation for this is the frequent correspondence between more dramatic molecular alterations and worse clinical outcomes [128,129]. To gain insights about possible biological interactions within each signature, we used the accompanying bibliographic results provided by the STRING database [115] (see Material and Methods section). The literature suggests a broad overlap between signatures in terms of gene functions (cell growth, division, small RNA metabolism, protein synthesis, maturation and transport, and mitochondrial dysfunction). In the case of signature C1 (most genes down-regulated), the literature suggested *NOP56* (a core component of the small nucleolar ribonucleic protein) as a central element in the

signature; interacting with *MKKS, NAA20,* and *PTPRA* (genes with roles on mitotic division); *ESF1, SNRPB, SNRPB2, POLR3F* and *CRNKL1* (involved in small RNA processing), *PCNA* and *ITPA* (involved in DNA replication and repair), *UBOX5, RRBP1*, *RBCK1* and *NRSN2* (protein synthesis, maturation, and antigen presentation), *RBBPP9* (resistance to growth inhibition of TGF); *SIRPA* and *DSTN* (cell adhesion)[130–133]. In the signature C1, *NOP56* could be a candidate for future therapeutic intervention. Tumor suppressors *NRSN2* and *RBCK1* could also be considered.

The three downregulated genes from signature C4 were involved in small RNA maturation (*TDRD6*, micro-RNA expression, and maturation), cell proliferation (*PAQR8*, plasma membrane progesterone receptor), and DNA repair (*POLH*, DNA polymerase involved in DNA repair). From these genes, *PAQR8* and *TDRD6* could represent potential targets of therapy. Although neither of them has been directly related to cancer, other members of the PAQR family of progesterone receptors are known tumor suppressors, while *TDRD6* has been reported as frequently down-regulated in breast cancer, suggesting its potential use as a biomarker [134]. In the case of signature C6 (most genes upregulated), the literature suggests *CTTN* as interacting with two groups of genes within the signature, either by co-expression or co-localization in amplicons. One group consisted of invasion and anti-apoptotic related genes (e.g., *SHANK*, *PAK1*, *PPFIA1*) and ion transport (*ANO1* and *TPCN2*) [135,136]. The other group consisted of *CCND1* (cell cycle checkpoints), *LRPS* (protein synthesis), *RSF1* (chromatin remodeling), and *USP35* (protein turnover; through amplicon-mediated overexpression in breast and gynecological cancers) [137,138]. Thus, patients with signature C6 could perhaps benefit from *ANO1* inhibitory therapy [136].

Signature C7 was characterized by multiple genes co-expressing with *KLHDC3* (involved in homologous recombination): *MEA1* (spermatogenesis), *CNPY3* (protein folding, antigen presentation), *PPP2R5D* (direct catalytic activity), *RRP36* (small RNA synthesis), *CCND3* (cyclin, cell cycle checks points), and *MED20* (transcription). *KLHDC3* also belongs to the protein turnover and antigen presentation pathway, together with *CUL7* and *UBR2*. The literature also suggests another group of co-expressing genes within signature C7, consisting of *RPL7L* (ribosome), *MRPL2,* and *MRPS10* (mitochondrial ribosome). These genes have also been found to interact in cell culture [140,141] physically. Signature C8 genes remarkably overlapped with signature C6 genes but exhibited opposite regulation (i.e., down- instead of up-regulated). Additionally, the literature suggests the interaction between *CCND1*, *NADSYN1,* and *MRPL20* in signature C8 [139,140]. *NADSYN1* has been proposed as a target of inhibitory therapy in cancer [141], while *MRPL20* has been suggested as a biomarker for gastric cancers [142,143].

The molecular classification of tumors generated clusters with clear differences in prognosis and severity, with C3 exhibiting better outcomes than the remaining clusters. C3 also resembled a previously reported “inflammatory” type in terms of immune infiltration and cancer type composition (enriched for prostate adenocarcinoma, thyroid, and pancreatic carcinomas and having elevated values of markers for CD4+ Th17 and Th1 cells and low genomic instability) [108]. Although the remaining clusters were clearly distinguished in terms of altered molecular processes, they were highly similar clinical and demographic characteristics. C3 also differed from the remaining clusters by lacking large CNV. In C3, we do not observe drastic genome alterations being systematically linked with worse cancer outcomes, either by causing loss of tumor-suppressing activities

(e.g., mitotic checkpoints, DNA instability sensing, pro-apoptotic activity), or gain of oncogenic function (e.g., duplication of mitotic factors). In either case, large CNV have been associated with worsened clinical outcomes, in contrast with those characterizing C3. This observation is somewhat supported by less aggressive cancers in C3 (e.g., a high frequency of prostate and thyroid cancers), co-located with low severity cases of more aggressive tumor types. Another example of less aggressive tumors in C3 are Her2+ breast cancer and proximal inflammatory lung adenocarcinomas, tumors of less severe outcomes than their luminal/basal and proximal proliferative subtypes, respectively [144][145]. Since similar signaling deregulation can arise in different cancers (e.g., dysregulated PI3K/AKT/mTOR pathway in gynecologic cancer) [146], further research on the link between shared molecular signatures within tumors in the same cluster could shed light on the development of novel therapies, or the repurpose and combination of existing ones. Given their small molecular weights, targeting oncogenes with common monoclonal antibodies and small-molecule tyrosine kinase inhibitors could aid in the treatment of tumors with overexpressed oncogenes [147]. For instance, tumors with signature C6 could benefit from combined therapy with indirubin and Ani1, inhibitors of *CCND1* and *ANO1* [148,149]. On the other side of the spectrum, targeting tumor suppressors on signatures of downregulated genes also presents exciting opportunities. For instance, tumors with signature C1 could benefit from target therapy for tumor suppressors *NRSN2* and *RBCK1*. Classic approaches for targeting tumor suppressor genes include re-activation by either re-introducing a functional copy (e.g., gene therapy) or diminishing the repressive action of other players through small-molecule inhibition [150].

Nevertheless, given the technical challenges of targeting loss of tumor-suppressing function, signatures exhibiting up-regulation could have more pharmacological potential. Similarly, signatures could also rapidly address differences in tumor heterogeneity (e.g., C8 and C5 were notoriously more heterogeneous than the rest). Finally, differences in immune infiltration (C6 with the lowest activated natural killers' infiltration and C8 with the lowest lymphocytic one) could also imply the potential use of signatures to aid immunotherapeutic decisions.

Our results included genes frequently duplicated in cancer (e.g., *KRAS*, *CCND1*). However, other frequently duplicated genes, like *ERBB2*, *MYC*, and *FGFR1,* were not present in our selected set of features. One possibility for this unexpected result could be a limitation of the EN penalty as a feature selection criterion. For example, an EN parameter of 0.5 could have been too stringent, and groups of correlated and relevant features could have been left out during the selection process. Another possibility is that the effects of these frequent duplication were washed out by the tissue correction. While removing the tissue "environment" evident in the omics, treating tissue as a systematic effect in a linear model, a different way of having defined tissue effects would have been to use tumor histology markers (e.g., mesenchymal, epithelial). Additionally, while removing dominant differences on tissues that may be unrelated to cancer, it could also eliminate differences that may allow certain cancers to progress in that specific tissue. The third possibility is that the effects of these events were not essential for the cluster partition.

Given the possibility of unveiling different biological channels altered in tumors of similar clinical and molecular characteristics, we believe this novel pan-cancer classification

could aid in identifying therapies for cancers without a standard of care. However, extrapolation of results herein should be exerted with the following caution. Although our data included information from multiple studies, sexes, ages, and ethnicity, our results could be strongly influenced by factors such as the country of origin of each study and biased on demographic characteristics. Further application of our methods to tumors from patients from diverse populations and ages would be essential for an effective generalization of our results.

CHAPTER 4

PHENOMIC DATA INTEGRATION IN THE UK BIOBANK REVEALS GENETIC VARIANTS INVOLVED IN ENERGY BALANCE

4.1. ABSTRACT

Excessive calorie intake and low physical activity contribute to a positive energy balance (EB), leading to obesity. Although EB is affected by several environmental and socioeconomic factors, a sizable part of its variation is still due to genetics. The study of EB as a target phenotype is challenging, as EB combines several layers of data, including whole-body size and composition, food selection and amount ingested engagement in physical activity, and metabolic profiles. Many genes (e.g., *FTO*, *MCR4*, *ANKRD33*, *FIGNL2*) have been identified as contributing to the variation observed in different components of EB. However, a complete set of genes is still missing. We integrated several phenotypes from the UK Biobank related to EB to increase our knowledge on EB's genetic and molecular basis. We have used sparse factor analysis to define patterns of energy balance (PEB) and determined genomic regions of interest with potential causal effects in EB.

4.2. INTRODUCTION

Obesity is affected by multiple environmental and socioeconomic factors [151]. However, it is also considered a heritable condition [152,153]. As such, obesity has been linked to the variation of many genes affecting energy balance (EB, the difference between energy intake and energy expenditure), including *FTO*, *MCR4*, *ANKRD33*, *FIGNL2* [154,155]. Nevertheless, due to the multiple phenotypes involved in EB, the complete set of genes affecting EB is still not available [156,157].

EB is determined by multiple variables affecting the thermic effects of feeding (e.g., frequency and quantity of different food items), resting metabolic rate (e.g., body size and composition), and physical activity (PA, e.g., intensity, frequency, and duration) [157]. Factor analysis methods can capture the complex covariation among all these different types of variables [158,159]. Under this framework, multiple phenotypes can be summarized by a few factors representing patterns of EB. For example, [160,161] have used sparse latent factor models to estimate dietary patterns and variations across an exercise intervention. On the other hand, [162] used principal functional components to estimate the temporal variation of PA, while [163] used reduced rank regression to combine metabolite and dietary patterns and their link with type I diabetes. Once factors related to EB are derived, they can be used as responses in genome-wide association studies (GWAS) [154,164]. Advantages of factors as GWAS responses include lower type I errors, increased power, and faster computations [165–167].

Several studies support the association between EB components and genomic variation [154,155]. For instance, selection experiments in mice have demonstrated strain variation in the predisposition to engage in physical activity of different intensity [168–170] and

preference of certain foods over others [171,172]. In addition, human studies on the genetics of PA have reported heritability values as high as 63% [173–176]. Furthermore, association studies have shown the putative roles of many genes involved in food preference and adherence to physical activity [164,177–179]. Moreover, specific genes (like the fat mass and obesity-associated protein, *FTO)* have also been shown to have pleiotropic effects on diet, physical activity, body size, and obesity risk [180–183].

In this work, we have used extensive phenomics and genomic information from the UK Biobank cohort to derive patterns of EB (PEB). These PEB captured distinct and relevant aspects of food preference, body size and composition types, physical activity tendencies, and metabolic profiles. Furthermore, the association between PEB and genomic variation has revealed novel genes affecting multiple components of EB.

4.3. MATERIAL AND METHODS

4.3.1. Cohort

This study used data from a subset of 219,049 participants from The UK Biobank (UKB) cohort [14] to derive energy intake and expenditure patterns and study their associations with genotypic information. Edition and quality control criteria for the phenotypic variables criteria included the exclusion of participants of non-Caucasian/European ancestry, exclusion of related individuals (KING's kinship value lower than 0.03 to exclude 1st, 2nd, and 3rd degrees of relationship) [184], exclusion of participants with abnormal biomarker levels, and exclusion of participants reporting unrealistic values of food intake and energy expenditure (Figure 4.1).

Phenotypes**:** We used 28 phenotypic variables with possible roles at affecting an individual's energy requirements and expenditure, split into four groups: *1) blood biomarkers*, including glucose, creatinine, triglycerides, and cholesterol levels, *2) body composition*, including body weight, standing height, waist circumference, fat mass and lean mass, *3) diet*, recorded with the ACE touch screen questionnaire on frequency of consumed items in the last year, including cooked and raw vegetables, fresh fruit, oily and non-oily fish, poultry, beef, pork, lamb or lamb, and processed meats; cheese, bread, cereal, tea, coffee, water, and alcohol (see detail in [185], and *4) physical activity* (as intensity, frequency, and duration).

To correct the effects of drugs affecting blood biomarkers, we followed [186]. First, we used self-reported data on medication taken for cholesterol, blood pressure, diabetes, or exogenous hormones (data field 6153, ACE Touchscreen). Next, drug effects were estimated using the subset of participants that reported taking medicine after the first

recorded instance. For these participants, a drug effect correction factor was calculated as the mean difference between biomarker levels before and after taking the drug. The correction factor was then used to multiply the biomarker levels of those participants taking medicine at the time of recruitment. After the adjustment, any remaining values outside three absolute median deviations (MADs) were considered outliers and excluded from the study.

Diet variables were either numerical (representing the number of servings by a period) or categorical (representing consumption or not of a specific item). The numerical variables were cooked and raw vegetable intake, fresh fruit, bread, cereal, tea, coffee, and water. The categorical variables corresponded to intakes of poultry, cheese, oily and non-oily fish, beef, pork, mutton/lamb, and alcohol. If frequencies for the numerical diet variables were outside 3 MADs, the participant was excluded. If any class within each categorical dietary variable was underrepresented after quality controls, we fused it with the previous ones until all classes had more than one percent of the total sample size. We fused the last three categories into *oily fish*, *non-oily fish*, *beef*, *lamb/mutton*, and *pork* intakes. For *poultry* and *processed meat* intakes, we only fuse the last two categories into one.

An estimate of energy intake is present within the UKB for a subset of participants that answered a web-based 24-hour dietary recall in addition to the ACE touchscreen questionnaire (data field **100002**). We used this subset of individuals to fit a linear model using this energy estimate as response and ACE's derived food items consumption as predictors. Finally, we estimated the total energy intake in the remaining participants missing calory intake estimations. Individuals with an estimated caloric intake larger than 3 MADs were excluded.

We additionally estimated the number of calories expended during PA data derived from IPAQ following [187]. We obtained these values by dividing the product between PA intensity and duration by PA frequency, dividing by 24 to get energy values within the calories/day range. Participants with PA energy expenditure higher than 3 MADs were excluded from our study.

We have used a set of covariables to account for possible confounders of the associations between phenotypes and SNPs. The list included participants sex (**31**, *reported by individual*), age (**21022**, *age of the participant on the day they attended an Initial Assessment Centre, truncated to whole year*), age^2 (*the square of the participant age*), the interaction between sex and age, the UKB assessment center (**54**, *center at which individuals consented to participate in the UKB study*), the first five principal components of the genotypic matrix, Townsend Deprivation Index (**189**, *calculated immediately before participant joining UK Biobank, based on the primary national census output areas, and assigned depending on the output area in which the participant postal code is located*), as a measurement of socioeconomic status, and type of genotyping array used.

Genotypic data: Details on the genotypic data and quality controls provided by the UKB project can be found in [14]. Briefly, genotyping was done using two closely related arrays: UK BiLEVE (~50,000 participants) and the UK Biobank Axiom arrays (~450,000 participants). The UKB Axiom array had over 820,000 SNPs and indel markers, and the UK BiLEVE array was very similar with over 95% shared content. Quality control controls included: excluding low-quality SNPs (missing call rates, low DNA concentration), multi-allelic SNPs (SNP with more than two allelic variants), SNPs departing from Hardy-Weinberg expected frequencies, with sex effects., the array, and batch effects (all test

based on a rejection threshold on the p-values equal to 10^{-12}). SNP genotypes were arranged in a BGData R object [188], where each cell with observed data coded as 0, 1, or 2, indicating the number of copies of the reference allele for each locus at each individual.

4.3.2. Statistical analysis

Derivation of PEB latent variables: Since our set of phenotypes consisted of several continuous and categorical variables, we derived PEB variables by first using Factor Analysis of Mixed Data (FAMD; [189]) and imposed sparsity on the contribution of each phenotype to each PEB using sparse singular value decomposition (sSVD [69]). Essentially, FAMD can be seen as an application of standard singular value decomposition (SVD) on a transformed version of the data that assures the variance of categorical variables does not artificially dominate the construction of latent factors. The transformation involves standardization of each numerical variable (i.e., center to zero means and scale to unit variance) and redefinition of each categorical variable as a set of modified dummy variables. In these dummy variables, ones and zeros are replaced by $1-\frac{n_{cat}}{n}$ and $-\frac{n_{cat}}{n}$, respectively, were n_{cat} is the total number of samples in each category, while n is the total samples size. The transformed phenotypes were adjusted by covariables as described in [190]. We used the following model to derive PEB as latent factors of the transformed phenotypes:

$$\mathbf{Q} = \mathbf{YU\Sigma} + \mathbf{e} \quad \textbf{[Eq.4-1]}$$

To estimate the elements of $\mathbf{Q}$, $\mathbf{U}$, and $\mathbf{\Sigma}$, we minimized the following loss function:

$$L = \|\mathbf{Q} - \mathbf{YU\Sigma}\|_2^2 + \lambda_U(\alpha_U\|\mathbf{U}\|_1 + (1 - \alpha_U)\|\mathbf{U}\|_2^2) \quad \textbf{[Eq.4-2]}$$

Where $\mathbf{Q}_{nxq}$ represents the PEB variables as a set of factors obtained as linear projections of the matrix $\mathbf{Y}_{nxs}$ (with rows representing participants and columns representing the pre-adjusted and transformed phenotypes). The columns of the matrix $\mathbf{U}_{sxq}$ contain the loadings of each phenotype onto the columns of $\mathbf{Q}$ (i.e., values representing the contribution of each phenotypic variable to the construction of each factor), and $\mathbf{\Sigma}_{qxq} = \boldsymbol{diag}\{\sigma_1 \cdots \sigma_q\}$, where σ_k is the k-th singular value, such as a $\sigma_1 > \cdots > \sigma_q > 0$. The the matrix $\mathbf{e}_{nxq}$ contains the projection errors that depended on the selected number of latent factors (the value of q). The second term at the right-hand side of [Eq.4-2] is an Elastic Net (EN) penalty on the elements of $\mathbf{U}$. The EN penalty balances well-established techniques to select variables (zeroing out the noise and redundant signal between omic features) and shrinkage (to account for the high number of omic features that often exceed the number of samples). The expressions $||.||_2$ and $||.||_1$ correspond to the L2 and L1 norms, respectively. The parameter λ_U is a real positive number controlling the amount of sparsity in elements of $\mathbf{U}$, while α_U is any number between zero and one. The value of α_U balances shrinking and variable selection [68].

Heuristic methods typically select the value of q for which the trajectory of $\sigma_1 \cdots \sigma_q$ explains a relatively large amount of variance, or at which the trajectory of $\sigma_1 \cdots \sigma_q$ bends drastically (e.g., elbow rule). We complemented these heuristics by fitting the model [Eq.4-1] 100 thousand times via bootstrap resampling to avoid distributional assumptions. We then used the bootstrap distribution of $\sigma_1 \cdots \sigma_q$ to estimate 95 % confidence intervals for q. Once the value of *q* was estimated, the remaining missing data after quality controls was imputed following [191], as implemented in the function *imputeFAMD* from R package *missMDA* [192]. The values of the hyperparameter λ_U was estimated using the heuristic

proposed in [190], whilst the value of α_U was set to 0.5. The bootstrap distribution for [Eq.4-1] also yield 95% confidence intervals for the sparse solution of $\mathbf{U}$ to determine what phenotyeps exactly contributed to the formation of each PEB.

Association studies: We conducted two types of genome-wide associations studies (GWAS) in two types of variables. The first one involved the PEB variables (columns of matrix $\mathbf{Q}$) as responses, while the second used transformed and covariates-adjusted original phenotypes (columns of $\mathbf{Y}$). Since normality was not assured, we applied the Rank-Based Inverse Normal Transformation (RIN) to both for $\mathbf{Y}$ and $\mathbf{Q}$. Considering the columns of the matrix $\mathbf{X}_{nxp}$ represent the genotypes of p SNPs across the n participants; the following models were adjusted, one marker and variable at a time:

$$f(\mathbf{Q}_k) = \mathbf{x}_j \gamma_{jk} + \boldsymbol{\epsilon}_{kj} \quad \textbf{[Eq.4-3]}$$

and

$$f(\mathbf{Y}_r) = \mathbf{x}_j \beta_{jr} + \mathbf{e}_{rj} \quad \textbf{[Eq.4-4]}$$

Where f represent the RIN transformation, $\mathbf{x}_j$ is the j -th column of $\mathbf{X}$ representing the j -th SNP (j =1, …, p), γ_{jk} and β_{jr} are the effects of the j -th SNP on the k-th PEB and the r-th phenotype, respectively; and $\boldsymbol{\epsilon}_{kj}$ and $\mathbf{e}_{rj}$ are vectors of models residuals. We called models [Eq.4-3] and [Eq.4-4] **PEB-GWAS** and **ORIG-GWAS**, respectively. Estimates of γ_{jk} and β_{jr} Moreover, p-values were obtained via the omnibus RNI omnibus test implemented in the R Package *RNOmni* [193]. In addition to the PEB-GWAS and ORIG-GWAS analyses, we studied the pleiotropic effect of each SNP on the original phenotypes contributing to each PEB, that is, those variables with loadings different from zero after imposing sparsity on the element of $\mathbf{U}$. For this, we used the fast sequential test of pleiotropy proposed in [194]. We called the analysis **PEB-PLEIO**.

Summary of GWAS results: A standard threshold of $1x10^{-8}$ was used to define an SNP as significant. GWAS results were also summarized in "peaks", defined based on SNP's p-values and linkage disequilibrium (LD) among them. LD decay was calculated as the coefficient of determination R^2 between significant and adjacent SNP within half megabase. GWAS peaks were defined imposing by imposing a threshold of $R^2 \geq 0.01$. The SNP with the lowest p-value within a peak was chosen as "lead". Peaks were annotated using functional information from the G37 genome assembly annotation provided by the UKB project and complemented by ENCODE annotation obtained via Bioconductor package *biomaRt* [195,196]. Additionally, R package *LDlink* [197] was used to retrieve information of the overlap between peaks and significant expression-QTL (eQTL) from the Genotype-Tissue Expression (GTEx; [198]). Enrichr [202] and Ingenuity Pathway Analysis (IPA; QIAGEN Inc) were used to determine the overlap between genes and pre-existing gene sets.

Cluster analysis: To determine if the PEB would induce a separation of participants in biologically meaningful groups, we first embedded the columns of $\mathbf{Q}$ onto two dimensions using Uniform Manifold Approximation and Projection (UMAP; [199])**.** UMAP is a non-linear embedding technique suitable for large data sets, producing convenient two-dimensional representations of clusters existing in higher dimensions. UMAP creates a graph in multiple dimensions and projects it onto a lower number of dimensions attempting to conserver its structure. For this, the number of neighbors and the minimum distance among them defining a local neighborhood must be tuned. Tuning was done for 5, 10, and 20 neighbors, and minimum distances of 0.1%, 1%, and 10% of the average Euclidean distances between rows of $\mathbf{Q}$. UMAP was fitted using the R package *umap*

[199]. The map rendering the best clusters separation was considered the optimal one. Cluster delineation was done following [190]. To determine if clusters had biological meaning, we fitted the following linear models:

$$f(\mathbf{Y}_r) = \mathbf{Z}\boldsymbol{\tau}_r + \boldsymbol{\zeta}_r$$

where $\mathbf{Z}$ is an indicator matrix representing the membership of a participant to a specific cluster, $\boldsymbol{\tau}_r$ is the vector of effects of cluster membership on $f(\mathbf{Y}_r)$ such as $\sum_{c=1}^{\#Clusters} \tau_{cr} = 0$, and $\boldsymbol{\zeta}_r$ is the vector of model residuals. Normality of $\boldsymbol{\zeta}_r$ was assumed by employing transformation f. Analysis of variance (ANOVA) was used to test for significant differences among clusters. Tukey test was used to test for significant pairwise differences between clusters means, using a significant level of $\frac{0.05}{\binom{\#Clusters}{2}}$ [200].

4.4. RESULTS

The primary purpose of this study was to identify genetic elements influencing energy balance. For this, we integrated variables potentially affecting the body's energy requirements, including physical activity, diet, body size and composition, and blood metabolites. Hence, we analyzed these data layers with a sparse latent factor. The derived factors exhibited commonalities between the layers. Finally, we studied genetic polymorphisms associated with these latent factors. Figure 4.1 shows the processing steps and resulting sample size of this study.

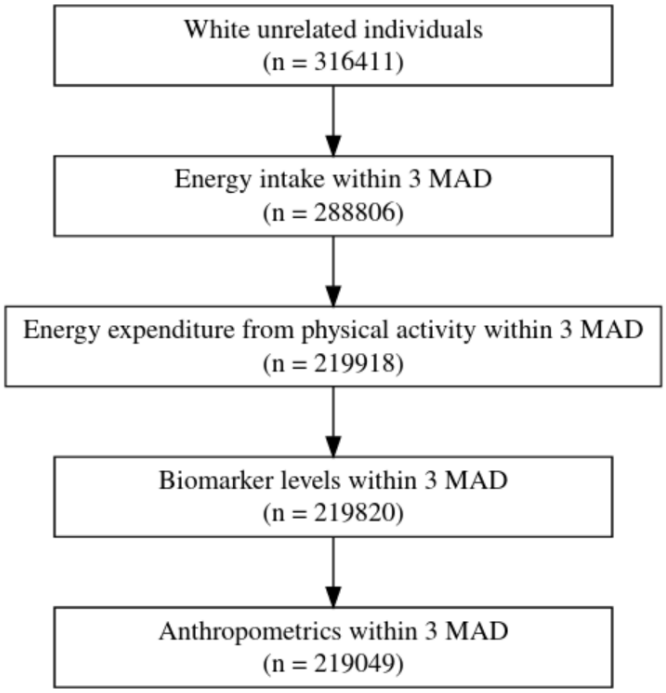

Figure 4.1: Inclusion criteria and sample size. Sample size (**n**) after each of the inclusion criteria. **MAD** = median absolute deviation.

4.4.1. PEB variables were associated with specific groups of phenotypes.

Summary statistics by phenotypic variables are presented in Table 4.1. In the following description of results, we will refer to these as "original" variables to distinguish them from

the PEB variables (i.e., the derived or latent variables). The variables considered were food consumption patterns from the touchscreen questionnaire, blood biomarkers (glucose, triglycerides, cholesterol, and creatinine blood levels), body size (weight, height, waist circumference), body composition (fat mass and lean mass), and physical activity measuring intensity (intensity of walking, moderate, and vigorous exercise) and periodicity of exercise (frequency and duration of exercise). Diet variables included low-caloric (water, vegetables, fresh fruit intakes), fish and meats (beef, pork, lamb, poultry, oily, and non-oily fish), low processed foods (coffee and tea intakes), moderately processed foods (bread and alcohol intakes), and highly processed foods (cereal, cheese, and processed meats intakes).

Table 4.1: Descriptive statistics of phenotypical variables. Numerical variables are summarized based on the median and median absolute deviation (**MAD**) on their original scales (***units***). Categorical variables are summarized by the percentage of samples within each category (**%**). For all categorical variables, except alcohol intake, the following coding was used to represent intake frequency: **0**= Never, **1**=Less that once a week, **2**=Once a week, **3**=Two or three times a week, **4**=Five or six times a week, **5**=Once or more daily. For alcohol, the following code was used: **1**=Daily or almost daily, **2**=Three or four times a week, **3**=Once or twice a week, **4**=One to three times a month, **5**=Special occasions only, **6**=Never. Missing data for all variables are expressed as a percentage of the total sample size. **PA**: physical activity.

Numerical variables	**Median (MAD)**	**Missing**
Glucose (mmol/L)	4.9 (0.49)	18.0
Cholesterol (mmol/L)	5.8 (1.1)	5.5
Creatinine (umol/L)	70 (14)	6.0
Triglycerides (mmol/L)	1.5 (0.72)	9.6
Weight (Kg)	77 (16)	0.4
Height (cm)	170 (10)	0.3
Waist circumference (cm)	91 (13)	0.2
Fat mass (Kg)	51 (13)	1.9
Lean mass (Kg)	24 (8.5)	2.1
Cooked vegetable	2 (1.5)	4.9
Salad/raw vegetable	2 (1.5)	8.5
Fresh fruit (pieces/day)	2 (1.5)	4.6
Bread (slices/week)	10 (5.9)	9.5
Cereal (bowls/week)	5 (3)	5.0
Tea (cups/day)	3 (3)	4.5
Coffee (cups/day)	2 (1.5)	10.0
Water (glasses/day)	2 (1.5)	12.0
Intensity of moderate PA	240 (360)	28.0
Intensity of vigorous PA	0 (0)	26.0
Intensity of walking	690 (680)	25.0
Intensity of all PA	1200 (1200)	26.0
Frequency of PA (days)	9 (4.4)	22.0
Duration of PA (min)	80 (67)	25.0
Categorical variables	**Category: %**	**Missing**
Oily fish (serv)	**0**:11, **1**:34, **2**:38, **3**:17	6.0
Non-oily fish (serv)	**0**:04, **1**:29, **2**:50, **3**:17	4.0
Processed meat (serv)	**0**:08, **1**:30, **2**:30, **3**:28, **4**:04	17.0
Poultry (serv)	**0**:04, **1**:11, **2**:37, **3**:46, **4**:02	21.0
Beef (serv)	**0**:09, **1**:46, **2**:33, **3**:12	4.2
Lamb mutton (serv)	**0**:17, **1**:57, **2**:22, **3**:04	6.6
Pork (serv)	**0**:15, **1**:59, **2**:23, **3**:03	6.2
Cheese (serv)	**0**:03, **1**:16, **2**:21, **3**:45, **4**:09, **5**:06	2.4
Alcohol (g)	**1**:21, **2**:23, **3**:26, **4**:11, **5**:11, **6**:08	1.0

PEB variables were derived from the first ten latent dimensions from the sparse FAMD. Since there was not a precise inflection point in the singular values to define a cutoff (see Figure C.1-A), we based the selection of latent variables on the threshold at which the rate of change in singular values, and therefore the size of phenotypic variances explained by each PEB, reached a plateau (Figure C.1 panel C and D). This value represented 20% of the total interindividual phenotypic variability (Fig C.1, panel B). Although differing at being derived from a smaller group of phenotypes, these PEB latent variables were highly correlated with their *dense* counterpart (i.e., latent factors from a FAMD without sparsity) and explained a similar proportion of variance (Figure C.2).

The latent variables were constructed, imposing sparsity on the loadings of the FAMD. Thus, values near zero were excluded, limiting the possible number of phenotypes contributing to each PEB. The relationship between PEB and original phenotypes was studied using non-zero FAMD loadings and marginal correlations (Figure 4.2). Since only the first five PEB variables were related to phenotypes representing multiple components of EB, we retained them for further analysis. We indexed these PEB variables from one to five, with PEB 1 and 5 representing the highest and lowest amounts of variance explained, respectively.

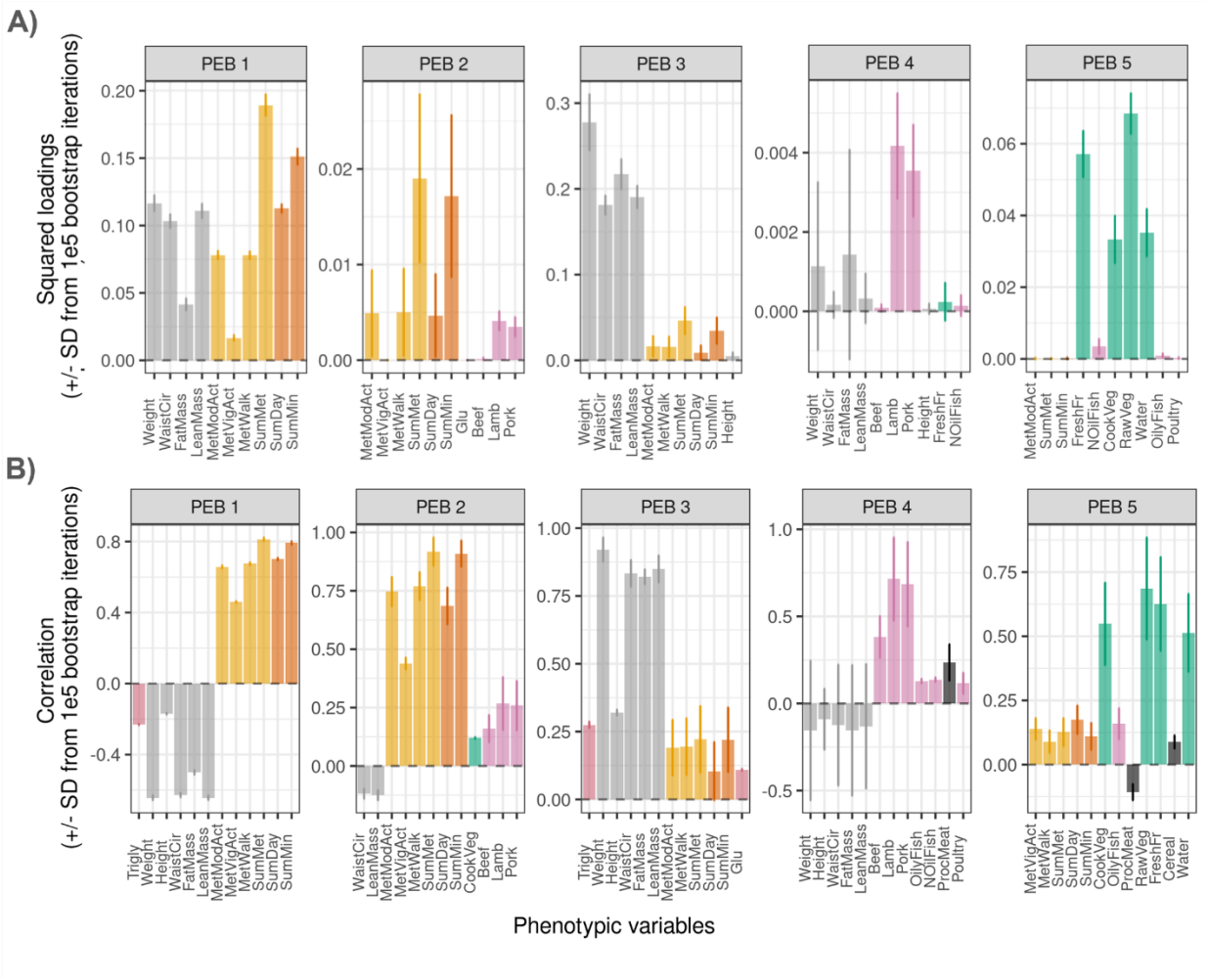

Figure 4.2: Relationship between PEB variables and original phenotypes. A) Squared phenotype loadings. The contribution of each of the original phenotypes was obtained from bootstrap repetitions of the sparse factor analysis. Each panel shows the original phenotypes that contribute to each PEB variable. **B) Marginal correlation between PEB and phenotypes.** Similarly, the Bootstrap correlations and CI are shown for each pair of PEB and phenotypes.

Higher values of PEB 1 represented active individuals of small body sizes and low blood triglycerides. The opposite was true for lower values of PEB 1. Lower PEB 1 values represented sedentary individuals of larger body sizes and higher values of blood triglyceride. Similarly, higher PEB 2 represented subjects with small body sizes and active diets, including red meats and vegetables. Higher PEB 3 represented subjects of large

body size, physically active, and with high levels of blood triglycerides and blood glucose. Higher values of PEB 4 represented individuals of average body size, predominantly meat-eaters. Finally, higher values of PEB 5 represented active individuals largely vegetarian.

Table 4.2: Labels for PEB variables based on correlation with original phenotypes.

PEB	Label	Correlation with original phenotypes	
		Positive	**Negative**
PEB 1	*Active, small body size, low blood triglycerides*	Physical activity (PA)	Body size and mass// Blood triglycerides
PEB 2	*Active, small body size, meat, and veggies intake*	PA // meat and vegetable consumption	Waist circumference, lean body mass
PEB 3	*Active, large body size, high blood triglycerides, and glucose*	Bodyweight and mass, waist circumference// PA//Blood triglycerides and glucose	--
PEB 4	*Average body size, meat intake*	Meat consumption	--
PEB 5	*Active, largely vegetarian*	Vegetables, fruit, water, cereal, and oily fish consumption // PA.	Processed meat consumption

4.4.2. PEB-induced aggrupation of phenotypically distinct participants.

The PEB variables can also induce a separation of individuals into phenotypically distinct groups. Figure 4.3 shows clusters of participants (Figure 4.3-A) and the association between clusters and original phenotypic variables (Figure 4.3-B). This representation was done by embedding the latent PEB variables to two dimensions using Uniform

Manifold Approximation and Projection (UMAP)[199]. There were seven distinct clusters of participants. The cluster with the highest degree of separation from the rest was Cluster 1. This cluster was dominated by participants exhibiting low body fat, blood biomarkers, meat consumption, poultry, and processed foods. Participants within this cluster also had high levels of physical activity and consumption of low caloric foods. Contrastingly, Cluster 5 was characterized by higher body fat levels, blood biomarkers, and alcohol, while Cluster 4 was characterized by higher consumption of meats and processed foods and relatively low physical activity. The remaining clusters had fewer striking differences. Cluster 2 followed a pattern like Cluster 1, but less marked, particularly having less consumption of low-calorie food. Cluster 3 followed Cluster 1 and 2, but even less markedly, characterized by more processed food consumption. Finally, clusters 6 and 7 represented average values across all contrasts, differing only at the consumption of alcohol, processed food (higher in Cluster 6), and low-calorie food (higher in Cluster 7).

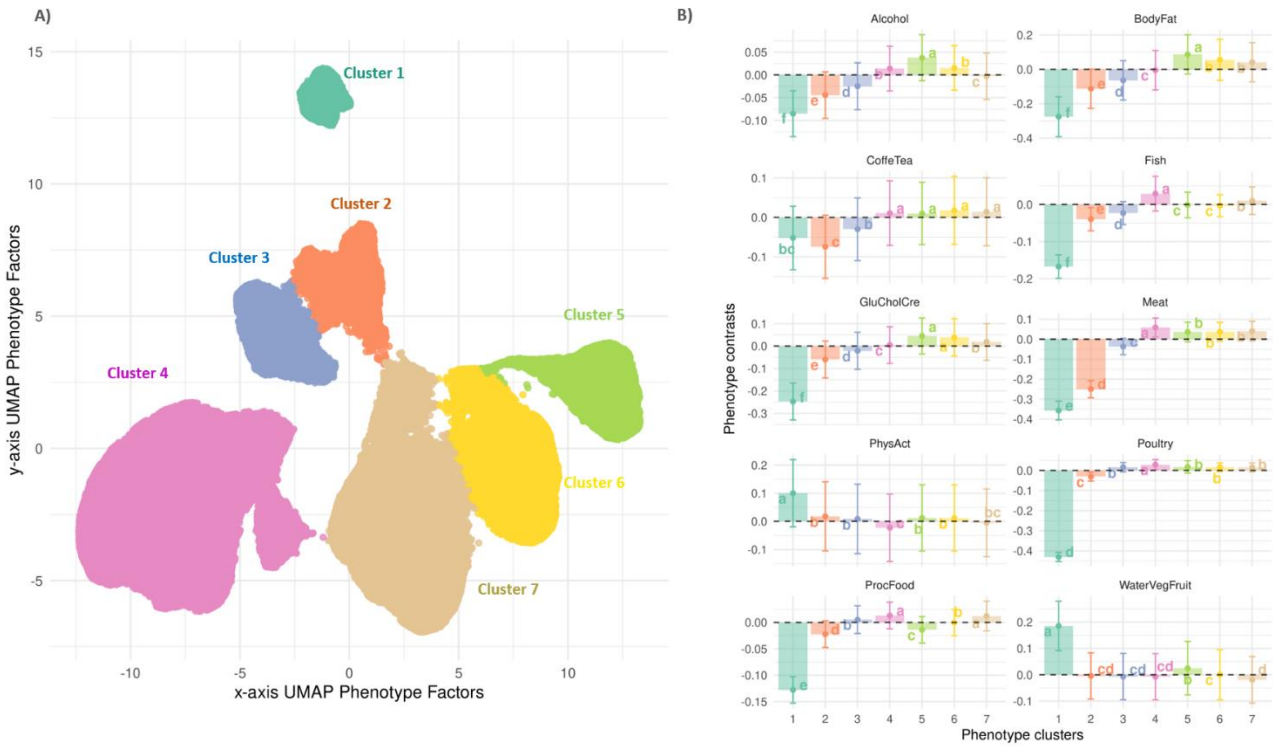

Figure 4.3: PEB-based cluster analysis. A) UMAP projection of PEB variables. A two-dimensional UMAP projection of PEB was obtained from Factor Analysis of Mixed Data on the original set of phenotypic variables involved in energy balance and resulted in seven clusters**. B) Contrasts between phenotypes and clusters**. Each panel represents a statistical contrast between phenotypes and clusters. The colored rectangles and points represent the means of each contrast. The vertical bars represent plus and minus one standard deviation of each contrast. **Abbreviations**: *Alcohol* (consumption of alcohol), *BodyFat* (body fat, waist circumference, weight, and blood triglycerides), *CoffeTea* (consumption of tea or coffee), *fish* (consumption of oily or non-oily fish), *GluChoCre* (level of glucose, cholesterol, or creatinine in the blood), *Meat* (consumption of meat), *PhysAct* (levels of physical activity), *poultry* (consumption of poultry), *ProcFood* (consumption of processed food), and *WaterVegFruit* (consumption of water, vegetable, or fruit).

4.4.3. Genomic variants associated with PEB.

We conducted three analyses to determine the relationship between PEB variables and the original phenotypes with genetic markers. Firstly, we conducted separated GWAS on the PEB variables (PEB-GWAS). Secondly, we studied the overlap between PEB-GWAS results and separated GWAS on the original variables (ORIG-GWAS). Lastly, we used

the information from PEB-GWAS to study pleiotropic effects on the groups of phenotypes contributing to each PEB (PEB-PLEIO).

The PEB-GWAS resulted in significant associations for all PEB variables. The significant hits (using a threshold of $5x10^{-8}$) were 2113, 116, 12348, 324, and 354. These significant hits were further grouped into 50, 3, 185, 4, and 67 peaks defined by markers in linkage disequilibrium. Annotated Manhattan plots for PEB-GWAS analyses are presented in Figures C.3-7.

Most of the PEB-GWAS peaks (15, 3, 60, 3, 3, and 7) were not present on ORIG-GWAS, demonstrating that combining correlated phenotypes increased the power to find these regions (Figure 4.4). Figure 4.4 shows the annotation information for PEB-GWAS, ORIG-GWAS, and PEB-PLEIO analyses, emphasizing the overlap with genic regions. Identified genes were enriched for five different gene sets: obesity, metabolic disease, cardiovascular disease, connective tissue development and function, and carbohydrates metabolisms (FDR p-value < 0.01) (Figure 4.4, see Materials and Methods).

The gene associated with the highest number of PEB latent variables (PEB 1, PEB 3, and PEB 5) was the Fat Mass and Obesity-Associated Protein (*FTO*) (Figure 4.4). This peak for *FTO* was present in the ORIG-GWAS for weight, height, waist circumference, lean and fat mass. In addition, *FTO* had three leading SNPs (defined here as the SNP with the lowest p-value within an LD block): rs56094641, rs1421085, and rs11642015, respectively. All these SNPs were present in the PEB-PLEIO for weight and body mass.

A group of genes mapped for both PEB 1 "Active, small body size, low blood triglycerides" and PEB 3 "Active, large body size, high blood triglycerides, and glucose". This group consisted of Protein lin-7 Homolog C (*LIN7C*), Centrosomal Protein POC5 (*POC5*),

INO80 complex subunit E (*INO80E*), Protein Lifeguard 2 (*FAIM2*), and G1/S-specific Cyclin-D2 (*CCND2*). Gene *LIN7C* had a peak present in the ORIG-GWAS of all body measurement variables. The leading PEB-GWAS SNPs at this peak were rs962369 and rs2049045, respectively. These peaks were also present ORIG-GWAS and PEB-PLEIO for all body measurement variables. The same peak was also present in the individual ORIG-GWAS for the intensity of moderate physical activity. Gene *POC5* also had a peak (leading PEB-GWAS SNPs rs2307111 and rs67570751, respectively) present in the ORIG-GWAS for all body measurement variables. Gene *INO80E* had a peak present in all body measurements ORIG-GWAS, except for height. The leading PEB-GWAS SNPs in this peak were rs7190185 and rs35105141, respectively. This peak was present in the PEB-PLEIO of weight, waist circumference, and body mass. Gene *CCND2* had a peak with rs76895963 as the leading SNP. This peak was present in the ORIG-GWAS of weight, fat mass, and lamb consumption. Also, the peak was present in the PEB-PLEIO for weight and fat mass.

Several genes had PEB-GWAS peaks mapping exclusively for "Active, large body size, high blood triglycerides, and glucose". We focus the description on peaks that were present in the ORIG-GWAS and PEB-PLEIO analyses. Golgin A6 Family Like 5 - Pseudogene (*GOLGA6L5P*) had a PEB-GWAS peak with rs11259919 as lead SNP. This peak was also present in the PEB-PLEIO of body measurement variables, lamb, and pork consumption. The peak was also present in the ORIG-GWAS of weight, fat mass, and waist circumference but absent for height, lamb, and pork consumption. The peak at Leiomodin 1 (*LMOD1*, leading SNP rs2820322) followed the same behavior as *GOLGA6L5P*, except that lamb consumption was not part of the PEB-PLEIO results. The

peaks at Glutathione S-Transferase Mu 4 (*GSTM4*, leading SNP rs7550711) and Vacuolar Protein Sorting-Associated Protein 29 (*VPS29*, with a non-reference leading SNP at chromosome 12 and 110852370 bp) were both presents in the ORIG-GWAS of weight, height, and fat mass. This peak was present in the PEB-PLEIO of weight, height, fat mass, and intensity of moderate physical activity, while the lead SNP at *VPS29* was present in the PEB-PLEIO of weight, height, fat mass, and physical activity from walking. The peak at Hedgehog Interacting Protein (*HHIP*, leading SNP rs4240326) was present in the ORIG-GWAS of weight, waist circumference, body mass, and lamb consumption. This peak was also present in the PEB-PLEIO of weight, waist circumference, and body mass.

Some genes had peaks collectively present in the ORIG-GWAS and PEB-PLEIO of all body measurements and variables, including Hydroxycarboxylic Acid Receptor 1 (*HCAR1*, leading SNP rs7133768), BTB/POZ Domain-Containing Protein 7 (*BTBD7*, leading SNP rs4381522), Hydroxysteroid 17-Beta Dehydrogenase 12 (*HSD17B12*, leading SNP rs1061810), SCM Polycomb Group Protein Homolog 1 (*SCMH1*, lead SNP rs61780439), Family 5 Member 11 (*SLC5A11*, leading SNP rs12923476), Regulatory Associated Protein of MTOR Complex 1 (*RPTOR*, leading SNP rs11150745), and tRNA Splicing Endonuclease Subunit 15 (*TSEN15*, leading SNP rs74767794). A peak at SUMO2 Pseudogene 17 (*SUMO2P17*, leading SNP rs8082345) was exclusively present in the ORIG-GWAS and PEB-PLEIO of weight, waist circumference, and fat mass. Peaks at Axis Inhibition Protein 2 (*AXIN2*, leading SNP rs757558), Centrosomal Protein 120 (*CEP120*, leading SNP rs34732995), and p53 Tumor Suppressor Protein (*TP53*, leading SNP rs78378222) were present in the individual ORIG-GWAS and PEB-PLEIO of weight,

height, and fat mass. A peak at Protein Disulfide Isomerase Family A Member 2 (*PDIA2*, lead SNP rs12926311) was present in ORIG-GWAS of weight, height, and fat mass, and PEB-PLEIO of weight and fat mass. A peak at RAD9 Checkpoint Clamp Component A (*RAD9A*, leading SNP rs34560402) was present in ORIG-GWAS of waist circumference. Finally, peaks at Mitogen-Activated Protein Kinase 5 (*MAP2K5*, leading SNP rs12050481) and Potassium Voltage-Gated Channel Subfamily E Regulatory Subunit 4 (*KCNE4*, leading SNP rs1607246) were present in the ORIG-GWAS of height. Finally, a group of genes had peaks exclusively for PEB-GWAS of "Active, small body size, low blood triglycerides". Peaks at Cytochrome P450 Family 3 Subfamily A Member 5 (*CYP3A5* lead SNP rs13311457) and UHRF1 Binding Protein 1 (*UHRF1BP1* lead SNP rs2744977) were present for the PEB-GWAS and PEB-PLEIO of weight, waist circumference, and fat mass. Peaks at SUMO1 Activating Enzyme Subunit 1 (*SAE1* lead SNP rs3810291) and Neurexin 2 (*NRXN2* lead SNP rs12273892) were present in PEB-GWAS and PEB-PLEIO of weight, waist circumference, and lean mass. Three peaks were present at putative-uncharacterized proteins. A peak at *RP11-866E20.3* (lead SNP rs12967135) was present for PEB-PLEIO of waist circumference, the intensity of moderate physical activity and walking, and ORIG-GWAS for the last two variables only. A peak at *RP11-251G23.2* (lead SNP rs73190105) was present in ORIG-GWAS of weight, fat mass, and intensity of vigorous activity and present in PEB-PLEIO of the first two variables. A peak at *RP4-635E18.7* (lead SNP rs5019466) was present in the ORIG-GWAS of weight, waist circumference, fat mass, and PEB-PLEIO of weight and waist circumference. Another gene was Zinc Finger and BTB Domain Containing Protein 38 (*ZBTB38*). Gene *ZBTB38*

had a significant peak in PEB-GWAS for "Active, small body size, low blood triglycerides" (lead SNP rs6785012).

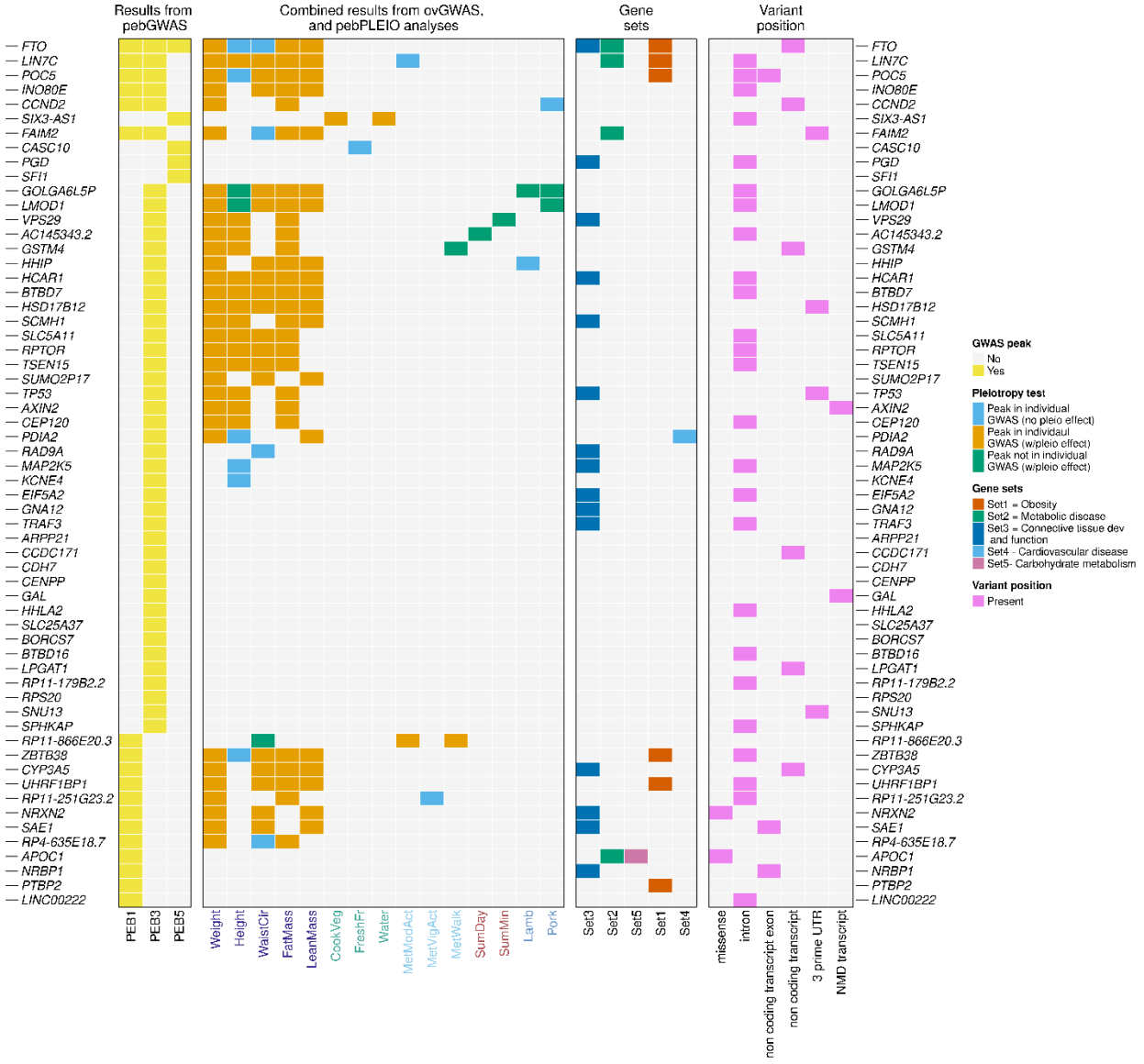

Figure 4.4: Summary of GWAS loci within genic regions. The panels above summarize the overlap between linkage disequilibrium blocks containing significant SNP from GWAS between PEB latent variable and genic regions. The panels represent, in order, the overlap between significant peaks from PEB-GWAS and genic region, overlap between ORIG-GWAS and PEB-PLEIO analyses, and overlap with gene sets and regulatory regions.

As illustrated in Figure 4.4, pleiotropy was most abundant for body size and composition variables, followed by components of PA and dietary intake variables. Regarding enrichment results, most genes with significant peaks were involved in connective tissue development and function, followed by obesity and metabolic disease. The most significant peaks were also located in intronic regions of the gene.

4.5. DISCUSSION

Excessive calorie intake and low physical activity generate a positive EB that results in obesity. We integrated several phenotypes from the UK Biobank related to EB. We used this information to get novel insights into the physiological reasons that lead to a positive EB. Sparse factor analysis of mixed data (sFAMD) and bootstrap inference was used to define patterns of energy balance (PEB) and to determine associations with common SNP. From these associations, a set of genes with possible links to EB patterns were identified.

Previous studies have applied similar phenotypes-integration techniques to determine individuals' diet or physical activity patterns [160–163]. In contrast, our study focused on all aspects of EB, solving the challenges of including random variables of different phenotypic scales (e.g., numerical biomarkers values and body measurements versus categorical food frequencies) by combining conventional FAMD [189] with sparse SVD [201]. The application of sparse FAMD (sFAMD), followed by bootstrap, clearly determined what phenotypes contributed to each PEB. This analysis generated PEB as factors that capture data characteristics beyond the ones expected by naïve integration (e.g., averaging subjectively defined groups of phenotypes) and effectively captured different aspects of EB.

Besides capturing variability due to EB's multiple components, PEBs were also associated with variability at the genomic level. For example, some PEB-GWAS peaks mapped on genes previously reported affecting obesity and metabolic diseases, such as *FTO*, *ZBTB38*, and *POC5*. For instance, the SNP rs56094641at *FTO* had effects on PEB 1 "Active, small body size, low blood triglycerides", PEB 3 "Active, large body size, high

blood triglycerides and glucose" and PEB 5 "Active, largely vegetarian". This SNP has also been previously reported as a cis-QTL in muscle (Fig C.8), as well as being present in GWAS of childhood obesity [202], body fat [203], and metabolic disease [204]. Additionally, the lead SNP rs6785012 at *ZBTB38*, mapping on PEB 1, has been associated with eczema and red blood count [205] and has also been reported as cis-eQTL in many tissues, including testis, adipose tissue, whole blood, spleen, pituitary, and thyroid glands (Figure C.8). Moreover, rs6785012 had pleiotropic effects on body measurements found here agree with a previously reported association of *ZBTB38* with BMI and waist circumference [206]. Lastly, the SNP rs2307111 at *POC5*, mapping on PEB 1 and 3, had pleiotropic effects on body measurement variables. This finding can be supported by the presence of this SNP in previous GWAS studies on body measurements [205,207]. However, previously reported associations between this SNP and cholesterol were not confirmed in PLEIO-GWAS or OIRG-GWAS analyses.

Some genes had pleiotropic effects on both body measurements and physical activity variables. For example, a non-synonym SNP at chromosome 12 and 110852370 bp mapping on *VPS29* was associated with height, weight, fat mass, and duration of physical activity. Variations in *VPS29* and other genes in the retrosome (part of the retrograde transport from the endosome to the Golgi apparatus) are typically linked with neurodegenerative disorders [208]. However, in mice models, variation in the retrosome components has been indirectly linked with normal growth and mammal development and muscle response to exercise and training through the Wnt and *β*-Catenin pathway [209,210]. The potential role of the variability of this SNP on muscle growth and response to exercise can be explained by the significant association with the expression of *VPS29*

in muscle (GTeX data presented in Figure C.8). Similarly, a peak on *GSTM4* with rs7550711 as lead SNP had a pleiotropic effect on weight, height, fat mass, and walking intensity. This SNP has been previously associated with BMI in physically active adults [206] and phospholipid fatty acids in plasma [211].

A group of PEB-GWAS peaks was located in genes involved in carbohydrate metabolism and cell cycle, a general category with known effects on body composition [212] and physical activity performance [213]. One of these genes, *SIX3,* had a PEB-GWAS peak for PEB5 "Active, healthy diet" variable, with rs4953152 as lead SNP. This SNP had pleiotropic effects on cooked vegetables and water consumption. These associations have not been previously registered for either cooked vegetables or water consumption. However, evidence of an association between variations in rs4953152 and cognitive processes has been observed [214,215]. In addition, rs4953152 has been reported as a cis-eQTL of *SIX3-As1* in the brain (Figure C.8). Lastly, there was a peak at *TP53* (lead SNP rs78378222), with pleiotropic effects on weight, height, fat mass. Variations in *TP53* are primarily linked to multiple cancer types [122].

Nevertheless, rs78378222 has also been reported as associated with fat mass [216,217] and cis-eQTL of TP53 adipose tissue (Figure C.8). Another SNP with pleiotropic effects on different groups of variables was rs11259919 at *LMOD1*. This SNP was associated with weight, fat mass, and consumption of pork. Variations in rs11259919 have been linked with the expression of *LMOD1* in many tissues, including the digestive tract, nerve, muscle, brain, adrenal gland, thyroid, artery, and heart (Figure C.8) well as with appendicular fat [217]. Although a previous link with food consumption has not been reported, variations in *LMOD1* expression in skeletal muscle have been linked with

differential susceptibility to cardiovascular disease in response to high-fat diets [218] and body composition [219].

Many of the putative QTL detected in this study have previous associations with phenotypes contributing to energy balance. For a relatively small number of phenotypes, like the ones used here, studying all possible combinations of pleiotropic effects could have been a possibility (i.e., studying all possible pleiotropic groups among variables). However, at the scale of the UK Biobank, this renders computationally prohibitive. For that reason, the alternative of sparse FMD used here offered a way to *a)* inform what groups of variables are most likely to collaborate to form the PEB variables and *b)* inform the study of pleiotropic effects in a more manageable number of phenotypes. Although most SNP had relatively minor effects on each PEB, the large sample size of this study was instrumental in detecting them.

One limitation of this study is the lack of functional validation, especially for novel genes associated with EB. Although we have used data from GTeX to confirm the association between the change of allele and gene expression, an association between gene expression and phenotypes related to EB was not confirmed (e.g., *VPS29* variant causing differential fat mobilization within the adipocyte).

The results generated here contribute to understanding the complex biology of energy balance and the interrelation between its related traits [220,221].

CHAPTER 5

CONCLUSIONS

Increasing sizes and data density become a constant computational and statistical challenge for existing data integration methods [56,66]. To face these challenges, we introduced a new method of data integration, Multi-Omic Integration via Sparse Singular Value Decomposition (MOSS). MOSS exploits the benefits of large data sizes (i.e., many samples to increase power and many features to discover biologically relevant signals) while maximizing computational performance. We have written MOSS as an R package that can be freely available Comprehensive R Archive Network (CRAN). We review the capabilities and limitations of the MOSS package in chapter 2. However, three caveats of MOSS remain to be discussed, as they arise depending on the application.

The first caveat involved convergency properties. MOSS relies on the NIPALS algorithm to extract SVD solutions, and therefore, convergence in supervised problems (e.g., PLS) is not always assured [222]. Although empirical results suggest that convergence is reached in most practical situations [222], future research on the analytical properties of MOSS should include a thorough analysis of its convergency properties.

The second caveat involves the lack of statistical inference for the results of the features selection process (i.e., basing feature selection in, for example, confidence intervals). By considering the elements of SVD as random variables, both factors and loadings can be thought of as drawn from probabilistic distributions (examples of this are [223] [224] and

[225]). Future work in MOSS could benefit from considering the data integration model as a probabilistic one to evaluate the significance of its solutions.

Finally, the third caveat comes from using elastic net (EN) penalty as a feature selection process. Fundamentally, EN zeroes out the noisy features while retaining correlated "signal" [68]. The intention is to find an optimal set of features representing biologically relevant groups (e.g., genes in pathways, food items in dietary patterns). Unfortunately, the performance of EN strongly depends on the tuning of the hyperparameter "α" [111]. In its current version, MOSS uses the fast heuristics in [69] to tune the degree of sparsity, but not α. Future versions of MOSS would require an alternative to tune all hyperparameters without compromising computational efficiency.

Cancer research is one of the primary disciplines where the performance of omic integration tools will continue to be challenged. Thanks to progress in sequencing techniques, cancer genomics has advanced at an extraordinary pace. This pace has been evident ever since the early days of cancer genomics, where microarray experiments were rapidly complemented by new generation sequencing techniques in less than a decade [226]. These techniques have been essential at creating several large international repositories of cancer multi-omic data, like The Cancer Genome Atlas (TCGA) [95], the International Cancer Genome Consortium (ICGC) [227], and the Cancer Cell Line Encyclopedia (CCLE) [228]. Cancer research has benefited from these data in many applications, most notoriously the enhancement of risk predictions models [14–16] and improving tumor classification with molecular subtypes [72,102,108].

Chapter 3 has also benefited from these repositories. We have used data from TCGA consisting of approximately six thousand tumor samples and sixty thousand features

representing genome-wide gene expression, copy number variants, and DNA methylation values. In chapter 3, we used MOSS to detect shared molecular features acting across clusters of tumors. These clusters formed beyond the restrictions of the site of origin and exhibited similar clinical and immunologic characteristics, supporting the role of common molecular signatures across cancer types [88,229].

Despite these exciting findings, future applications of our method in cancer data would require a more robust estimation of tissue effects, for example, by using markers of tumor histology (e.g., mesenchymal, epithelial). In addition, newer classification efforts must emphasize validation avenues, such as knock-out and gene-drug interaction models.

Another area of fast-growing pace is phenomics. Once deemed prohibitively expensive and time-consuming, gathering phenomic data is now a reality [22]. Advances in imaging techniques [25], mass spectrometry [230], and automated data loggers [26] in the last years have been instrumental in the creation of large-scale phenotyping projects. One of the most extensive ongoing efforts is the UK Biobank (UKB), with data across several phenomic layers and genotypic information for more than five hundred thousand individuals. Integration of UKB's phenomic and genomic data has been conducted for multiple complex traits, including neurodegenerative disorders [231], and cardiovascular disease [232], and dietary habits [233].

Similarly, we focused our chapter 4 on integrating the UKB phenome to infer variation associated with different aspects of energy balance (EB) (e.g., the tendency of lean body types to engage in regular physical and healthy diets). We studied the association between this variation and genetics. We found known genes involved with EB (e.g., *FTO*, *POC5*, *ZBTB38*, *INO80E*) and novel ones (e.g., *VPS29, SIX3, LMOD1*) not present in the

GWAS of separated phenotypes. This work is the first to integrate all EB components from a phenomic point of view to the extent of our knowledge.

Regardless of these compelling results, chapter 3 used only a small set of the phenomic data in the UKB. Although only information from a few blood metabolites is currently present in the UKB, ongoing efforts to produce detailed metabolic profiles for a large sample of individuals are on their way [234]. Moreover, further incorporation of metabolomic profiles in our set of phenotypes could profoundly impact our definition of EB factors and associated genes since metabolomics has been shown to efficiently complement and improve the assessment of dietary patterns [235].

Following, we propose some avenues for future research on omic integration.

We have stated that MOSS does not rely on distributional assumptions and how this limits the possibility of statistical inference. One possibility for incorporating inference within MOSS is the adoption of Bayesian methods. Bayesian alternatives to SVD [223,225] could be extended to incorporate FBM to deal with large data sets. Additionally, adopting Bayesian methods would allow the incorporation of different prior distributions on the features' loadings coming from different omic layers. The study of the choice of prior distribution on the overall performance could aid researchers in deciding what set of assumptions better describes their data.

Another possible line of research is the study of alternative forms of sparsity on the performance of omic integration. Omic integration might not be as robust in biological scenarios where specific molecular events are not as drastic as in cancer. Therefore, the evaluation of sparsity (e.g., on pre-defined groups of features) and their impact on

performance could help decide what methods are best for problems where the signal-to-noise ratio is low.

Alternatively, since cancer is dominated by extreme, sometimes widespread, molecular events, we foresee omic integration to continue being useful for cancer research. An exciting line of research can focus on evaluating methods' performance and the proposal of new algorithmic shortcuts to handle huge data sets (e.g., recently generated data from the Pan-Cancer Analysis of Whole Genomes) [236].

APPENDICES

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Table A.1: Code templates to replicate examples.

```
# Install and load MOSS.
install.packages("MOSS")

library("MOSS")
# Example of unsupervised omic integration with MOSS
out_sig <- moss_signatures(out_sim,out_moss$selected_items,
                    clus_lab = out_moss$clus_plot$dbscan.res$cluster,
                    only.candidates = TRUE,
                    plot=TRUE,
                    th=0.05)
# Example of supervised omic integration with MOSS
set.seed(345)
out_moss <- moss(data.blocks = out_sim,
              method = "pls",
              resp.block = 1,
              scale.arg = T,
              norm.arg = T,
              K.X=10,
              nu.v = seq(1,500,by=2),
              nu.u = seq(1,100,by=2),
              alpha.v = 0.5,
              axes.pos = 1:2,
              alpha.u=0.5,
              exact.dg = TRUE,
              use.fbm = TRUE,
              nu.parallel = TRUE,
              tSNE=list("perp"=30,
                    "n.iter"=1e3,
                    "n.samples"=1),
              cluster = list(eps_range=c(0,1),
                        eps_res=10,
                        min_clus_size=2),
              plot=TRUE,
              lib.thresh = TRUE)
```

Table A.1 (cont'd)

```
# Extract heatmap.
moss_heatmap(out_moss$B,
          SVD = out_moss$sparse,
           axes.pos = c(1,3),
           col.lab = "miRNA-seq & ",
           row.lab = "RNA-seq")
```

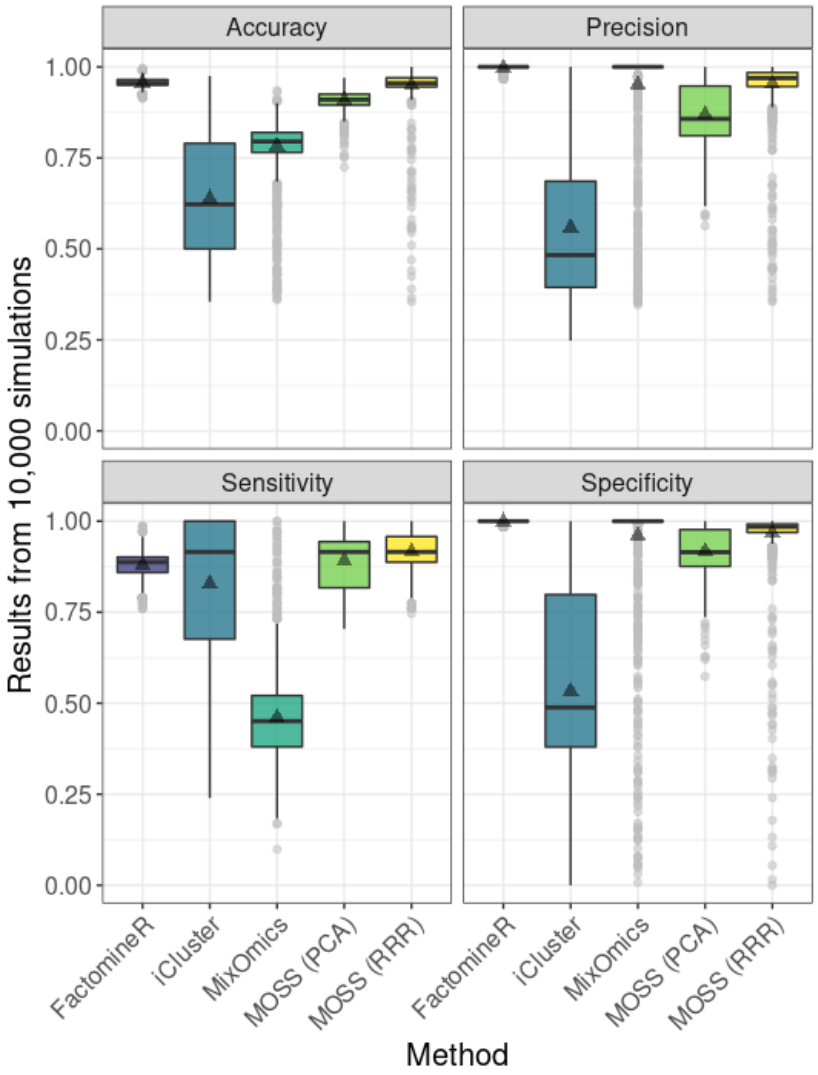

Figure A.1: Analytical performance of several omic integration methods. Each panel corresponds to the accuracy (**top-left**), precision (**top-right)**, sensitivity (**bottom-left**), and specificity (**bottom-right**) of each omic-integration method at detecting informative features (i.e., features with signal across and between omic blocks). Methods compare were **FactoMineR** (function *MFA*), **iCluster** (function *tune.iCluster2*), **mixOmics** (function *tune.splsda*) and **MOSS** (function *moss* with method="pca" and method="lrr", respectively). Results were obtained across 10,000 random simulations.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Table B.1: List of genes significantly deregulated in at least one pan-cancer cluster. All the genes significantly different in at least one cluster are sorted by chromosome and genomic position. ANOVA's p-values, adjusted for multiple comparisons, are displayed on the last column of the table (**p-value**).

	Average enrichment scores by cluster								
Gene	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**p-value**
C1orf159	0.02 b	0.10 ab	-0.07 b	0.15 ab	0.21 ab	0.41 a	0.00 b	-0.09 b	6.36E-13
WASH2P	-0.29 b	0.21 a	0.00 b	0.05 ab	0.17 ab	-0.16 b	-0.04 b	0.05 ab	9.78E-14
RAB6C	0.08 ab	-0.04 bc	-0.01 b	-0.09 bc	-0.42 bc	0.43 a	-0.12 bc	-0.48 c	4.47E-17
ITM2C	0.07 ab	0.01 b	-0.06 b	0.15 ab	0.02 ab	0.41 a	0.07 ab	-0.07 b	4.94E-11
HSP90AB2P	-0.19 ab	0.02 ab	0.05 ab	0.28 a	0.12 ab	-0.03 ab	-0.51 b	-0.16 ab	2.62E-15
MCC	-0.11 c	-0.03 bc	-0.08 c	0.42 ab	0.22 abc	0.60 a	0.26 abc	0.29 abc	1.35E-33
STK38	-0.23 ab	-0.14 ab	0.09 a	-0.26 ab	-0.14 ab	0.12 a	-0.38 b	-0.01 ab	1.92E-18
PPIL1	-0.11 ab	-0.10 ab	0.09 a	-0.23 ab	-0.27 ab	-0.04 ab	-0.44 b	-0.12 ab	3.22E-15
C6orf89	0.07 ab	0.10 a	-0.03 ab	0.05 ab	0.19 a	0.05 ab	-0.33 ab	0.14 a	5.86E-06
MTCH1	-0.09 ab	0.08 a	0.00 ab	0.12 a	0.06 ab	0.11 a	-0.45 b	0.08 ab	5.02E-09
TBC1D22B	-0.03 a	0.06 a	0.03 a	-0.05 ab	-0.15 ab	-0.03 a	-0.50 b	0.07 a	2.49E-09
RNF8	-0.15 ab	0.00 a	0.05 a	-0.19 ab	-0.14 ab	0.13 a	-0.46 b	-0.01 ab	1.05E-11
CMTR1	-0.03 ab	-0.02 a	0.05 a	-0.04 ab	-0.10 ab	-0.06 ab	-0.46 b	-0.02 ab	1.40E-08
ZFAND3	-0.04 ab	0.06 a	0.02 ab	0.05 ab	0.14 a	-0.09 ab	-0.43 ab	-0.03 ab	1.92E-07

Table B.1 (cont'd)

BTBD9	-0.03 a	0.13 a	0.03 a	0.29 a	-0.10 a	0.05 a	-1.19 b	0.03 a	1.04E-61
GLO1	-0.21 ab	0.07 a	0.05 a	0.00 ab	0.02 ab	-0.08 ab	-0.50 b	-0.09 ab	1.67E-14
SAYSD1	-0.17 ab	0.01 a	0.04 a	0.17 a	0.21 a	-0.02 a	-0.48 ab	0.18 a	1.60E-12
LRFN2	-0.02 b	0.12 ab	0.05 b	0.49 a	-0.14 b	0.05 b	-1.60 c	0.05 b	1.10E-118
UNC5CL	0.02 ab	-0.05 ab	0.01 ab	0.00 ab	0.02 ab	0.18 a	-0.36 b	0.10 ab	5.13E-06
OARD1	0.18 a	0.13 a	-0.03 a	0.19 a	0.21 a	0.17 a	-0.87 b	0.01 a	1.16E-37
NFYA	-0.09 a	0.09 a	0.04 a	0.06 a	-0.14 ab	-0.04 a	-0.77 b	0.05 a	4.39E-25
FOXP4	0.02 a	0.07 a	0.01 a	0.27 a	0.27 a	-0.04 a	-0.67 b	0.10 a	1.26E-19
TFEB	0.16 ab	0.09 ab	-0.07 ab	0.34 a	0.31 ab	0.31 ab	-0.31 b	0.15 ab	2.71E-18
FRS3	-0.01 b	0.16 b	0.05 b	0.76 a	0.02 b	0.06 b	-2.04 c	0.13 b	9.81E-207
PRICKLE4	0.13 a	0.06 a	-0.03 a	0.02 a	-0.07 a	0.13 a	-0.29 a	0.13 a	1.07E-05
TOMM6	-0.11 a	0.03 a	0.08 a	0.32 a	-0.13 a	0.02 a	-1.24 b	-0.03 a	7.35E-70
USP49	-0.01 a	0.03 a	0.04 a	0.00 a	-0.05 ab	-0.11 ab	-0.65 b	0.06 a	2.95E-17
MED20	-0.05 a	0.02 a	0.09 a	0.15 a	-0.06 a	-0.08 a	-1.28 b	-0.06 a	2.09E-72
BYSL	-0.17 a	-0.03 a	0.10 a	0.19 a	-0.28 ab	0.00 a	-1.09 b	-0.18 a	3.39E-58
CCND3	-0.01 a	0.14 a	-0.02 a	0.40 a	0.12 a	0.16 a	-0.77 b	0.09 a	1.24E-30
TAF8	0.08 a	0.12 a	0.00 a	0.29 a	0.02 a	0.10 a	-1.06 b	0.16 a	2.53E-50
GUCA1A	-0.03 b	0.15 b	0.06 b	0.91 a	0.00 b	0.05 b	-2.16 c	0.04 b	5.46E-242
GUCA1B	0.06 ab	0.10 ab	-0.04 ab	0.36 a	-0.12 ab	0.06 ab	-0.29 b	0.13 ab	3.81E-08
MRPS10	-0.11 a	0.03 a	0.08 a	0.14 a	-0.18 a	0.02 a	-1.22 b	-0.15 a	5.96E-67
TRERF1	-0.01 b	0.15 b	0.05 b	0.92 a	-0.03 b	0.10 b	-2.05 c	0.09 b	6.77E-218

Table B.1 (cont'd)

UBR2	0.03 a	0.11 a	0.03 a	0.22 a	0.11 a	-0.11 a	-0.94 b	0.06 a	4.27E-38
TBCC	-0.08 a	0.06 a	0.04 a	0.40 a	-0.05 a	0.09 a	-1.01 b	0.04 a	2.47E-46
BICRAL	0.10 a	0.13 a	-0.02 a	0.28 a	0.32 a	-0.10 a	-0.60 b	0.16 a	9.26E-20
RPL7L1	-0.10 a	0.09 a	0.04 a	0.29 a	0.08 a	-0.01 a	-1.01 b	0.04 a	2.92E-45
C6orf226	0.07 a	0.04 a	0.00 a	0.29 a	-0.02 ab	0.07 a	-0.68 b	0.22 a	2.97E-20
CNPY3	0.06 ab	0.09 ab	-0.01 b	0.54 a	0.28 ab	0.12 ab	-0.96 c	0.05 ab	3.29E-46
GNMT	0.09 abc	0.26 ab	-0.10 c	0.48 a	0.18 abc	0.02 abc	-0.09 bc	0.21 abc	3.01E-22
PEX6	0.03 a	0.15 a	-0.02 a	0.34 a	-0.04 ab	0.06 a	-0.64 b	0.07 a	1.66E-21
PPP2R5D	-0.08 a	0.09 a	0.05 a	0.32 a	-0.03 a	0.00 a	-1.18 b	-0.05 a	2.00E-61
MEA1	-0.08 a	0.05 a	0.06 a	0.36 a	-0.08 a	-0.01 a	-1.22 b	0.00 a	3.13E-66
KLHDC3	-0.02 a	0.12 a	0.02 a	0.43 a	-0.10 a	0.17 a	-1.14 b	-0.04 a	3.64E-60
RRP36	-0.07 a	0.06 a	0.06 a	0.30 a	0.02 a	0.02 a	-1.18 b	-0.03 a	2.84E-61
CUL7	-0.01 a	0.05 a	0.04 a	0.25 a	0.05 a	-0.12 a	-0.92 b	0.20 a	6.30E-37
KLC4	-0.02 a	0.06 a	0.00 a	0.39 a	0.28 a	0.25 a	-0.86 b	0.15 a	3.94E-37
MRPL2	-0.08 a	-0.01 a	0.06 a	0.39 a	-0.29 a	0.17 a	-1.15 b	-0.03 a	2.73E-61
PTK7	0.00 a	0.01 a	0.02 a	0.18 a	-0.28 ab	0.09 a	-0.56 b	0.05 a	1.39E-12
SRF	0.11 a	0.09 a	-0.02 a	0.30 a	0.26 a	0.15 a	-0.75 b	0.06 a	6.28E-27
CUL9	0.09 a	0.08 a	0.01 a	0.33 a	-0.03 a	-0.01 a	-0.89 b	0.07 a	1.75E-34
DNPH1	0.02 a	-0.05 a	0.02 a	0.33 a	-0.12 ab	0.21 a	-0.74 b	0.28 a	1.48E-27
CRIP3	0.16 ab	0.07 ab	-0.08 ab	0.30 a	0.25 ab	0.29 a	-0.12 ab	0.34 a	2.50E-16
ZNF318	0.00 a	0.08 a	0.04 a	0.16 a	0.08 a	-0.11 a	-0.85 b	0.02 a	3.81E-30
ABCC10	-0.13 a	-0.02 a	0.07 a	0.25 a	0.00 a	-0.02 a	-0.90 b	0.05 a	4.20E-37

Table B.1 (cont'd)

TJAP1	-0.07 a	0.12 a	0.05 a	0.36 a	-0.26 a	-0.04 a	-1.15 b	-0.04 a	1.10E-60
LRRC73	0.09 ab	0.02 ab	-0.01 ab	0.38 a	0.22 ab	-0.11 ab	-0.35 b	0.11 ab	1.32E-08
POLR1C	-0.12 a	0.00 a	0.06 a	0.37 a	-0.02 a	0.09 a	-1.15 b	0.09 a	3.73E-60
YIPF3	-0.01 b	0.11 ab	0.02 b	0.47 a	0.21 ab	0.04 ab	-1.09 c	0.06 ab	9.49E-55
XPO5	-0.17 a	-0.07 a	0.12 a	0.18 a	-0.11 a	-0.14 a	-1.15 b	-0.08 a	9.67E-68
POLH	-0.01 b	0.16 b	0.06 b	0.94 a	-0.02 b	0.04 b	-2.14 c	0.06 b	4.63E-239
GTPBP2	-0.16 b	-0.01 ab	0.10 a	0.06 ab	-0.35 bc	-0.02 ab	-1.00 c	-0.11 ab	4.93E-48
MAD2L1BP	-0.09 a	0.09 a	0.05 a	0.32 a	0.02 a	0.10 a	-1.11 b	-0.15 a	1.38E-55
MRPS18A	0.02 a	0.01 a	0.04 a	0.46 a	-0.07 a	0.11 a	-1.14 b	-0.04 a	3.73E-59
VEGFA	-0.16 ab	0.00 ab	0.04 ab	0.14 a	0.07 ab	0.05 ab	-0.50 b	-0.03 ab	6.75E-12
MRPL14	-0.16 b	0.04 ab	0.05 ab	0.37 a	-0.25 bc	0.16 ab	-0.98 c	0.00 ab	8.68E-47
TMEM63B	-0.04 a	0.06 a	0.07 a	0.20 a	-0.35 ab	-0.16 a	-1.01 b	-0.21 a	1.97E-48
CAPN11	0.13 ab	0.15 ab	-0.07 b	0.42 a	-0.07 ab	0.03 ab	-0.09 b	0.25 ab	3.67E-12
SLC29A1	0.00 a	-0.02 a	0.06 a	0.14 a	-0.30 ab	-0.10 a	-0.74 b	0.08 a	6.62E-24
HSP90AB1	-0.19 a	0.03 a	0.10 a	0.30 a	-0.18 a	-0.15 a	-1.13 b	-0.19 a	1.20E-65
SLC35B2	-0.09 a	0.01 a	0.05 a	0.33 a	-0.06 a	0.09 a	-1.12 b	0.15 a	7.46E-56
NFKBIE	-0.06 ab	0.00 a	0.01 a	0.25 a	-0.04 ab	0.21 a	-0.64 b	0.22 a	5.80E-20
AARS2	-0.06 a	0.02 a	0.05 a	0.33 a	-0.25 ab	0.07 a	-1.04 b	-0.02 a	8.22E-48
CDC5L	-0.07 a	0.08 a	0.04 a	0.25 a	0.03 a	0.04 a	-1.06 b	-0.06 a	1.54E-48
SUPT3H	0.12 a	0.04 a	0.00 a	0.22 a	0.20 a	0.06 a	-0.70 b	0.04 a	7.20E-21
SLC25A27	-0.05 b	0.00 b	-0.03 b	0.23 ab	0.23 ab	0.34 a	-0.11 b	0.11 ab	1.55E-08

Table B.1 (cont'd)

TDRD6	-0.02 b	0.09 b	0.05 b	1.29 a	-0.12 b	-0.06 b	-1.72 c	-0.06 b	1.46E-176
CD2AP	-0.05 ab	0.04 a	0.03 a	0.23 a	0.02 ab	-0.04 ab	-0.68 b	0.11 a	3.75E-19
MUT	-0.03 ab	0.08 a	0.01 a	0.39 a	-0.01 ab	0.01 a	-0.57 b	-0.01 ab	1.43E-15
CENPQ	-0.21 a	-0.05 a	0.10 a	0.10 a	-0.21 ab	-0.06 a	-0.69 b	-0.26 ab	1.37E-29
MCM3	-0.16 a	-0.05 a	0.10 a	0.21 a	-0.20 ab	-0.07 a	-0.81 b	-0.28 ab	3.84E-36
PAQR8	-0.08 b	0.16 b	0.02 b	1.52 a	-0.11 b	0.06 b	-1.79 c	0.02 b	1.29E-209
EFHC1	0.05 ab	0.04 ab	-0.01 ab	0.35 a	0.29 ab	0.02 ab	-0.52 b	0.16 ab	2.00E-13
TRAM2	0.05 abc	0.10 ab	-0.01 bc	0.45 a	0.08 abc	-0.07 bc	-0.56 c	0.08 abc	1.72E-17
TMEM14A	-0.08 ab	-0.03 ab	0.09 a	0.28 a	-0.14 abc	-0.15 abc	-0.79 c	-0.36 bc	1.78E-35
GSTA4	0.11 ab	0.08 ab	-0.02 ab	0.30 a	0.34 a	-0.11 ab	-0.42 ab	0.11 ab	1.99E-11
ICK	-0.02 ab	0.10 ab	0.02 ab	0.44 a	0.23 ab	-0.25 bc	-0.66 c	0.04 ab	4.90E-26
FBXO9	0.02 bc	0.10 bc	-0.02 bc	0.61 a	0.13 abc	-0.06 bc	-0.55 c	0.13 ab	1.42E-21
GCM1	-0.12 a	-0.06 a	0.03 a	0.35 a	-0.06 a	-0.10 a	-0.17 a	0.11 a	1.37E-05
ELOVL5	0.04 ab	0.11 ab	-0.02 b	0.51 a	0.28 ab	-0.20 b	-0.28 b	0.03 ab	1.81E-12
GCLC	0.02 ab	-0.01 ab	0.06 ab	0.21 a	0.06 ab	-0.26 ab	-0.70 b	-0.12 ab	1.37E-25
KLHL31	-0.03 ab	0.01 ab	0.02 ab	0.32 a	-0.02 ab	-0.14 ab	-0.36 b	0.13 ab	5.98E-08
LRRC1	-0.03 a	0.02 a	0.05 a	0.14 a	-0.13 ab	-0.24 ab	-0.59 b	0.08 a	2.03E-17
DST	0.09 a	0.06 a	0.02 ab	0.03 ab	-0.10 ab	-0.13 ab	-0.40 b	-0.18 ab	4.53E-08
KIAA1586	-0.06 ab	0.08 ab	0.01 ab	0.19 a	0.12 ab	0.00 ab	-0.58 b	0.16 a	2.59E-14
ZNF451	-0.04 a	0.10 a	0.01 a	0.36 a	-0.12 ab	-0.03 a	-0.55 b	0.07 a	6.26E-15
BAG2	0.06 ab	0.11 ab	-0.06 b	0.17 ab	0.08 ab	0.24 a	-0.16 b	0.13 ab	4.47E-08
PRIM2	-0.18 ab	0.05 a	0.05 a	0.21 a	-0.05 ab	-0.05 ab	-0.53 b	-0.16 ab	1.93E-15

Table B.1 (cont'd)

GUSBP4	-0.03 a	0.02 a	0.03 a	0.27 a	-0.07 ab	-0.03 a	-0.57 b	0.00 a	1.83E-13
PHF3	0.01 a	0.13 a	-0.02 a	0.01 a	0.35 a	-0.06 a	-0.24 a	0.14 a	2.09E-05
THSD7A	0.04 ab	0.20 a	-0.13 ab	0.42 a	0.55 a	0.35 a	0.20 ab	0.18 ab	1.35E-31
INTS4P2	0.00 b	0.00 b	-0.05 b	0.13 ab	-0.15 b	0.58 a	-0.01 b	-0.23 b	6.07E-21
ARHGEF 10	-0.17 a	0.06 a	-0.03 a	0.14 a	0.38 a	0.20 a	0.15 a	0.03 a	1.05E-06
HTRA4	-0.17 b	-0.07 ab	0.03 ab	-0.10 ab	-0.09 ab	0.23 a	-0.11 ab	-0.14 ab	2.05E-06
DIP2C	0.02 ab	-0.01 b	-0.02 b	-0.12 b	-0.01 b	0.35 a	-0.04 b	-0.13 b	1.46E-06
DNA2	0.02 b	0.00 b	-0.06 b	0.00 b	0.06 ab	0.55 a	0.06 b	0.03 b	7.72E-18
HKDC1	-0.06 b	-0.07 b	-0.04 b	0.06 ab	0.03 ab	0.55 a	0.11 ab	0.23 ab	4.08E-20
CTBP2	0.10 b	0.08 b	-0.10 b	0.13 ab	0.08 b	0.51 a	0.17 ab	0.09 b	5.97E-23
FAM196 A	-0.04 b	-0.10 b	0.02 b	-0.09 b	0.04 ab	0.39 a	-0.21 b	-0.33 b	9.00E-14
CSNK2A 3	-0.51 c	0.38 a	-0.08 bc	0.23 ab	0.18 abc	0.22 ab	0.20 ab	0.21 ab	8.41E-53
FOLH1	0.06 a	0.14 a	-0.06 a	0.31 a	0.28 a	0.10 a	-0.13 a	0.10 a	6.03E-08
MRPL16	-0.12 b	-0.13 b	0.00 b	-0.13 b	0.01 ab	0.50 a	0.00 b	0.09 ab	8.72E-17
EIF1AD	0.23 ab	0.07 ab	-0.09 b	0.27 ab	0.02 ab	0.28 a	0.18 ab	-0.21 b	5.57E-17
SF3B2	0.11 a	-0.03 a	-0.03 a	0.21 a	-0.20 a	0.21 a	0.07 a	-0.27 a	4.31E-06
PACS1	-0.11 b	-0.07 b	0.01 b	0.06 ab	-0.15 b	0.32 a	0.02 ab	-0.26 b	4.39E-08
KLC2	0.11 b	-0.03 b	-0.06 b	0.15 ab	-0.06 b	0.58 a	0.05 b	-0.18 b	1.75E-22
RAB1B	0.04 b	-0.08 b	0.00 b	0.06 ab	-0.29 b	0.42 a	-0.06 b	-0.40 b	5.73E-14
YIF1A	0.17 ab	0.07 ab	-0.05 bc	0.00 abc	-0.03 abc	0.31 a	0.04 abc	-0.42 c	3.28E-13
BRMS1	0.15 ab	0.05 ab	-0.07 b	0.16 ab	0.03 ab	0.38 a	0.06 ab	-0.22 b	1.47E-13
MRPL11	0.16 ab	0.10 ab	-0.07 b	0.19 ab	0.00 ab	0.28 a	0.06 ab	-0.16 b	4.33E-11

Table B.1 (cont'd)

PELI3	-0.02 b	-0.06 b	-0.03 b	0.06 b	-0.23 b	0.60 a	0.06 b	-0.34 b	1.59E-23
DPP3	0.17 b	0.09 b	-0.10 b	0.19 ab	-0.19 b	0.54 a	0.06 b	-0.17 b	2.36E-27
BBS1	-0.12 b	-0.09 b	0.04 ab	0.06 ab	-0.51 b	0.28 a	-0.07 ab	-0.43 b	1.69E-12
ZDHHC24	0.08 b	0.04 b	-0.05 bc	-0.03 bc	-0.23 bc	0.63 a	-0.04 bc	-0.52 c	1.25E-30
CCDC87	-0.02 b	-0.04 b	-0.01 b	0.00 ab	-0.27 b	0.40 a	0.07 ab	-0.29 b	1.31E-10
CCS	0.04 ab	-0.08 bc	0.01 ab	-0.01 abc	-0.13 bc	0.31 a	-0.11 bc	-0.50 c	2.67E-11
RBM4	-0.01 b	-0.09 b	0.00 b	0.07 ab	-0.30 b	0.40 a	-0.01 b	-0.29 b	2.04E-11
C11orf80	0.12 b	0.06 b	-0.06 b	0.12 ab	0.14 ab	0.49 a	-0.07 b	-0.31 b	2.06E-19
RCE1	0.09 b	0.10 b	-0.11 b	0.31 ab	0.07 b	0.69 a	0.10 b	-0.22 b	4.34E-40
PC	0.06 ab	-0.12 b	-0.02 b	0.03 ab	-0.07 b	0.43 a	0.05 ab	-0.12 b	5.98E-12
LRFN4	0.19 ab	0.09 b	-0.09 b	0.01 b	0.21 ab	0.47 a	-0.01 b	-0.11 b	8.26E-21
RHOD	-0.04 b	-0.01 ab	-0.01 b	0.13 ab	0.10 ab	0.30 a	-0.03 b	-0.28 b	2.78E-06
KDM2A	0.03 b	-0.02 b	-0.06 b	0.20 b	-0.15 b	0.82 a	0.02 b	-0.40 b	6.02E-46
GRK2	0.09 b	-0.05 b	-0.03 b	-0.11 b	-0.23 b	0.56 a	-0.11 b	-0.26 b	1.20E-20
ANKRD13D	0.11 b	0.01 b	-0.08 b	0.17 b	0.11 b	0.75 a	-0.06 b	-0.29 b	1.54E-39
SSH3	0.09 b	0.04 b	-0.07 b	0.21 ab	0.13 ab	0.63 a	-0.10 b	-0.34 b	1.02E-29
RAD9A	0.18 b	0.04 bc	-0.09 bc	0.19 b	0.10 bc	0.76 a	-0.02 bc	-0.45 c	1.33E-47
POLD4	-0.02 bc	-0.06 bc	0.00 b	-0.17 bc	-0.37 bc	0.56 a	-0.13 bc	-0.49 c	1.67E-25
PPP1CA	0.12 b	-0.01 bc	-0.06 bc	0.06 bc	-0.13 bc	0.75 a	-0.05 bc	-0.41 c	1.86E-39
RPS6KB2	0.13 b	-0.01 b	-0.06 b	0.08 b	0.13 ab	0.61 a	0.04 b	-0.49 b	7.68E-30
CORO1B	0.07 b	-0.08 bc	-0.02 bc	-0.03 bc	-0.42 bc	0.66 a	0.00 bc	-0.53 c	1.20E-33
CABP4	-0.04 b	-0.10 b	-0.07 b	-0.04 b	-0.40 bc	1.64 a	-0.15 b	-1.06 c	7.53E-214

Table B.1 (cont'd)

TMEM134	-0.05 b	-0.10 b	-0.01 b	0.07 b	-0.35 b	0.70 a	-0.07 b	-0.43 b	2.53E-34
AIP	0.06 b	-0.04 bc	-0.02 b	-0.03 bc	-0.08 bc	0.54 a	-0.09 bc	-0.56 c	2.70E-24
PITPNM1	-0.08 b	-0.10 b	-0.05 b	-0.02 b	-0.32 bc	1.51 a	-0.14 b	-1.03 c	6.17E-181
CDK2AP2	0.02 b	0.01 b	-0.03 bc	0.00 bc	-0.18 bc	0.57 a	-0.03 bc	-0.58 c	5.03E-26
NDUFV1	0.06 b	0.00 b	-0.03 b	-0.04 bc	-0.06 bc	0.58 a	-0.08 bc	-0.54 c	3.38E-26
NUDT8	-0.14 b	-0.06 ab	0.02 ab	0.10 ab	0.02 ab	0.27 a	-0.01 ab	-0.33 b	1.31E-07
ALDH3B2	0.01 b	-0.01 b	-0.05 b	0.25 ab	0.12 ab	0.54 a	-0.03 b	-0.08 b	2.38E-18
UNC93B1	0.01 b	-0.08 b	-0.07 b	-0.02 b	-0.40 b	1.66 a	-0.15 b	-1.22 c	6.81E-230
ALDH3B1	-0.01 b	-0.09 b	-0.07 b	-0.02 b	-0.37 b	1.72 a	-0.16 b	-1.28 c	6.74E-250
NDUFS8	0.05 b	0.00 b	-0.04 bc	-0.03 bc	-0.02 bc	0.70 a	-0.06 bc	-0.60 c	1.45E-37
TCIRG1	0.00 b	-0.10 b	0.01 b	-0.07 b	0.03 ab	0.40 a	-0.17 b	-0.37 b	1.32E-13
CHKA	0.07 ab	0.00 b	-0.02 b	-0.21 bc	-0.16 bc	0.41 a	0.06 ab	-0.66 c	1.82E-20
KMT5B	-0.01 b	-0.11 b	-0.01 b	0.15 b	-0.41 bc	0.74 a	0.00 b	-0.73 c	9.19E-48
C11orf24	0.01 bc	-0.02 bc	-0.06 bc	0.07 bc	-0.02 bc	0.78 a	0.13 b	-0.54 c	1.21E-44
LRP5	-0.05 b	-0.12 b	-0.05 b	-0.08 b	-0.36 b	1.64 a	-0.11 b	-1.35 c	1.47E-233
PPP6R3	0.04 b	-0.06 b	-0.05 b	0.16 b	-0.23 bc	0.89 a	0.02 b	-0.68 c	2.65E-61
TESMIN	0.07 b	0.01 b	-0.07 b	-0.05 b	-0.33 b	0.82 a	-0.01 b	-0.38 b	3.03E-45
CPT1A	-0.03 b	-0.14 b	-0.09 b	-0.08 b	-0.03 b	1.85 a	-0.04 b	-1.04 c	3.87E-270
MRPL21	0.02 b	-0.10 b	-0.11 b	0.01 b	-0.19 b	1.95 a	-0.06 b	-1.03 c	2.94E-303
IGHMBP2	-0.03 b	-0.11 b	-0.09 b	0.01 b	-0.15 bc	1.76 a	-0.12 b	-0.91 c	1.02E-236
MRGPRD	-0.02 ab	-0.11 ab	-0.01 ab	0.36 a	-0.10 ab	0.35 a	-0.08 ab	0.05 ab	1.44E-10

Table B.1 (cont'd)

TPCN2	-0.04 b	-0.12 b	-0.10 b	0.08 b	-0.18 b	1.97 a	-0.06 b	-1.05 c	0
CCND1	0.01 b	-0.15 b	-0.12 b	-0.11 b	-0.20 bc	2.14 a	-0.12 b	-0.86 c	0
ORAOV1	-0.01 b	-0.12 bc	-0.12 bc	-0.08 bc	-0.06 bc	2.00 a	-0.11 bc	-0.76 c	2.27E-302
ANO1	-0.08 b	-0.14 b	-0.12 b	0.05 b	-0.02 b	2.12 a	-0.02 b	-0.82 c	0
FADD	0.07 b	-0.05 bc	-0.11 bc	0.18 b	-0.04 bc	1.43 a	0.01 b	-0.59 c	1.82E-147
PPFIA1	0.01 b	-0.11 b	-0.14 b	0.17 b	-0.10 b	2.10 a	0.03 b	-0.86 c	0
CTTN	0.02 b	-0.13 b	-0.13 b	0.08 b	-0.02 b	2.02 a	0.00 b	-0.86 c	0
SHANK2	-0.10 b	-0.09 b	-0.11 b	0.25 b	0.17 b	1.78 a	0.07 b	-0.84 c	3.59E-244
DHCR7	0.13 b	0.08 b	-0.09 bc	0.00 bc	-0.21 bc	0.73 a	0.14 b	-0.41 c	5.83E-43
NADSYN1	-0.09 b	-0.13 b	-0.06 b	0.08 b	-0.03 b	1.67 a	-0.09 b	-1.25 c	1.60E-233
KRTAP5-7	-0.01 b	-0.04 b	-0.04 b	0.08 ab	-0.10 b	0.59 a	0.02 b	-0.13 b	2.17E-20
KRTAP5-8	-0.06 b	-0.03 b	-0.03 b	0.07 ab	0.02 ab	0.50 a	-0.01 b	-0.03 b	4.14E-14
KRTAP5-9	-0.09 b	-0.04 b	-0.01 b	0.06 ab	-0.09 b	0.45 a	-0.05 b	-0.22 b	4.90E-12
KRTAP5-10	-0.01 b	-0.04 b	-0.01 b	-0.10 b	-0.05 b	0.44 a	-0.06 b	-0.26 b	1.00E-11
FAM86C1	-0.02 b	0.04 b	-0.07 b	0.15 b	-0.25 bc	0.93 a	0.07 b	-0.75 c	3.71E-70
RNF121	-0.05 b	-0.03 b	-0.03 b	0.18 b	-0.22 bc	0.79 a	-0.02 b	-0.94 c	6.34E-60
NUMA1	-0.11 b	-0.14 b	-0.02 b	-0.01 b	-0.45 b	1.55 a	-0.10 b	-1.71 c	1.49E-246
LRTOMT	-0.12 b	-0.08 b	0.02 b	0.02 b	-0.40 bc	0.61 a	-0.11 b	-0.81 c	4.42E-40
LAMTOR1	-0.10 b	-0.06 b	-0.01 b	0.08 b	-0.36 bc	0.75 a	0.01 b	-0.94 c	1.23E-56
ANAPC15	-0.11 b	-0.03 b	0.02 b	0.02 b	-0.19 bc	0.46 a	-0.13 b	-0.72 c	3.16E-25
INPPL1	0.16 b	-0.06 b	-0.05 b	0.01 b	-0.13 bc	0.76 a	0.01 b	-0.75 c	5.98E-51
CLPB	0.10 bc	0.14 b	-0.10 cd	0.14 bc	0.01 bcd	0.74 a	0.09 bc	-0.57 d	2.31E-50

Table B.1 (cont'd)

ARAP1	-0.08 bc	-0.06 bc	0.01 b	-0.11 bc	-0.04 bc	0.59 a	-0.15 bc	-0.78 c	7.45E-36
STARD10	0.03 b	-0.05 b	-0.01 b	-0.09 bc	-0.23 bc	0.66 a	-0.09 bc	-0.67 c	8.25E-37
ATG16L2	-0.06 b	-0.03 b	0.02 b	-0.01 b	0.11 ab	0.34 a	-0.28 b	-0.52 b	1.68E-15
FCHSD2	-0.14 bc	-0.08 b	0.03 b	0.02 b	-0.09 bc	0.48 a	-0.17 bc	-0.72 c	3.54E-28
ARHGEF17	-0.16 bc	-0.10 bc	0.05 ab	-0.11 bc	-0.14 bc	0.31 a	-0.08 abc	-0.48 c	4.06E-15
RELT	0.03 b	0.05 b	-0.05 b	0.11 ab	0.10 ab	0.41 a	-0.01 b	-0.25 b	1.60E-11
RAB6A	0.07 b	0.00 b	-0.06 b	0.15 b	-0.30 bc	0.73 a	0.14 b	-0.68 c	5.61E-46
MRPL48	0.06 b	0.06 b	-0.05 b	0.04 b	-0.18 bc	0.55 a	0.16 ab	-0.78 c	5.02E-35
COA4	0.07 b	0.03 b	-0.05 b	0.03 b	-0.24 bc	0.68 a	0.11 b	-0.94 c	9.29E-51
PAAF1	0.01 b	0.07 b	-0.03 b	-0.08 b	-0.16 bc	0.54 a	0.00 b	-0.89 c	3.25E-36
UCP3	-0.16 b	0.00 ab	0.01 ab	0.03 ab	-0.15 ab	0.30 a	-0.01 ab	-0.46 b	8.15E-11
C2CD3	-0.07 b	-0.02 b	-0.04 b	0.15 b	-0.32 bc	0.74 a	0.14 b	-0.69 c	8.74E-47
PPME1	0.07 b	0.05 b	-0.09 b	0.12 b	-0.25 bc	0.88 a	0.19 b	-0.66 c	4.14E-64
LIPT2	-0.02 b	0.03 b	-0.04 b	0.01 b	-0.45 bc	0.52 a	0.21 ab	-0.65 c	1.00E-28
POLD3	-0.03 b	0.02 b	-0.06 b	0.32 ab	-0.12 bc	0.56 a	0.22 ab	-0.56 c	5.18E-31
RNF169	-0.06 bc	-0.05 bc	-0.03 bc	0.25 ab	-0.26 bc	0.63 a	0.11 b	-0.46 c	1.43E-30
XRRA1	-0.01 b	-0.08 b	0.00 b	0.01 b	-0.42 bc	0.66 a	0.05 b	-0.91 c	3.25E-47
SPCS2	0.04 b	-0.01 b	-0.01 b	0.02 b	-0.24 bc	0.55 a	-0.06 b	-1.04 c	4.36E-44
NEU3	0.01 b	-0.02 b	-0.02 b	0.04 ab	-0.20 b	0.45 a	0.03 ab	-0.43 b	4.20E-15
ARRB1	-0.08 b	-0.06 b	-0.02 b	-0.11 b	-0.42 b	1.31 a	-0.02 b	-1.77 c	9.70E-200
RPS3	-0.05 b	-0.03 b	0.02 b	-0.08 b	-0.39 bc	0.41 a	-0.05 b	-0.91 c	3.08E-30
UVRAG	-0.10 b	0.02 b	-0.04 b	0.24 ab	-0.48 bc	0.69 a	0.11 b	-0.72 c	1.14E-44

Table B.1 (cont'd)

THAP12	0.01 b	-0.02 b	-0.02 b	0.04 b	-0.44 bc	0.60 a	0.13 b	-0.83 c	4.51E-39
EMSY	-0.06 bc	-0.07 bc	-0.03 b	0.15 b	-0.31 bc	0.69 a	0.08 b	-0.53 c	4.76E-36
LRRC32	-0.06 b	-0.05 b	-0.04 b	-0.05 b	-0.45 b	1.34 a	-0.03 b	-1.51 c	4.02E-180
TSKU	-0.11 bc	-0.12 bc	0.02 bc	0.02 abc	-0.06 bc	0.46 a	0.09 ab	-0.53 c	6.54E-21
ACER3	-0.06 bc	0.00 b	-0.04 b	0.14 b	-0.15 bc	0.66 a	0.11 b	-0.55 c	8.76E-34
MYO7A	-0.05 b	-0.05 b	-0.05 b	-0.03 b	-0.39 b	1.44 a	-0.03 b	-1.40 c	3.28E-189
PAK1	0.00 b	-0.02 b	-0.04 b	0.13 b	-0.55 c	0.70 a	0.06 b	-0.55 c	2.36E-38
CLNS1A	0.08 b	0.04 b	-0.08 b	0.14 b	-0.42 bc	0.82 a	0.14 b	-0.76 c	7.70E-61
AQP11	-0.01 b	-0.03 b	0.00 b	-0.16 b	-0.36 b	0.40 a	-0.02 b	-0.45 b	7.05E-14
RSF1	-0.05 b	-0.06 b	-0.05 b	-0.02 b	-0.43 b	1.40 a	0.02 b	-1.26 c	1.19E-171
AAMDC	-0.06 b	-0.03 b	0.00 b	-0.07 bc	-0.27 bc	0.51 a	-0.07 bc	-0.68 c	5.25E-26
INTS4	-0.05 b	-0.01 b	-0.05 b	0.14 b	-0.36 bc	0.99 a	0.06 b	-0.85 c	1.20E-79
NDUFC2	0.03 b	-0.05 b	-0.04 b	-0.07 b	-0.33 bc	0.87 a	0.01 b	-0.83 c	5.68E-64
ALG8	0.06 b	0.07 b	-0.09 b	0.02 b	-0.24 bc	1.05 a	0.05 b	-0.84 c	1.48E-91
KCTD21	0.04 b	-0.08 b	-0.03 b	0.07 b	-0.20 bc	0.80 a	-0.03 b	-0.74 c	2.07E-52
USP35	0.03 b	-0.04 b	-0.05 b	-0.11 b	-0.31 b	1.34 a	-0.07 b	-1.33 c	1.49E-165
NARS2	0.04 b	0.02 b	-0.03 b	0.09 ab	-0.48 bc	0.60 a	0.04 b	-0.75 c	4.23E-36
CCDC77	0.11 a	0.03 a	-0.05 a	0.34 a	-0.24 a	0.01 a	0.10 a	0.22 a	1.02E-05
WNK1	0.09 ab	-0.01 ab	-0.05 b	0.15 ab	-0.34 b	0.28 a	0.08 ab	0.11 ab	1.27E-06
ADIPOR2	0.12 ab	0.00 ab	-0.06 b	0.24 ab	-0.40 b	0.31 a	0.12 ab	0.14 ab	8.02E-11
DCP1B	-0.10 a	-0.03 a	0.05 a	0.00 a	-0.64 a	-0.08 a	-0.14 a	-0.08 a	1.78E-05
FKBP4	0.21 ab	-0.01 bc	-0.08 c	0.14 abc	-0.38 c	0.38 a	0.12 abc	0.37 ab	3.77E-20

Table B.1 (cont'd)

RHNO1	0.22 a	0.02 ab	-0.07 ab	0.24 a	-0.46 ab	0.21 ab	0.13 ab	0.05 ab	4.58E-11
TULP3	0.14 a	0.01 a	-0.05 a	0.30 a	-0.26 a	0.15 a	0.10 a	0.05 a	4.54E-06
TSPAN9	-0.08 a	-0.10 a	0.06 a	-0.07 a	-0.36 a	-0.01 a	-0.28 a	-0.06 a	4.44E-07
PARP11	-0.09 a	-0.06 a	0.05 a	-0.19 a	-0.68 a	0.02 a	-0.14 a	-0.11 a	1.08E-06
CCND2	-0.09 ab	0.00 a	0.08 a	-0.31 ab	-0.28 ab	-0.38 b	-0.22 ab	-0.20 ab	5.54E-17
RAD51AP1	0.25 ab	0.13 ab	-0.13 b	0.41 a	0.08 ab	0.25 ab	0.24 ab	0.26 ab	1.09E-27
DYRK4	-0.16 a	-0.03 a	0.06 a	-0.16 a	-0.48 a	-0.10 a	-0.15 a	-0.19 a	8.96E-08
AKAP3	-0.15 ab	-0.11 ab	0.09 a	-0.20 ab	-0.34 b	-0.24 b	-0.10 ab	0.01 ab	4.76E-12
TNFRSF1A	-0.02 a	-0.02 a	0.00 a	0.03 a	-0.86 b	0.22 a	-0.05 ab	0.06 a	2.66E-07
LTBR	0.14 a	0.07 a	-0.06 a	0.25 a	-0.25 a	0.21 a	0.08 a	-0.02 a	6.41E-08
NCAPD2	0.11 ab	0.10 ab	-0.10 ab	0.35 a	-0.16 ab	0.24 a	0.24 a	0.24 a	1.92E-16
GAPDH	0.16 a	0.11 a	-0.07 a	0.20 a	-0.14 a	0.07 a	0.08 a	0.24 a	1.14E-08
NOP2	0.15 ab	0.13 ab	-0.09 b	0.17 ab	-0.19 b	0.24 a	0.19 ab	0.12 ab	1.03E-13
ING4	-0.04 a	-0.06 a	0.04 a	0.25 a	-0.52 a	-0.18 a	-0.02 a	-0.04 a	2.19E-05
ZNF384	0.01 a	0.05 a	-0.05 a	0.22 a	-0.46 a	0.25 a	0.14 a	-0.02 a	3.32E-07
COPS7A	0.00 a	0.04 a	-0.04 a	0.21 a	-0.52 a	0.17 a	0.06 a	0.22 a	7.81E-06
MLF2	0.09 a	0.02 a	-0.04 a	0.06 a	-0.45 a	0.23 a	0.10 a	0.10 a	9.65E-06
CDCA3	0.27 a	0.17 ab	-0.14 b	0.31 a	0.23 ab	0.27 a	0.25 a	0.31 a	1.25E-33
USP5	0.06 a	0.09 a	-0.06 a	0.17 a	-0.36 a	0.15 a	0.13 a	0.27 a	9.67E-08
TPI1	0.17 a	0.13 a	-0.08 ab	0.25 a	-0.23 ab	0.00 ab	0.12 ab	0.25 a	1.25E-10
ATN1	0.03 a	-0.03 a	-0.02 a	0.08 a	-0.53 a	0.25 a	0.04 a	0.10 a	3.65E-05
C12orf57	-0.31 b	-0.22 b	0.11 a	-0.12 ab	-0.56 b	0.02 ab	-0.09 ab	-0.07 ab	4.00E-24

Table B.1 (cont'd)

SCARNA 12	0.14 a	0.12 a	-0.07 a	0.18 a	-0.13 a	0.15 a	0.06 a	0.01 a	1.25E-07
EMG1	0.18 a	0.11 a	-0.06 a	0.09 a	-0.19 a	0.08 a	0.08 a	-0.01 a	1.05E-06
LPCAT3	-0.07 b	0.05 ab	-0.04 b	0.13 ab	-0.37 b	0.31 a	0.12 ab	-0.03 ab	1.34E-07
NECAP1	0.11 ab	-0.05 ab	-0.01 ab	0.00 ab	-0.67 b	0.16 a	-0.02 ab	0.11 ab	2.30E-05
CLEC4A	0.01 ab	0.01 ab	-0.01 ab	-0.03 ab	-0.79 b	0.21 a	-0.04 ab	0.07 ab	4.06E-06
DDX12P	0.17 a	0.09 ab	-0.10 ab	0.33 a	0.10 ab	0.26 a	0.13 ab	0.14 ab	4.89E-15
GABARAPL1	-0.11 a	-0.08 a	0.05 a	-0.02 a	-0.65 a	-0.13 a	0.11 a	0.00 a	1.48E-06
MAGOHB	0.19 a	0.16 a	-0.10 ab	0.12 ab	-0.16 ab	0.20 a	0.06 ab	0.26 a	1.96E-15
LOH12CR2	-0.13 a	-0.03 a	0.05 a	0.13 a	-0.59 a	-0.12 a	-0.03 a	-0.22 a	1.54E-06
BORCS5	-0.01 a	0.04 a	-0.04 a	0.21 a	-0.61 a	0.13 a	0.16 a	0.20 a	2.00E-06
CREBL2	-0.10 a	-0.11 a	0.06 a	-0.06 a	-0.70 a	0.08 a	-0.12 a	-0.17 a	1.86E-08
GPR19	0.14 ab	0.14 ab	-0.11 b	0.13 ab	0.21 ab	0.23 a	0.21 ab	0.37 a	5.50E-19
DDX47	0.05 a	0.02 a	-0.04 a	0.22 a	-0.54 a	0.12 a	0.20 a	0.16 a	3.30E-06
FAM234B	-0.04 b	0.01 b	-0.05 b	0.12 ab	-0.11 b	0.41 a	0.07 ab	0.13 ab	2.99E-10
WBP11	0.12 ab	0.01 ab	-0.05 ab	0.15 ab	-0.60 b	0.23 a	0.10 ab	0.19 ab	1.13E-08
STRAP	0.15 a	0.08 a	-0.06 a	0.15 a	-0.51 a	0.10 a	0.17 a	0.17 a	1.35E-08
DERA	0.01 a	0.05 a	-0.03 a	0.37 a	-0.29 a	-0.04 a	0.22 a	-0.01 a	2.47E-05
RECQL	0.11 a	0.04 a	-0.05 a	0.30 a	-0.18 a	0.09 a	0.15 a	0.12 a	1.33E-05
GOLT1B	0.18 a	0.04 ab	-0.07 ab	0.32 a	-0.16 ab	0.13 ab	0.18 ab	0.14 ab	6.73E-10
CMAS	0.27 a	0.06 ab	-0.07 b	0.23 ab	-0.45 b	0.03 ab	0.07 ab	0.19 ab	1.98E-11
ETFRF1	-0.09 a	-0.10 a	0.07 a	0.17 a	-0.55 a	-0.12 a	-0.20 a	-0.12 a	6.68E-09
KRAS	0.12 a	0.02 a	-0.06 a	0.31 a	-0.26 a	0.15 a	0.21 a	0.15 a	3.64E-08

Table B.1 (cont'd)

INTS13	0.12 a	0.12 a	-0.07 a	0.31 a	-0.21 a	0.01 a	0.18 a	0.06 a	3.16E-08
MED21	0.04 a	0.02 a	-0.03 a	0.36 a	-0.51 a	0.05 a	0.13 a	0.07 a	2.47E-05
DDX11	0.08 a	0.09 a	-0.07 a	0.36 a	-0.15 a	0.17 a	0.11 a	0.17 a	2.73E-09
H3F3C	0.09 ab	0.00 ab	-0.03 ab	0.07 ab	-0.71 b	0.20 a	0.04 ab	0.05 ab	4.35E-06
DNM1L	0.13 a	0.03 a	-0.07 a	0.38 a	-0.34 a	0.19 a	0.16 a	0.11 a	3.68E-10
YARS2	0.16 a	0.09 a	-0.06 a	0.27 a	-0.22 a	0.05 a	0.06 a	0.07 a	2.30E-06
ALG10	0.01 ab	0.02 ab	-0.05 ab	0.36 a	-0.25 ab	0.26 a	0.16 ab	0.05 ab	1.38E-08
HOXC13	0.04 ab	0.22 ab	-0.14 b	0.43 a	0.38 ab	0.38 ab	0.17 ab	0.39 a	7.99E-37
HOXC11	-0.05 ab	0.21 a	-0.10 ab	0.33 a	-0.14 ab	0.28 a	0.19 ab	0.21 ab	1.84E-21
HOXC8	-0.03 ab	0.22 ab	-0.10 b	0.26 ab	-0.06 ab	0.31 a	0.06 ab	0.15 ab	3.51E-19
GPR132	-0.22 ab	0.09 ab	-0.04 ab	0.19 a	0.35 a	0.18 a	0.18 ab	0.17 ab	4.84E-10
WASH3P	-0.25 ab	0.22 a	-0.04 ab	0.24 a	0.36 a	-0.02 ab	0.01 ab	0.07 ab	4.07E-14
MYO1C	-0.15 b	0.12 ab	-0.03 ab	-0.06 ab	-0.27 b	0.26 a	0.06 ab	-0.04 ab	9.07E-08
NTN1	0.02 ab	0.05 a	0.03 a	-0.09 ab	-0.24 ab	-0.18 ab	-0.37 b	-0.06 ab	1.54E-07
TRIM16L	-0.21 b	0.08 ab	-0.05 b	0.15 ab	0.26 ab	0.29 a	0.43 a	-0.04 ab	1.67E-17
ANKRD13B	0.06 ab	0.25 a	-0.12 ab	0.27 a	0.31 a	0.20 a	0.10 ab	0.27 a	3.07E-23
SPACA3	-0.38 b	-0.01 ab	0.02 ab	0.12 ab	-0.08 ab	0.18 a	0.12 a	0.05 ab	4.45E-13
SCARNA17	-0.27 ab	0.12 a	0.01 ab	0.05 ab	0.14 a	-0.03 ab	0.04 ab	-0.22 ab	6.98E-08
ABCA7	0.02 ab	0.19 a	-0.08 ab	0.21 a	0.15 ab	0.21 a	0.06 ab	0.00 ab	8.46E-12
EPS15L1	-0.16 bc	-0.20 c	0.05 b	0.01 bc	-0.47 c	0.35 a	0.02 bc	-0.28 c	6.44E-19
CEACAM8	-0.32 ab	0.02 ab	0.01 ab	0.09 ab	0.33 a	0.13 a	0.19 a	-0.02 ab	5.35E-10
JOSD2	-0.01 b	0.00 b	-0.04 b	0.13 ab	-0.09 b	0.41 a	-0.01 b	-0.04 b	2.56E-09

Table B.1 (cont'd)

FKBP1AP1	-0.28 ab	0.19 a	-0.02 ab	0.05 ab	0.29 a	0.07 ab	0.03 ab	-0.08 ab	3.41E-11
DEFB126	-0.10 ab	0.10 ab	-0.07 ab	0.34 a	0.13 ab	0.31 a	0.19 ab	0.05 ab	7.72E-13
C20orf96	-0.58 b	0.36 a	0.00 b	0.02 ab	0.08 ab	-0.08 b	-0.07 b	-0.12 b	2.44E-47
ZCCHC3	-0.86 c	0.52 a	-0.04 b	0.18 ab	0.36 ab	-0.03 b	0.20 ab	0.09 ab	1.21E-111
SOX12	-0.53 c	0.35 a	-0.07 b	0.09 ab	0.35 ab	0.24 ab	0.24 ab	0.17 ab	5.60E-51
NRSN2	-0.54 c	0.28 a	-0.03 b	0.05 ab	0.26 ab	0.12 ab	0.17 ab	0.17 ab	1.29E-38
TRIB3	-0.14 ab	0.33 a	-0.10 ab	-0.12 ab	0.44 a	0.16 ab	0.26 a	0.25 ab	4.92E-31
RBCK1	-0.80 c	0.62 a	-0.06 b	0.05 b	0.36 ab	-0.01 b	0.13 b	0.09 b	1.42E-122
TBC1D20	-0.91 c	0.51 a	-0.04 b	0.11 b	0.21 ab	0.14 b	0.17 ab	0.09 b	6.61E-117
CSNK2A1	-0.77 d	0.58 a	-0.10 c	0.24 abc	0.26 abc	0.21 bc	0.28 ab	0.27 abc	1.93E-121
SRXN1	-0.44 b	0.39 a	-0.10 b	0.08 ab	0.31 ab	0.25 ab	0.42 a	0.21 ab	5.15E-57
SLC52A3	-0.24 c	0.22 a	-0.04 bc	0.06 abc	0.16 abc	0.17 ab	0.07 abc	-0.14 bc	1.34E-13
FAM110A	-0.51 b	0.39 a	-0.07 a	0.13 a	0.49 a	0.22 a	0.10 a	0.23 a	3.39E-53
PSMF1	-1.05 c	0.67 a	-0.07 b	0.11 b	0.44 ab	0.20 b	0.27 b	0.11 b	3.66E-186
TMEM74B	-0.12 ab	0.29 a	-0.05 ab	0.07 ab	0.32 a	-0.08 ab	0.06 ab	-0.10 ab	2.59E-15
C20orf202	-0.18 ab	0.21 a	-0.02 ab	-0.20 ab	0.21 a	-0.05 ab	0.08 ab	-0.05 ab	5.42E-09
SNPH	-1.77 e	1.12 a	-0.09 d	0.04 cd	0.78 ab	0.21 bc	0.24 bc	0.26 bc	0
SDCBP2	-0.33 b	0.17 a	0.00 ab	0.15 ab	0.14 ab	-0.10 ab	0.10 ab	0.01 ab	3.23E-13
FKBP1A	-0.76 b	0.44 a	-0.03 a	0.05 a	0.48 a	0.03 a	0.22 a	0.11 a	1.65E-83
NSFL1C	-1.02 c	0.64 a	-0.08 b	0.27 ab	0.56 ab	0.15 b	0.22 b	0.28 ab	1.93E-175
SIRPA	-0.25 b	0.13 a	0.00 ab	0.12 ab	0.02 ab	-0.04 ab	-0.01 ab	-0.11 ab	8.68E-07

Table B.1 (cont'd)

STK35	-0.63 c	0.56 a	-0.13 b	0.26 ab	0.31 ab	0.33 a	0.26 ab	0.39 a	4.34E-109
TGM6	-1.82 e	1.13 a	-0.09 d	0.02 cd	0.76 ab	0.26 bc	0.25 bc	0.28 bc	0
SNRPB	-0.61 c	0.58 a	-0.12 b	0.22 ab	0.64 a	0.25 ab	0.24 ab	0.30 ab	8.01E-109
ZNF343	-0.81 c	0.57 a	-0.09 b	0.21 b	0.34 ab	0.24 b	0.21 b	0.16 b	8.30E-123
TMC2	-0.30 b	0.18 a	0.00 ab	0.00 ab	0.10 ab	0.03 ab	-0.10 ab	-0.02 ab	3.89E-11
NOP56	-0.64 c	0.55 a	-0.10 b	0.08 b	0.49 ab	0.15 b	0.22 ab	0.29 ab	3.45E-96
IDH3B	-0.91 c	0.62 a	-0.06 b	0.18 b	0.42 ab	0.00 b	0.19 b	0.27 ab	1.62E-143
EBF4	-0.40 b	0.08 a	0.03 a	-0.13 ab	-0.20 ab	0.07 a	-0.07 ab	0.17 a	1.39E-14
PCED1A	-0.64 b	0.46 a	-0.03 a	0.00 a	0.52 a	0.03 a	0.02 a	-0.15 ab	6.69E-69
VPS16	-0.92 c	0.61 a	-0.08 b	0.16 b	0.56 ab	0.17 b	0.20 b	0.25 ab	2.78E-148
PTPRA	-0.88 c	0.52 a	-0.04 b	0.10 b	0.27 ab	0.18 b	0.11 b	0.12 b	6.66E-115
MRPS26	-0.76 c	0.57 a	-0.06 b	0.05 b	0.42 ab	0.08 b	0.15 b	0.07 b	3.16E-106
OXT	-1.81 e	1.15 a	-0.10 d	0.04 cd	0.76 ab	0.25 bc	0.24 bc	0.28 bc	0
UBOX5	-0.90 c	0.51 a	-0.03 b	-0.02 b	0.31 ab	0.22 ab	0.11 b	-0.05 b	1.43E-115
FASTKD5	-0.78 c	0.55 a	-0.07 b	0.16 ab	0.40 ab	0.12 b	0.20 ab	0.14 ab	1.38E-107
DDRGK1	-0.84 c	0.52 a	-0.01 b	-0.13 b	0.14 ab	0.02 b	-0.02 b	-0.02 b	3.47E-104
ITPA	-0.76 c	0.51 a	-0.05 b	0.18 ab	0.39 ab	0.02 b	0.17 ab	0.09 ab	1.49E-94
SLC4A11	-0.37 b	0.14 a	0.02 ab	-0.09 ab	0.04 ab	0.03 ab	-0.02 ab	-0.13 ab	5.10E-13
C20orf194	-0.56 ab	0.12 a	0.02 a	0.12 a	0.16 a	0.15 a	0.06 a	-0.07 ab	2.36E-28
ATRN	-0.74 c	0.36 a	-0.05 b	0.17 ab	0.17 ab	0.32 a	0.31 a	0.13 ab	2.33E-78
ADAM33	-1.83 e	1.13 a	-0.09 d	0.03 cd	0.74 ab	0.24 bc	0.24 bc	0.33 bc	0
HSPA12B	-1.74 d	1.11 a	-0.10 c	0.08 bc	0.73 ab	0.22 b	0.26 b	0.37 b	0

Table B.1 (cont'd)

C20orf27	-1.71 e	1.13 a	-0.11 d	0.07 cd	0.81 ab	0.23 bc	0.25 bc	0.32 bc	0
SPEF1	-0.36 b	0.23 a	0.00 b	0.03 ab	0.11 ab	-0.04 b	-0.03 b	-0.05 b	1.58E-16
CENPB	-0.88 c	0.57 a	-0.04 b	0.07 b	0.37 ab	0.06 b	0.06 b	0.16 ab	2.18E-121
CDC25B	-0.27 ab	0.45 a	-0.10 ab	0.18 ab	0.49 a	0.07 ab	0.14 ab	0.15 ab	4.42E-46
AP5S1	-0.87 c	0.60 a	-0.05 b	-0.01 b	0.47 ab	0.03 b	0.10 b	0.04 b	6.14E-128
MAVS	-0.73 d	0.52 a	-0.06 c	0.00 bc	0.17 abc	0.24 ab	0.17 abc	0.02 bc	1.68E-95
PANK2	-0.94 c	0.49 a	-0.03 b	0.19 ab	0.38 ab	0.01 b	0.19 ab	0.18 ab	1.77E-118
RNF24	-0.53 b	0.30 a	-0.05 b	0.15 ab	0.24 ab	0.21 ab	0.22 ab	0.06 ab	6.54E-42
SMOX	-1.79 e	1.14 a	-0.10 d	-0.01 cd	0.75 ab	0.23 bc	0.23 bc	0.30 bc	0
PRNP	-0.46 ab	0.17 a	0.01 ab	0.13 a	0.31 a	-0.04 ab	0.04 a	0.01 ab	6.22E-21
PRND	-1.79 e	1.13 a	-0.10 d	0.04 cd	0.77 ab	0.27 bc	0.21 bc	0.33 bc	0
SLC23A2	-0.54 b	0.18 a	-0.01 ab	0.16 ab	0.16 ab	0.22 a	0.04 ab	0.06 ab	2.17E-30
TMEM230	-0.87 c	0.58 a	-0.05 b	0.19 ab	0.18 ab	0.13 b	0.07 b	0.04 b	5.73E-124
PCNA	-0.48 b	0.50 a	-0.12 a	0.31 a	0.58 a	0.17 a	0.22 a	0.35 a	3.74E-78
CDS2	-0.82 c	0.51 a	-0.06 b	0.11 b	0.42 ab	0.28 ab	0.12 b	0.09 b	1.04E-108
PROKR2	-1.82 e	1.15 a	-0.10 d	0.05 cd	0.76 ab	0.27 bc	0.22 bc	0.27 bc	0
GPCPD1	-0.48 c	0.32 a	0.00 b	0.11 ab	-0.09 bc	-0.11 bc	-0.17 bc	0.05 ab	2.83E-35
C20orf196	-0.63 c	0.41 a	0.01 b	-0.08 b	-0.09 bc	-0.15 bc	-0.13 bc	0.02 ab	1.42E-59
CHGB	-0.24 ab	0.04 ab	0.05 a	-0.20 ab	-0.24 ab	-0.12 ab	-0.12 ab	0.05 a	1.59E-07
TRMT6	-0.65 c	0.63 a	-0.13 b	0.47 a	0.66 a	0.14 ab	0.23 a	0.37 a	2.67E-127
MCM8	-0.45 b	0.48 a	-0.14 a	0.39 a	0.53 a	0.29 a	0.28 a	0.33 a	3.12E-81
CRLS1	-0.74 c	0.56 a	-0.04 b	0.09 b	0.16 ab	-0.05 b	0.13 b	-0.06 b	3.67E-99

Table B.1 (cont'd)

LRRN4	-1.82 e	1.13 a	-0.09 d	0.02 cd	0.73 ab	0.25 bc	0.21 bc	0.25 bc	0
FERMT1	-1.71 d	1.08 a	-0.09 c	0.08 bc	0.70 ab	0.22 b	0.22 b	0.21 bc	0
BMP2	-0.31 b	0.14 a	0.03 ab	0.10 ab	0.07 ab	-0.23 b	-0.06 ab	-0.02 ab	6.35E-13
TMX4	-0.61 c	0.28 a	0.02 b	0.10 ab	-0.17 bc	-0.14 b	0.02 ab	0.03 ab	1.08E-42
PLCB1	-0.37 b	0.16 a	0.00 a	0.17 a	-0.16 ab	-0.01 ab	0.08 a	-0.03 ab	3.30E-14
PLCB4	-0.16 a	0.10 a	0.04 a	-0.06 a	-0.68 a	-0.22 a	-0.17 a	-0.09 a	6.38E-11
ANKEF1	-0.48 c	0.41 a	-0.08 bc	0.28 ab	0.17 abc	0.30 a	0.10 abc	0.06 abc	3.15E-56
SNAP25	-0.22 ab	0.05 a	0.04 a	0.12 a	-0.04 ab	-0.08 ab	-0.21 ab	-0.07 ab	3.55E-06
MKKS	-0.92 c	0.59 a	-0.06 b	0.20 ab	0.30 ab	0.09 b	0.19 b	0.22 ab	3.32E-139
SLX4IP	-0.43 c	0.31 a	-0.07 bc	0.29 ab	0.22 abc	0.32 a	0.18 abc	0.08 abc	4.54E-41
JAG1	-0.42 c	0.22 a	-0.02 bc	0.19 ab	0.16 abc	0.13 abc	-0.07 bc	0.13 abc	1.37E-22
BTBD3	-0.58 c	0.30 a	-0.02 bc	0.28 ab	0.14 abc	-0.03 bc	0.16 ab	0.12 abc	2.76E-44
SPTLC3	-0.44 b	0.10 a	0.06 a	0.16 a	-0.20 ab	-0.24 ab	-0.16 ab	-0.03 ab	4.76E-23
TASP1	-0.75 c	0.49 a	-0.05 b	0.26 ab	0.29 ab	0.08 b	0.11 b	0.08 b	2.45E-89
ESF1	-0.58 c	0.49 a	-0.10 b	0.32 ab	0.45 ab	0.20 ab	0.18 ab	0.21 ab	1.21E-80
NDUFAF5	-0.81 c	0.59 a	-0.06 b	0.21 ab	0.45 ab	0.00 b	0.08 b	0.18 ab	2.31E-117
MACROD2	-0.36 b	0.10 a	0.06 ab	0.06 ab	-0.09 ab	-0.22 ab	-0.12 ab	-0.20 ab	4.52E-17
FLRT3	-0.35 b	0.15 a	0.02 ab	0.20 a	0.06 ab	-0.11 ab	-0.15 ab	0.14 ab	8.51E-15
KIF16B	-0.53 c	0.39 a	-0.07 bc	0.22 ab	0.18 abc	0.25 a	0.13 abc	0.18 ab	1.10E-55
SNRPB2	-0.74 c	0.67 a	-0.10 b	0.26 ab	0.62 ab	0.00 b	0.17 b	0.24 ab	1.35E-133
BFSP1	-0.37 ac	0.35 a	-0.07 abc	0.26 ab	0.40 a	-0.03 abc	0.23 ab	0.26 ab	3.55E-38
DSTN	-0.66 c	0.36 a	-0.02 b	0.13 ab	0.18 ab	0.04 ab	0.12 ab	0.08 ab	1.05E-57

Table B.1 (cont'd)

RRBP1	-1.72 e	1.09 a	-0.09 d	0.07 cd	0.78 ab	0.22 bc	0.20 bc	0.26 bc	0
SNX5	-0.79 c	0.46 a	-0.02 b	0.06 b	0.30 ab	0.07 b	0.11 ab	-0.13 b	7.95E-87
SNORD17	-0.42 b	0.25 a	-0.04 ab	0.10 a	0.41 a	0.18 a	0.18 a	0.01 ab	1.50E-28
MGME1	-0.68 c	0.58 a	-0.13 b	0.41 a	0.71 a	0.25 ab	0.31 a	0.25 ab	9.72E-124
ZNF133	-0.81 c	0.55 a	-0.07 b	0.22 ab	0.53 ab	0.11 b	0.15 b	0.15 ab	5.22E-113
DZANK1	-0.51 c	0.42 a	-0.04 bc	0.14 ab	0.07 abc	0.01 bc	-0.10 bc	0.09 ab	5.46E-50
POLR3F	-0.75 c	0.60 a	-0.09 b	0.32 ab	0.45 ab	0.09 b	0.15 b	0.23 ab	1.19E-117
RBBP9	-0.74 c	0.43 a	-0.02 b	0.25 ab	0.10 ab	-0.03 b	0.02 b	0.06 ab	2.23E-76
SEC23B	-0.63 c	0.37 a	-0.03 bc	0.06 ab	0.02 abc	0.14 ab	0.08 ab	0.08 ab	1.19E-55
SMIM26	-0.90 c	0.58 a	-0.03 b	0.09 b	0.26 ab	-0.02 b	0.05 b	0.01 b	5.27E-126
DTD1	-0.58 c	0.46 a	-0.04 b	0.12 ab	0.26 ab	-0.18 b	0.16 ab	0.19 ab	1.65E-66
LINC00652	-0.16 b	0.30 a	-0.04 b	-0.04 b	-0.05 b	-0.07 b	-0.11 b	0.07 ab	1.24E-15
SLC24A3	-0.19 ab	0.09 a	0.03 ab	-0.04 ab	-0.02 ab	-0.01 ab	-0.30 b	-0.15 ab	2.92E-07
RIN2	-0.46 b	0.31 a	-0.04 b	0.15 ab	0.23 ab	0.05 ab	0.11 ab	0.07 ab	1.21E-32
NAA20	-0.61 c	0.60 a	-0.13 b	0.25 b	0.53 ab	0.27 b	0.22 b	0.22 b	3.61E-110
CRNKL1	-0.73 c	0.50 a	-0.07 b	0.23 ab	0.15 ab	0.20 ab	0.15 ab	0.15 ab	5.95E-93
DEFB127	-0.10 ab	0.10 ab	-0.07 ab	0.34 a	0.13 ab	0.31 a	0.19 ab	0.05 ab	-7.80E-11
C20orf97	-0.58 b	0.36 a	0.00 b	0.02 ab	0.08 ab	-0.08 b	-0.07 b	-0.12 b	-7.88E-11
ZCCHC4	-0.86 c	0.52 a	-0.04 b	0.18 ab	0.36 ab	-0.03 b	0.20 ab	0.09 ab	-7.96E-11
SOX13	-0.53 c	0.35 a	-0.07 b	0.09 ab	0.35 ab	0.24 ab	0.24 ab	0.17 ab	-8.03E-11
NRSN3	-0.54 c	0.28 a	-0.03 b	0.05 ab	0.26 ab	0.12 ab	0.17 ab	0.17 ab	-8.11E-11
TRIB4	-0.14 ab	0.33 a	-0.10 ab	-0.12 ab	0.44 a	0.16 ab	0.26 a	0.25 ab	-8.19E-11

Table B.1 (cont'd)

RBCK2	-0.80 c	0.62 a	-0.06 b	0.05 b	0.36 ab	-0.01 b	0.13 b	0.09 b	-8.27E-11
TBC1D21	-0.91 c	0.51 a	-0.04 b	0.11 b	0.21 ab	0.14 b	0.17 ab	0.09 b	-8.34E-11
CSNK2A2	-0.77 d	0.58 a	-0.10 c	0.24 abc	0.26 abc	0.21 bc	0.28 ab	0.27 abc	-8.42E-11
SRXN2	-0.44 b	0.39 a	-0.10 b	0.08 ab	0.31 ab	0.25 ab	0.42 a	0.21 ab	-8.50E-11
SLC52A4	-0.24 c	0.22 a	-0.04 bc	0.06 abc	0.16 abc	0.17 ab	0.07 abc	-0.14 bc	-8.57E-11
FAM110A	-0.51 b	0.39 a	-0.07 a	0.13 a	0.49 a	0.22 a	0.10 a	0.23 a	-8.65E-11
PSMF2	-1.05 c	0.67 a	-0.07 b	0.11 b	0.44 ab	0.20 b	0.27 b	0.11 b	-8.73E-11
TMEM74B	-0.12 ab	0.29 a	-0.05 ab	0.07 ab	0.32 a	-0.08 ab	0.06 ab	-0.10 ab	-8.81E-11
C20orf203	-0.18 ab	0.21 a	-0.02 ab	-0.20 ab	0.21 a	-0.05 ab	0.08 ab	-0.05 ab	-8.88E-11
SNPH	-1.77 e	1.12 a	-0.09 d	0.04 cd	0.78 ab	0.21 bc	0.24 bc	0.26 bc	-9E-11
SDCBP3	-0.33 b	0.17 a	0.00 ab	0.15 ab	0.14 ab	-0.10 ab	0.10 ab	0.01 ab	-9.04E-11
FKBP1A	-0.76 b	0.44 a	-0.03 a	0.05 a	0.48 a	0.03 a	0.22 a	0.11 a	-9.12E-11
NSFL1C	-1.02 c	0.64 a	-0.08 b	0.27 ab	0.56 ab	0.15 b	0.22 b	0.28 ab	-9.19E-11
SIRPA	-0.25 b	0.13 a	0.00 ab	0.12 ab	0.02 ab	-0.04 ab	-0.01 ab	-0.11 ab	-9.27E-11
MUC6	-0.11 b	0.15 ab	-0.06 b	0.43 a	0.17 ab	0.10 ab	-0.04 ab	0.32 ab	1.99E-12
KAT14	-0.83 c	0.50 a	-0.03 b	0.04 b	0.20 ab	0.08 b	0.10 b	0.14 ab	4.33E-101
LOC653566	-0.02 b	-0.03 b	-0.01 b	0.05 b	-0.29 bc	0.67 a	0.03 b	-1.04 c	2.59E-53
LOC100270804	-0.42 b	0.37 a	-0.04 b	0.13 ab	0.28 ab	0.03 b	-0.04 b	-0.05 b	2.85E-36
LOC100130987	0.01 b	-0.10 b	-0.01 b	0.10 ab	-0.27 b	0.54 a	-0.03 b	-0.49 b	2.12E-23
ProSAPiP1	-0.35 b	0.27 a	0.00 b	0.09 ab	-0.10 b	-0.16 b	0.06 ab	-0.25 b	9.78E-23
LOC645332	-0.11 bc	0.00 b	0.00 b	-0.08 bc	-0.40 bc	0.50 a	-0.06 bc	-0.53 c	2.62E-22
LOC100289673	-0.12 ab	0.12 ab	-0.07 ab	0.27 a	0.50 a	0.24 a	0.16 ab	0.23 ab	3.90E-14

Table B.1 (cont'd)

LOC730101	-0.06 ab	-0.08 ab	0.06 ab	0.28 a	0.23 ab	-0.10 ab	-0.47 b	-0.15 ab	4.63E-14
ATPGD1	-0.04 b	0.09 ab	-0.07 b	0.32 ab	0.27 ab	0.36 a	0.02 ab	0.09 ab	1.63E-13
WASH5P	-0.25 ab	0.21 a	-0.03 ab	0.23 a	0.17 ab	-0.04 ab	0.01 ab	0.04 ab	3.19E-12
C20orf46	-0.06 ab	0.17 a	-0.07 ab	0.29 a	-0.04 ab	0.19 a	0.10 ab	0.21 a	7.38E-12
C19orf22	-0.08 b	0.01 b	-0.04 b	0.30 ab	-0.05 b	0.38 a	0.00 b	-0.07 b	2.20E-10
SUV420H1	-0.13 b	-0.11 b	0.04 ab	-0.06 ab	-0.39 b	0.23 a	-0.01 ab	-0.38 b	2.32E-10
LINC01512	-0.02 ab	-0.01 ab	0.04 a	-0.02 ab	0.02 ab	-0.02 ab	-0.53 b	0.04 a	2.49E-10
LOC149837	-0.22 b	0.15 a	0.02 ab	-0.16 ab	-0.14 ab	0.00 ab	-0.14 ab	-0.16 ab	3.95E-08
SINHCAF	0.08 a	0.09 a	-0.04 a	0.40 a	-0.37 a	-0.11 a	0.17 a	0.05 a	6.33E-08
LOC100134868	-0.21 b	0.17 a	-0.01 ab	0.13 ab	0.10 ab	-0.10 ab	0.01 ab	-0.02 ab	2.18E-07
LOC642846	0.11 a	0.05 a	-0.07 a	0.25 a	0.10 a	0.18 a	0.11 a	0.11 a	1.25E-06
KIAA1026	0.10 a	-0.03 a	0.03 a	0.09 a	-0.03 a	-0.22 a	-0.29 a	0.01 a	2.81E-06
LOC374443	-0.09 b	-0.07 b	0.00 b	0.05 ab	-0.28 b	0.31 a	0.00 b	-0.02 b	6.53E-06

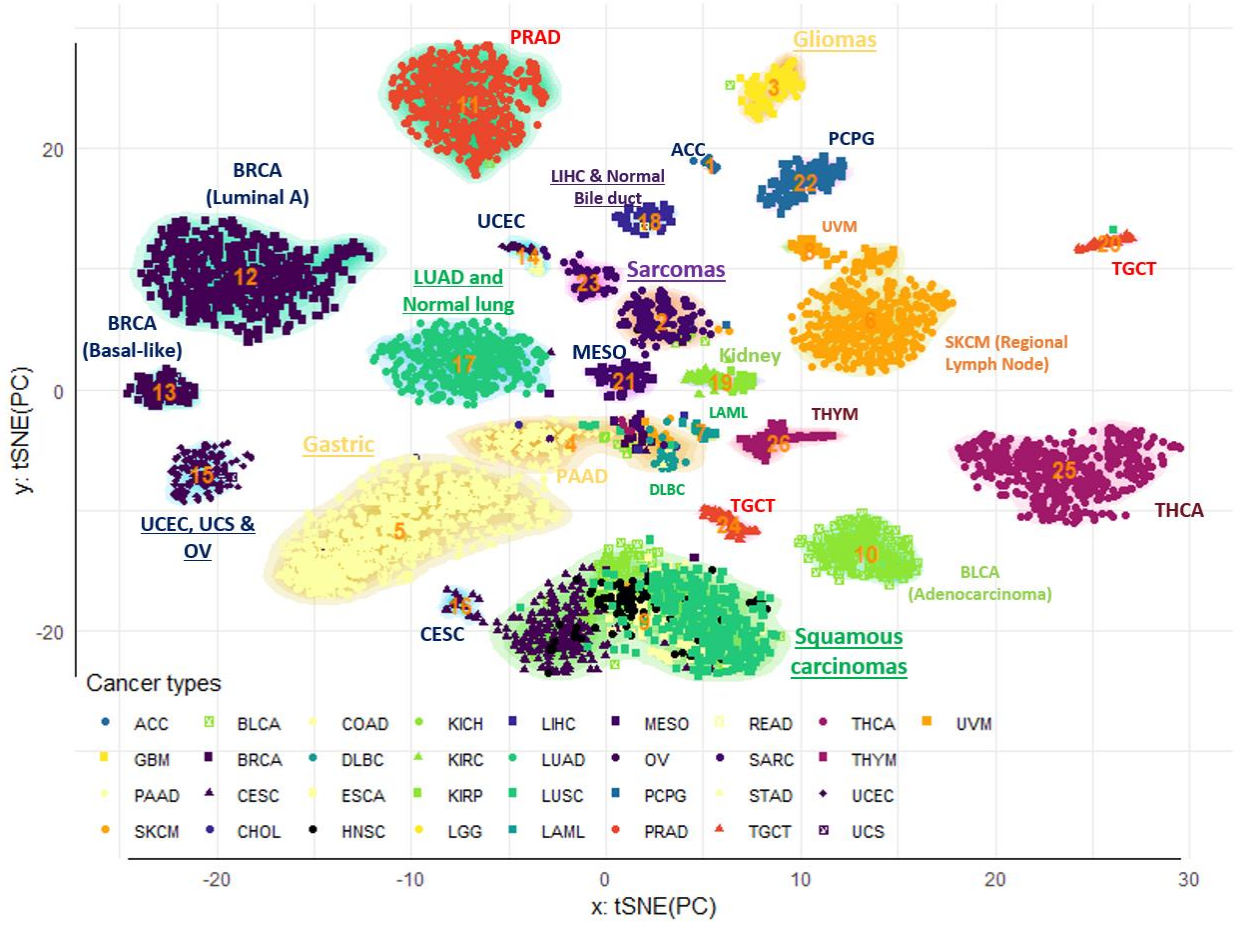

Figure B.1: Clustering of tumor samples (no constraints). Tumor clusters were obtained by sequential application of tSNE and DBSCAN algorithm for 5,408 samples from 33 cancer types. The contours reflect cluster membership, and the points' colors and shapes represent similar anatomical sites and cancer types, respectively. After removing the first two, the two-dimensional tSNE projection was obtained from the first 50 principal axes of the extended omic matrix. Extended omic matrix contained appended values of gene expression, DNA methylation, and copy number variant intensity. Integers represent individual clusters. Clusters were also annotated in terms of their most enriched histological/molecular subtypes.

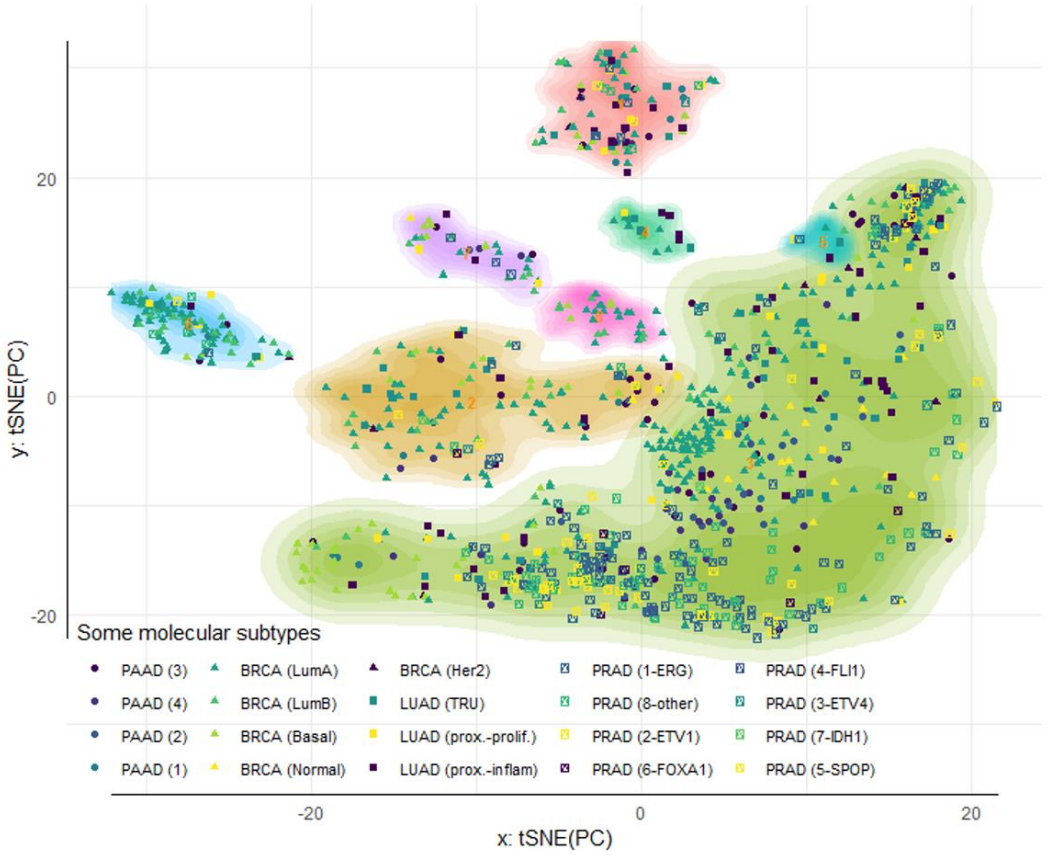

Figure B.2: Re-classification of tumors and previously reported molecular subtypes.

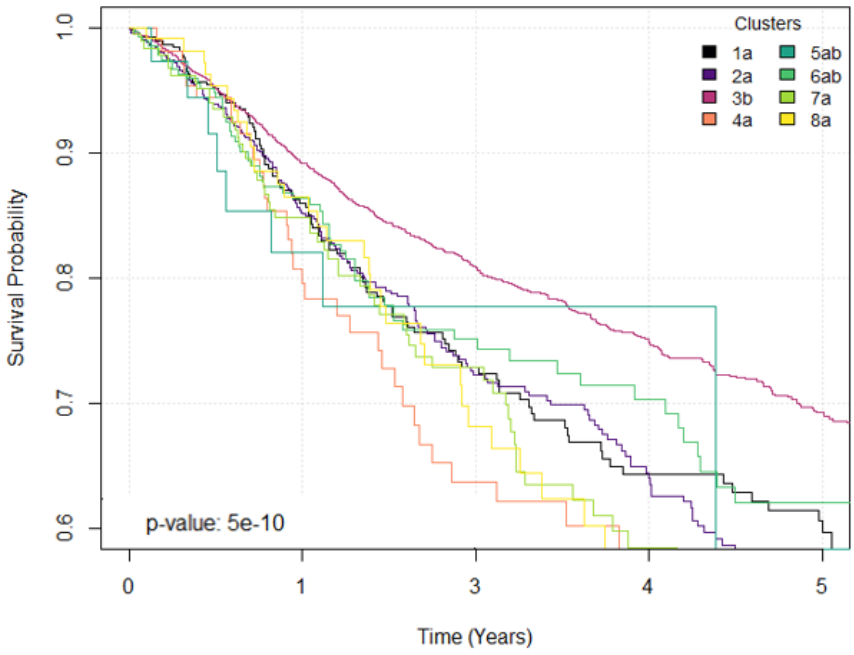

Figure B.3: Survival curves by pan-cancer tumor clusters. The figure shows Kaplan-Meier curves highlighting the survival probability by time in years for each cluster. Log-rank tests were performed to determined significant differences between curves. The legend shows the results of multiple comparisons between survival curves. Statistical differences are represented with different letters.

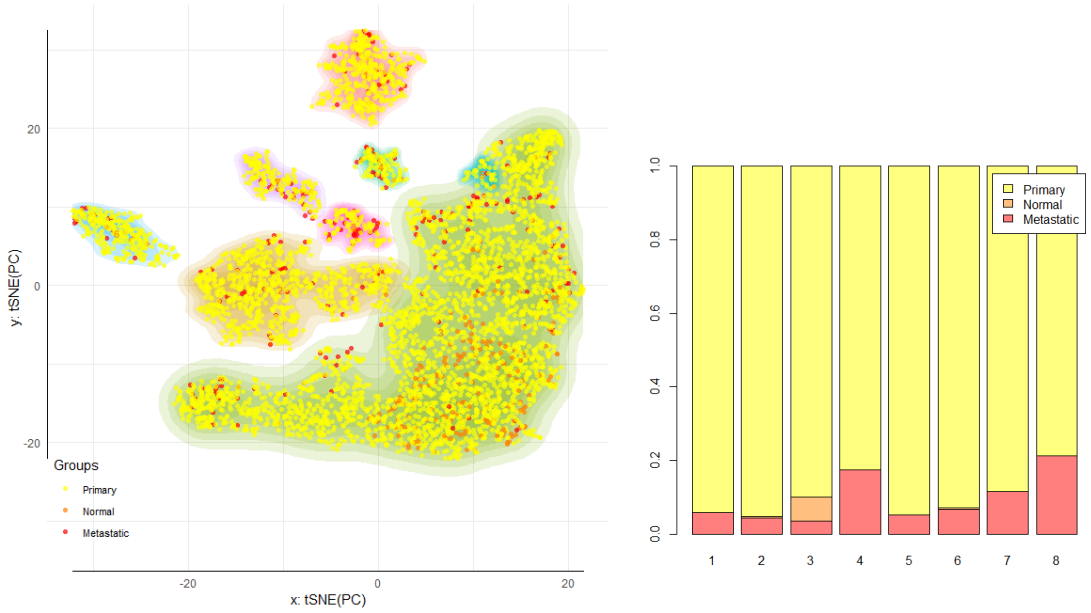

Figure B.4: Re-classification of tumors reveals differences in sample type. The relative position and number of primary, normal, and metastatic tissue samples are shown. The figure at the left shows the location of the samples by clusters. The figure at the right shows the relative proportion of sample types by cluster.

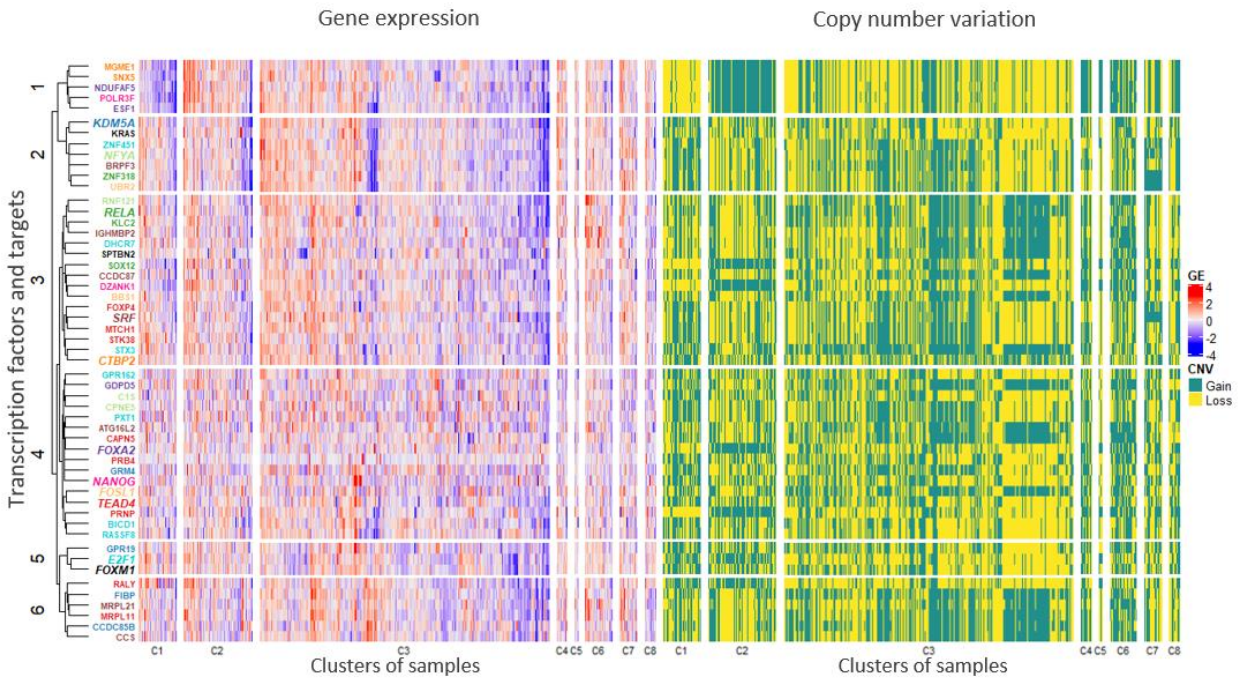

Figure B.5: Expression and copy numbers for transcription factors and targets. The expression and copy number variation by genes is shown by cluster (C1-8). The colors by gene name represent groups defined by different transcription factors and their targets (e.g., black represents the group of *FOXM1* and its targets *KRAS* and *SPTBN2*). TFs names are shown with italic and larger font sizes. The number at the left of the dendrogram represents a grouping of genes based on k-means clustering.

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

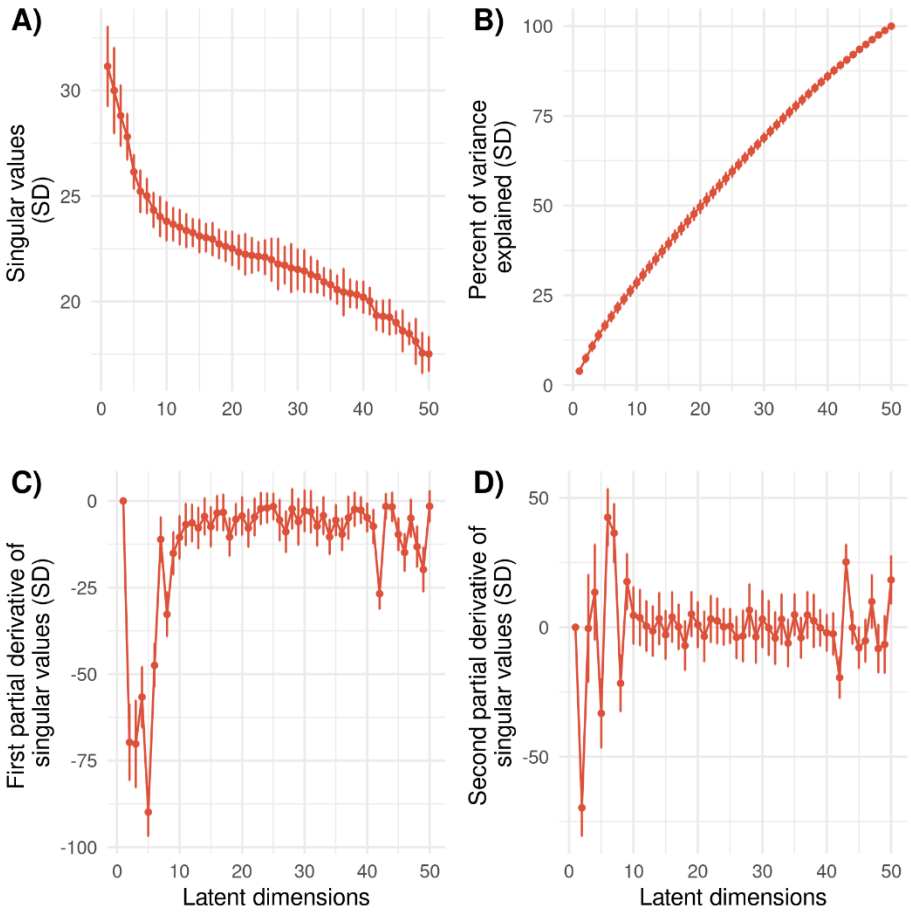

Figure C.1: Plots of singular values trajectories. A) Scree plot by latent SVD dimension. B) Cumulative proportion of variance explained by latent dimension. C) First empirical partial derivative of singular values by latent dimension. **D) Second empirical partial derivative of singular values by latent dimension.** The points and bars in each panel represent average and standard deviations (SD) from $1x10^5$ bootstrap repetitions of the sparse SVD applied to the matrix of phenotypic variables.

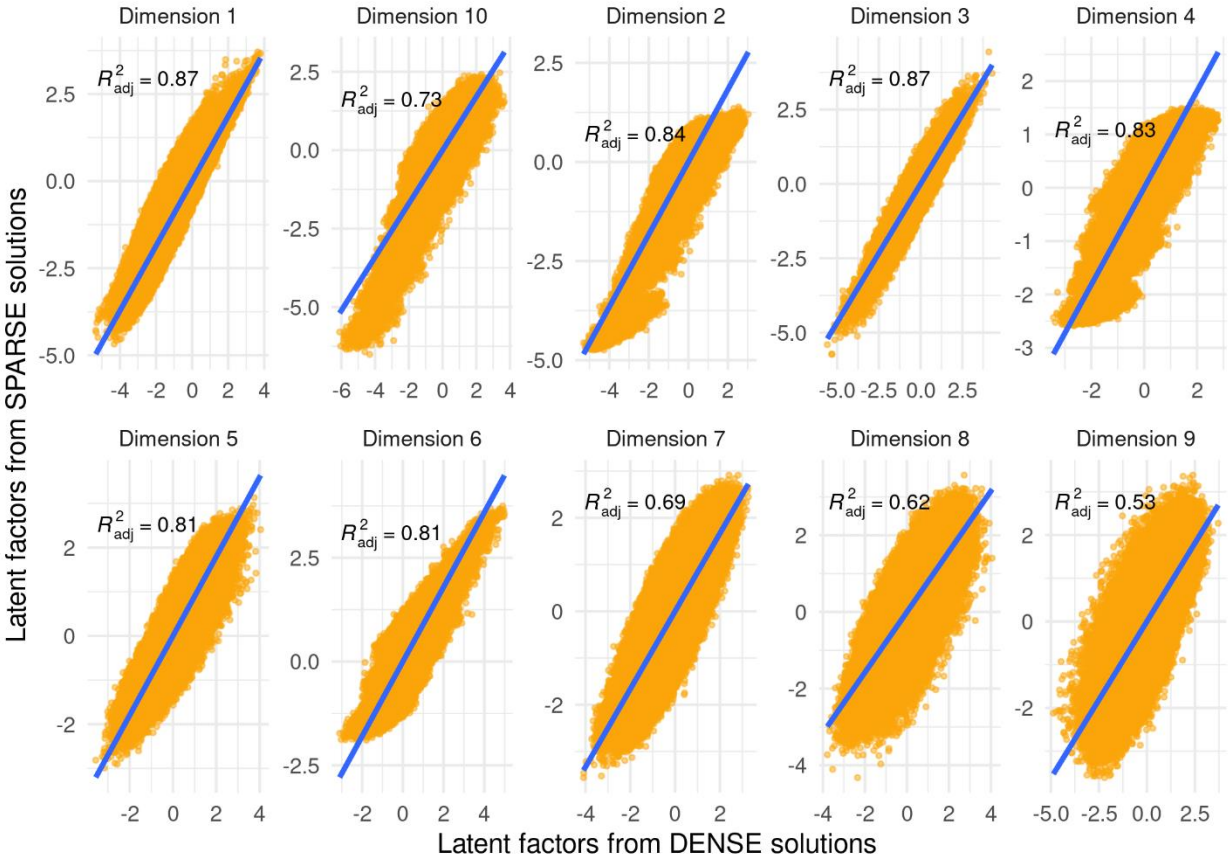

Figure C.2: Comparison between dense and sparse latent factors

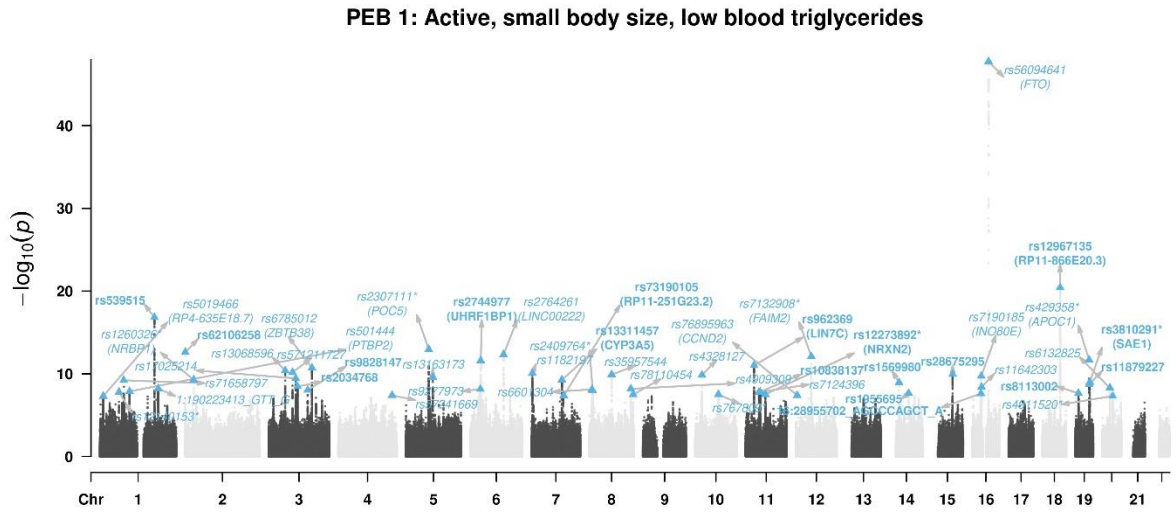

Figure C.3: Annotated Manhattan plot for PEB 1

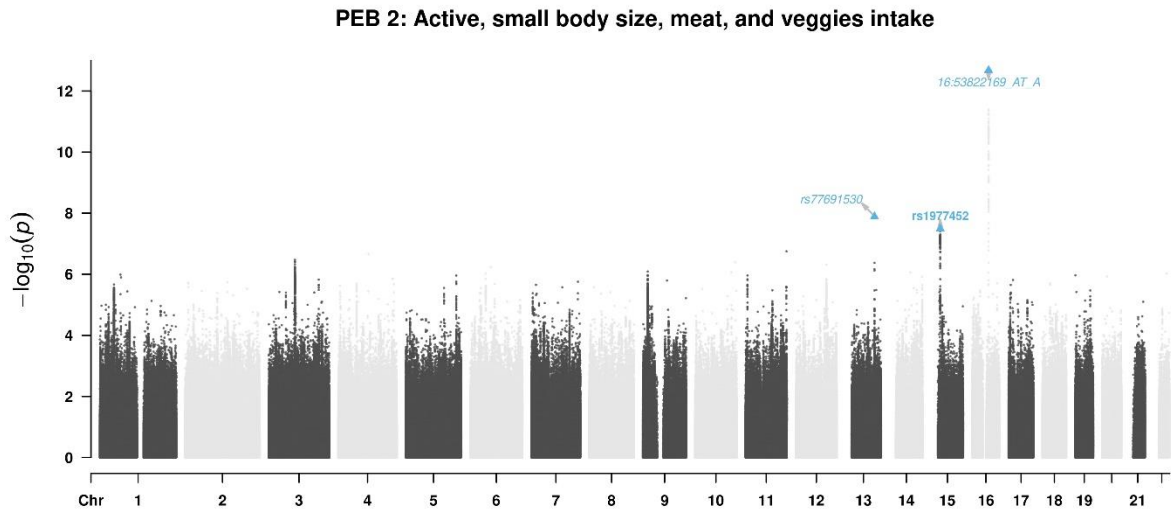

Figure C.4: Annotated Manhattan plot for PEB 2

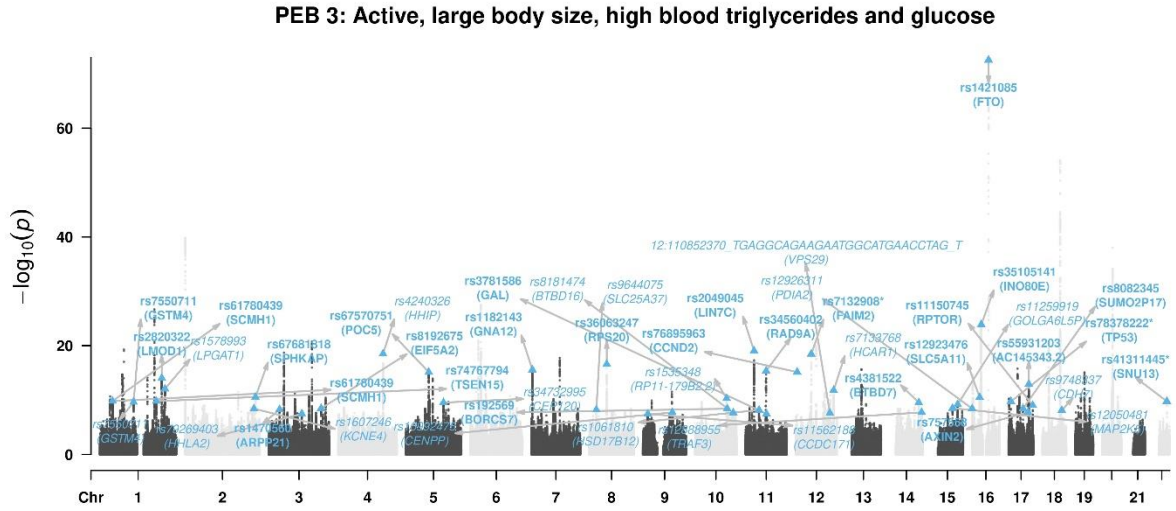

Figure C.5: Annotated Manhattan plot for PEB 3

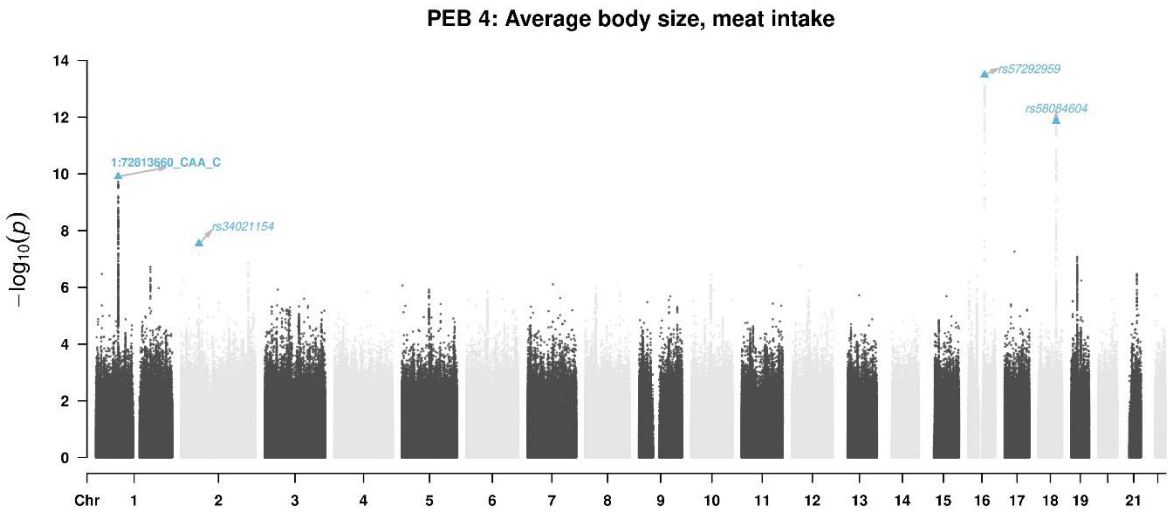

Figure C.6: Annotated Manhattan plot for PEB 4

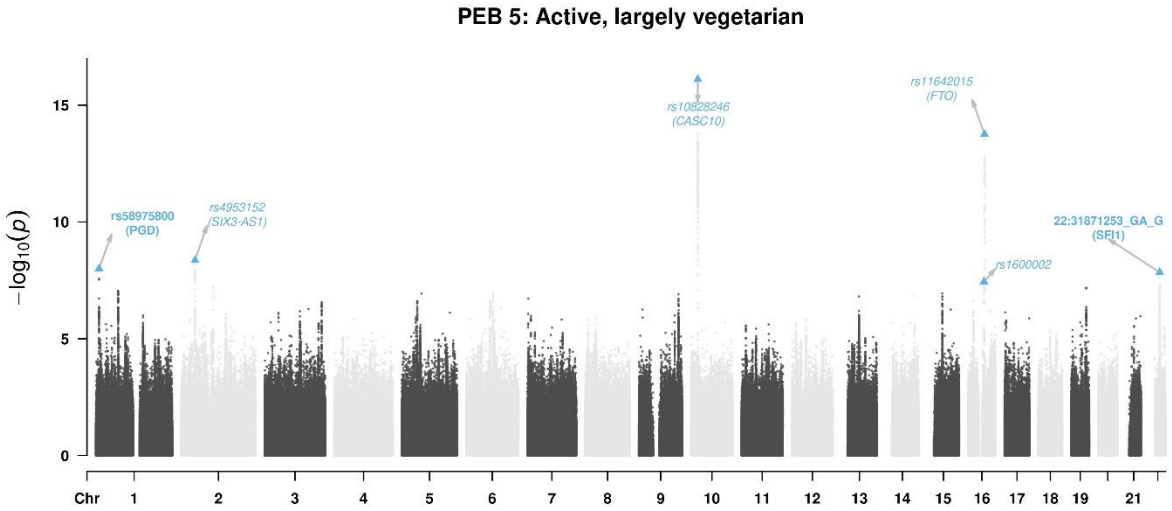

Figure C.7: Annotated Manhattan plot for PEB 5

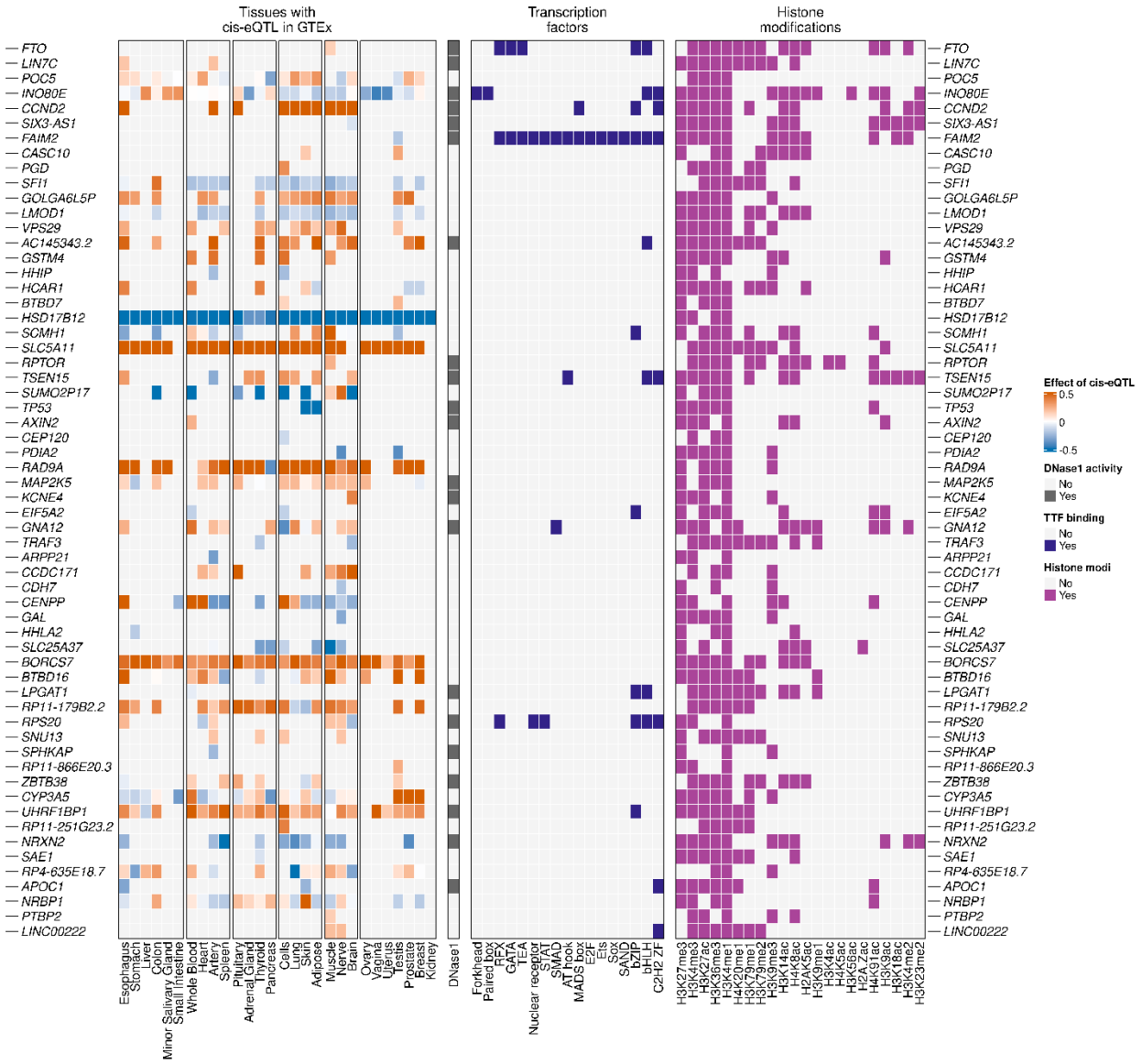

Figure C.8: Summary of previously reported information for genes in Figure 4.4. Each row in the heatmap corresponds to gene mapping onto a significant peak from the GWAS on each PEB variable. The first sets of panels summarize the average effects of significant GWAS peaks with cis-eQTL in GTEx. The color of each cell represents the magnitude and sign of the cis-eQTL effect on the expression of that gene in each of a set of tissues. The order of the subpanels corresponds to a broader aggrupation of tissues in main organs and systems: digestive, reproductive-urinary, neuro-muscular; skin, and adipose tissues, circulatory and glandular. The following panels correspond to results from epigenomic experiments conducted in previous studies. The second panel shows where the peak was in a zone of positive DNase1 activity (as captured by ATAC-seq assays). The Third panel shows whether transcription factors were bound to the region (as captured by CHIP-seq assays. Each column represents a broad classification of transcription factors within families representing binding motifs. The last panel shows the presence of evidence for histone modifications (as captured by CHIP-seq). Each column represents a different histone mark.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Vailati-Riboni M, Palombo V, Loor JJ. What Are Omics Sciences? *Periparturient Dis Dairy Cows A Syst Biol Approach*. 2017; 1–7.

2. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet*. 2018;19: 299–310.

3. Zelenin A V., Rodionov A V., Bolsheva NL, Badaeva ED, Muravenko O V. Genome: Origins and evolution of the term. *Mol Biol*. 2016;50: 542–550.

4. Yadav SP. The wholeness in suffix -omics, -omes, and the word om. *J Biomol Tech*. 2007;18: 277.

5. Eddy SR. The C-value paradox, junk DNA and ENCODE. *Curr Biol*. 2012;22: R898-9.

6. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74: 5463–7.

7. Saiki R, Gelfand D, Stoffel S, Scharf S, Higuchi R, Horn G, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (80-)*. 1988;239: 487–491.

8. Cohen SN, Chang AC, Boyer HW, Helling RB. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A*. 1973;70: 3240–4.

9. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986;321: 674–679.

10. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Zhang H, et al. The International HapMap Project. *Nature*. 2003;426: 789–796.

11. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (80-)*. 2004;306: 636–640.

12. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107: 1–8.

13. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature*. 2015;526: 68–74.

14. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562: 203–209.

15. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995;270: 484–7.

16. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLOS Comput Biol*. 2017;13: e1005457.

17. Eom EM, Lee JY, Park HS, Byun YJ, Ha-Lee YM, Lee DH. Comparison between SAGE and cDNA microarray for quantitative accuracy in transcript profiling analyses. *J Plant Biol*. 2006;49: 498–506.

18. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10: 57–63.

19. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez J-C, et al. From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. *Nat Biotechnol*. 1996;14: 61–65.

20. Tomizaki K, Usui K, Mihara H. Protein-protein interactions and selection: array-based techniques for screening disease-associated biomarkers in predictive/early diagnosis. *FEBS J*. 2010;277: 1996–2005.

21. Tollefsbol TO. Advances in epigenetic technology. *Methods Mol Biol*. 2011;791: 1–10.

22. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. *Nat Rev Genet 2010 1112*. 2010;11: 855–866.

23. Morgan H, Beck T, Blake A, Gates H, Adams N, Debouzy G, et al. EuroPhenome: a repository for high-throughput mouse phenotyping data. *Nucleic Acids Res*. 2010;38: D577–D585.

24. Pieruschka R, Schurr U. Plant phenotyping: Past, present, and future. *Plant Phenomics*. 2019;2019.

25. Poldrack RA, Congdon E, Triplett W, Gorgolewski KJ, Karlsgodt KH, Mumford JA, et al. A phenome-wide examination of neural and cognitive function. *Sci Data 2016 31*. 2016;3: 1–12.

26. Vyssotski AL, Serkov AN, Itskov PM, Dell'Omo G, Latanov A V., Wolfer DP, et al. Miniature neurologgers for flying pigeons: Multichannel EEG and action and field potentials in combination with GPS recording. *J Neurophysiol*. 2006;95: 1263–1273.

27. Montes JM, Melchinger AE, Reif JC. Novel throughput phenotyping platforms in plant genetic studies. *Trends Plant Sci*. 2007;12: 433–436.

28. Yugi K, Kubota H, Hatano A, Kuroda S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. *Trends Biotechnol*. 2016;34: 276–290.

29. 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.

30. Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nat Commun*. 2016;7: 13091.

31. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Trans Pattern Anal Mach Intell*. 2007;29: 40–51.

32. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. 2012;40: 9379–9391.

33. 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.

34. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013;7: 523–542.

35. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*. 2014;30: 1370–1376.

36. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. 2012;28: 3290–3297.

37. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics*. 2013;29: 2610–2616.

38. Wang J. Laplacian Eigenmaps. Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. pp. 235–247.

39. Bowen GJ. Isoscapes: Spatial Pattern in Isotopic Biogeochemistry. *Annu Rev Earth Planet Sci*. 2010;38: 161–187.

40. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*. 2000;290: 2323–6.

41. Wahba G. An introduction to reproducing kernel hilbert spaces and why they are so useful. *IFAC Proc Vol*. 2003;36: 525–528.

42. Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 1997. pp. 583–588.

43. 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: 333–337.

44. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*. 2015;31: i268-75.

45. Cun Y, Fröhlich H. Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics. Boccaletti S, editor. *PLoS One*. 2013;8: e73074. \

46. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*. 2014;30: 838–845.

47. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*. 2009;8: Article28.

48. Chen X, Liu H. An Efficient Optimization Algorithm for Structured Sparse CCA, with Applications to eQTL Mapping. *Stat Biosci*. 2012;4: 3–26.

49. Lin D, Zhang J, Li J, Calhoun VD, Deng H-W, Wang Y-P. Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*. 2013;14: 245.

50. Wangen LE, Kowalski BR. A multiblock partial least squares algorithm for investigating complex chemical systems. *J Chemom*. 1989;3: 3–20.

51. Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. 2012;28: 2458–2466.

52. Izenman AJ. Reduced-rank regression for the multivariate linear model. *J Multivar Anal*. 1975;5: 248–264.

53. Vounou M, Nichols TE, Montana G, Alzheimer's Disease Neuroimaging Initiative the ADN. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage*. 2010;53: 1147–59.

54. Goh G, Dey DK, Chen K. Bayesian sparse reduced rank multivariate regression. *J Multivar Anal*. 2017;157: 14–28.

55. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*. 2005;15: 945–53.

56. Tini G, Marchetti L, Priami C, Scott-Boyer M-P. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Brief Bioinform*. 2017 [cited 13 Feb 2019].

57. Chuanchao Zhang, Juan Liu, Qianqian Shi, Xiangtian Yu, Tao Zeng, Luonan Chen. Integration of multiple heterogeneous omics data. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2016. pp. 564–569.

58. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. Schneidman D, editor. *PLOS Comput Biol*. 2017;13: e1005752.

59. González-Reymúndez A, De Los Campos G, Gutiérrez L, Lunt SY, Vazquez AI. Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *Eur J Hum Genet*. 2017;25: 538–544.

60. Vazquez AI, Veturi Y, Behring M, Shrestha S, Kirst M, Resende MFR, et al. Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. *Genetics*. 2016;203: 1425–1438.

61. Vazquez A, Wiener H, Shrestha S, Tiwari H, de los Campos G. Integration of Multi-Layer Omic Data for Prediction of Disease Risk in Humans. Proceedings, 10th World Congress of Genetics Applied to Livestock Production. 2014. p. 6.

62. 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.

63. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology. BioMed Central Ltd.; 2017. pp. 1–15.

64. Müller H, Dagher G, Loibner M, Stumptner C, Kungl P, Zatloukal K. Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management. *Curr Opin Biotechnol*. 2020;65: 45–51.

65. Mangul S, Martin LS, Hill BL, Lam AKM, Distler MG, Zelikovsky A, et al. Systematic benchmarking of omics computational tools. *Nat Commun*. 2019;10: 1–11.

66. Chiroma H, Abdullahi UA, Abdulhamid SM, Abdulsalam Alarood A, Gabralla LA, Rana N, et al. Progress on Artificial Neural Networks for Big Data Analytics: A Survey. *IEEE Access*. 2019;7: 70535–70551.

67. Lorber A, Wangen LE, Kowalski BR. A theoretical foundation for the PLS algorithm. *J Chemom*. 1987;1: 19–31.

68. Zou H, Zou H, Hastie T. Regularization and variable selection via the Elastic Net. *J R Stat Soc Ser B*. 2005;67: 301–320.

69. Shen H, Huang JZ. Sparse principal component analysis via regularized low rank matrix approximation. *J Multivar Anal*. 2008;99: 1015–1034.

70. Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. 1996. pp. 226–231.

71. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20: 53–65.

72. Taskesen E, Huisman SMH, Mahfouz A, Krijthe JH, de Ridder J, van de Stolpe A, et al. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci Rep*. 2016;6: 24949.

73. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *J Mach Learn Res*. 2008;9: 2579–2605.

74. Linderman GC, Steinerberger S. Clustering with t-SNE, provably. *arXiv.org*. 2017 [cited 28 Nov 2018].

75. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*. 2018;34: 2781–2787.

76. Fisher RA. The use of multipole measurements in taxonomic problems. *Ann Eugen*. 1936;7: 179–188.

77. Hahsler M, Piekenbrock M. dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. CRAN; 2017.

78. Bengtsson H. A Unifying Framework for Parallel and Distributed Processing in R using Futures. 2020 [cited 27 May 2021].

79. Wickham H. Ggplot2 : elegant graphics for data analysis. Springer; 2009.

80. Tarazona S, Martinez C. Bioconductor - MOSimMulti-Omics Simulation (MOSim). 2021.

81. Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I, Ramirez RN, Company C, Schmidt A, et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data 2019 61*. 2019;6: 1–15.

82. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11: 367.

83. el Bouhaddani S, Uh HW, Jongbloed G, Hayward C, Klarić L, Kiełbasa SM, et al. Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics*. 2018;19: 371.

84. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32: 2847–2849.

85. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data*. 2019;6: 1–4.

86. Zhang T. On the Consistency of Feature Selection using Greedy Least Squares Regression. *J Mach Learn Res*. 2009.

87. Zhang JM, Harman M, Guedj B, Barr ET, Shawe-Taylor J. Perturbation Validation: A New Heuristic to Validate Machine Learning Models. 2019 [cited 19 Jun 2021].

88. González-Reymúndez A, Vázquez AI. Multi-omic signatures identify pan-cancer classes of tumors beyond tissue of origin. *Sci Rep*. 2020;10: 8341.

89. Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. Translational Implications of Tumor Heterogeneity. *Clin Cancer Res*. 2015;21: 1258–1266.

90. Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499: 214–218.

91. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*. 2013;501: 338–45.

92. Langlands FE, Horgan K, Dodwell DD, Smith L. Breast cancer subtypes: response to radiotherapy and potential radiosensitisation. *Br J Radiol*. 2013;86: 20120601.

93. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168: 613–628.

94. Abdullah LN, Chow EK-H. Mechanisms of chemoresistance in cancer stem cells. *Clin Transl Med*. 2013;2: 3.

95. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45: 1113–1120.

96. Behring M, Shrestha S, Manne U, Cui X, Gonzalez-Reymundez A, Grueneberg A, et al. Integrated landscape of copy number variation and RNA expression associated with nodal metastasis in invasive ductal breast carcinoma. *Oncotarget*. 2018;9: 36836–36848.

97. Vazquez AI, Veturi Y, Behring M, Shrestha S, Kirst M, Resende MF, et al. Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multi-omic Profiles. *Genetics*. 2016; genetics–115.

98. Bernal Rubio YL, González Reymúndez A, Wu K-HH, Griguer CE, Steibel JP, de Los Campos G, et al. Whole-Genome Multi-omic Study of Survival in Patients with Glioblastoma Multiforme. *G3 (Bethesda)*. 2018; g3.200391.2018.

99. González-Reymúndez A, de los Campos G, Gutiérrez L, Lunt SY, Vazquez AI. Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *Eur J Hum Genet*. 2017 [cited 13 Mar 2017].

100. Sánchez-Vega F, Gotea V, Margolin G, Elnitski L. Pan-cancer stratification of solid human epithelial tumors and cancer cell lines reveals commonalities and tissue-specific features of the CpG island methylator phenotype. *Epigenetics Chromatin*. 2015;8.

101. Hoadley KA, Yau C, Stuart JM, Benz CC, Correspondence PWL. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*. 2018;173: 291–304.

102. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*. 2014;158: 929–944.

103. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*. 2016;375: 1109–1112.

104. Zhu Y, Qiu P, Ji Y. TCGA-Assembler: open-source software for retrieving and processing TCGA data. *Nat Methods*. 2014;11: 599–600.

105. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9: 559.

106. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11: 587.

107. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform*. 2013;14: 469–490.

108. Thorsson VV, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, et al. The Immune Landscape of Cancer. *Immunity*. 2018;48: 812-830.e14.

109. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *J Am Stat Assoc*. 1952;47: 583–621.

110. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol Biol*. 2018;1711: 243–259.

111. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J. Evaluation of the la sso and the elastic net in genome-wide association studies. *Front Genet*. 2013;4: 270.

112. Baglama J, Reichel L, Lewis BW. irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. CRAN R project; 2018.

113. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res*. 2002;12: 996–1006.

114. Jawaid W. enrichr: Gene enrichment using Enrichr in enrichR: Provides an R Interface to "Enrichr." CRAN R project; 2017.

115. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015;43: D447–D452.

116. Dunn OJ. Multiple Comparisons Using Rank Sums. *Technometrics*. 1964;6: 241–252.

117. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

118. Krijthe JH. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. CRAN R project; 2015.

119. Worrall SF. TNM Classification of Malignant Tumours. *Br J Oral Maxillofac Surg*. 2000;38: 244.

120. Yang X, Gao L, Zhang S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief Bioinform*. 2016; bbw063.

121. Mishra S, Whetstine JR. Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive. *Mol Cell Biol*. 2016;36: 1050–63.

122. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*. 2013;45: 1134–1140.

123. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. *Hum Mol Genet*. 2009;18: R1-8.

124. Gao Y, Widschwendter M, Teschendorff AE. DNA Methylation Patterns in Normal Tissue Correlate more Strongly with Breast Cancer Status than Copy-Number Variants. *EBioMedicine*. 2018;31: 243–252.

125. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet*. 2018;19: 129–147.

126. Maloney R, Budiman M, Korshunova Y, Monte J, Bacher B, Lakey N, et al. Tissue-specific DNA methylation patterns are frequent targets of epigenetic change in multiple cancer types. *Cancer Res*. 2008;68: LB-256.

127. Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. *Genome Med*. 2014;6: 66.

128. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144: 646–674.

129. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*. 2011;144: 27–40.

130. Shen AL, Moran SA, Glover EA, Drinkwater NR, Swearingen RE, Teixeira LB, et al. Association of a Chromosomal Rearrangement Event with Mouse Posterior Polymorphous Corneal Dystrophy and Alterations in Csrp2bp, Dzank1, and Ovol2 Gene Expression. Anderson MG, editor. *PLoS One*. 2016;11: e0157577.

131. Xu M-D, Liu S-L, Feng Y-Z, Liu Q, Shen M, Zhi Q, et al. Genomic characteristics of pancreatic squamous cell carcinoma, an investigation by using high throughput sequencing after in-solution hybrid capture. *Oncotarget*. 2017;8: 14620–14635.

132. Pei Y-F, Ren H-G, Liu L, Li X, Fang C, Huang Y, et al. Genomic variants at 20p11 associated with body fat mass in the European population. *Obesity*. 2017;25: 757–764.

133. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*. 2007;3: 89.

134. Shah MA, Denton EL, Arrowsmith CH, Lupien M, Schapira M. A global assessment of cancer genomic alterations in epigenetic mechanisms. *Epigenetics Chromatin*. 2014;7: 29.

135. Wanitchakool P, Wolf L, Koehl GE, Sirianant L, Schreiber R, Kulkarni S, et al. Role of anoctamins in cancer and apoptosis. *Philos Trans R Soc B Biol Sci*. 2014;369: 20130096.

136. Ayoub C, Wasylyk C, Li Y, Thomas E, Marisa L, Robé A, et al. ANO1 amplification and expression in HNSCC with a high propensity for future distant metastasis and its functions in HNSCC cell lines. *Br J Cancer*. 2010;103: 715–726.

137. Wang X, Sheu JJ-C, Lai M-T, Yin-Yi Chang C, Sheng X, Wei L, et al. RSF-1 overexpression determines cancer progression and drug resistance in cervical cancer. *BioMedicine*. 2018;8: 4.

138. Sircoulomb F, Bekhouche I, Finetti P, Adélaïde J, Hamida A Ben, Bonansea J, et al. Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer*. 2010;10: 539.

139. Peña-Chilet M, Blanquer-Maceiras M, Ibarrola-Villava M, Martinez-Cadenas C, Martin-Gonzalez M, Gomez-Fernandez C, et al. Genetic variants in PARP1 (rs3219090) and IRF4(rs12203592) genes associated with melanoma susceptibility in a Spanish population. *BMC Cancer*. 2013;13: 160.

140. Hao J-J, Shi Z-Z, Zhao Z-X, Zhang Y, Gong T, Li C-X, et al. Characterization of genetic rearrangements in esophageal squamous carcinoma cell lines by a combination of M-FISH and array-CGH: further confirmation of some split genomic regions in primary tumors. *BMC Cancer*. 2012;12: 367.

141. Chowdhry S, Zanca C, Rajkumar U, Koga T, Diao Y, Raviram R. NAD Metabolic Dependency Determines Therapeutic Sensitivity in Cancer. *Cancer Discov*. 2019;9: OF14–OF14.

142. Kim H-J, Maiti P, Barrientos A. Mitochondrial ribosomes in cancer. *Semin Cancer Biol*. 2017;47: 67–81.

143. Sotgia F, Lisanti MP, Sotgia F, Lisanti MP. Mitochondrial biomarkers predict tumor progression and poor overall survival in gastric cancers: Companion diagnostics for personalized medicine. *Oncotarget*. 2017;8: 67117–67128.

144. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511: 543–550.

145. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*. 2014;511: 543–550.

146. Janku F, Wheler JJ, Westin SN, Moulder SL, Naing A, Tsimberidou AM, et al. PI3K/AKT/mTOR inhibitors in patients with breast and gynecologic malignancies harboring PIK3CA mutations. *J Clin Oncol*. 2012;30: 777–782.

147. Hoelder S, Clarke PA, Workman P. Discovery of small molecule cancer drugs: Successes, challenges and opportunities. Molecular Oncology. John Wiley and Sons Ltd; 2012. pp. 155–176.

148. Bonelli P, Tuccillo FM, Borrelli A, Schiattarella A, Buonaguro FM. CDK/CCN and CDKI alterations for cancer prognosis and therapeutic predictivity. *Biomed Res Int*. 2014;2014.

149. Seo M, Seo M, Goldschmidt-clermont PJ, West M. Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Ann Appl Stat*. [cited 12 Feb 2018].

150. Guo X, Ngo B, Modrek A, Lee W-H. Targeting Tumor Suppressor Networks for Cancer Therapeutics. *Curr Drug Targets*. 2014;15: 2–16.

151. Bray GA, Kim KK, Wilding JPH. Obesity: a chronic relapsing progressive disease process. A position statement of the World Obesity Federation. *Obes Rev*. 2017;18: 715–723.

152. Loos RJF. Recent progress in the genetics of common obesity. *Br J Clin Pharmacol*. 2009;68: 811–829.

153. MO G. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. *lancet Diabetes Endocrinol*. 2018;6: 223–236.

154. Jiang L, Penney KL, Giovannucci E, Kraft P, Wilson KM. A genome-wide association study of energy intake and expenditure. *PLoS One*. 2018;13.

155. Romieu I, Dossus L, Barquera S, Blottière HM, Franks PW, Gunter M, et al. Energy balance and obesity: what are the main drivers? *Cancer Causes Control*. 2017;28: 247–258.

156. Diels S, Berghe W Vanden, Hul W Van. Insights into the multifactorial causation of obesity by integrated genetic and epigenetic analysis. *Obes Rev*. 2020;21: e13019.

157. Hall KD, Heymsfield SB, Kemnitz JW, Klein S, Schoeller DA, Speakman JR. Energy balance and its components: implications for body weight regulation. *Am J Clin Nutr*. 2012;95: 989–994.

158. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14: e8124.

159. Ubbens J, Cieslak M, Prusinkiewicz P, Parkin I, Ebersbach J, Stavness I. Latent space phenotyping: Automatic image-based phenotyping for treatment studies. *Plant Phenomics*. 2020;2020.

160. Joo J, Williamson SA, Vazquez AI, Fernandez JR, Bray MS. Advanced Dietary Patterns Analysis Using Sparse Latent Factor Models in Young Adults. *J Nutr*. 2018;148: 1984–1992.

161. Joo J, Williamson SA, Vazquez AI, Fernandez JR, Bray MS. The influence of 15-week exercise training on dietary patterns among young adults. *Int J Obes*. 2019; 1.

162. Xu SY, Nelson S, Kerr J, Godbole S, Johnson E, Patterson RE, et al. Modeling temporal variation in physical activity using functional principal components analysis. *Stat Biosci*. 2019;11: 403–421.

163. Johnson RK, Vanderlinden L, DeFelice BC, Kechris K, Uusitalo U, Fiehn O, et al. Metabolite-related dietary patterns and the development of islet autoimmunity. *Sci Rep*. 2019;9: 1–11.

164. Guénard F, Bouchard-Mercier A, Rudkowska I, Lemieux S, Couture P, Vohl M-C. Genome-Wide Association Study of Dietary Pattern Scores. *Nutrients*. 2017;9: 649.

165. Tang CS, Ferreira MAR. A gene-based test of association using canonical correlation analysis. *Bioinformatics*. 2012;28: 845–850.

166. Basu S, Zhang Y, Ray D, Miller MB, Iacono WG, McGue M. A rapid gene-based genome-wide association test with multivariate traits. *Hum Hered*. 2014;76: 53–63.

167. Aschard H, Vilhjálmsson BJ, Greliche N, Morange PE, Trégouët DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet*. 2014;94: 662–676.

168. Lightfoot JT, Turner MJ, Daves M, Vordermark A, Kleeberger SR. Genetic influence on daily wheel running activity level. *Physiol Genomics*. 2004;19: 270–276.

169. Koteja P, Garland T, Sax JK, Swallow JG, Carter PA. Behaviour of house mice artificially selected for high levels of voluntary wheel running. *Anim Behav*. 1999;58: 1307–1318.

170. Lerman I, Harrison BC, Freeman K, Hewett TE, Allen DL, Robbins J, et al. Genetic variability in forced and voluntary endurance exercise performance in seven inbred mouse strains. *J Appl Physiol*. 2002;92: 2245–2255.

171. West DB, York B. Dietary fat, genetic predisposition, and obesity: lessons from animal models. *Am J Clin Nutr*. 1998;67: 505S-512S.

172. Ellacott KLJ, Morton GJ, Woods SC, Tso P, Schwartz MW. Assessment of feeding behavior in laboratory mice. *Cell Metab*. 2010;12: 10–17.

173. Barron R, Bermingham K, Brennan L, Gibney ER, Gibney MJ, Ryan MF, et al. Twin metabolomics: the key to unlocking complex phenotypes in nutrition research. *Nutr Res*. 2016;36: 291–304.

174. Rintala M, Lyytikäinen A, Leskinen T, Alen M, Pietiläinen KH, Kaprio J, et al. Leisure-time physical activity and nutrition: a twin study. *Public Health Nutr*. 2011;14: 846–852.

175. Beunen G, Thomis M. Genetic determinants of sports participation and daily physical activity. *Int J Obes*. 1999;23: S55–S63.

176. Maia JAR, Thomis M, Beunen G. Genetic factors in physical activity levels: a twin study. *Am J Prev Med*. 2002;23: 87–91.

177. Lightfoot JT. Current understanding of the genetic basis for physical activity. *J Nutr*. 2011;141: 526–30.

178. Herring MP, Sailors MH, Bray MS. Genetic factors in exercise adoption, adherence and obesity. *Obes Rev*. 2014;15: 29–39.

179. Kwon SM, Cho H, Choi JH, Jee BA, Jo Y, Woo HG. Perspectives of integrative cancer genomics in next generation sequencing era. *Genomics Inform*. 2012;10: 69–73.

180. Rankinen T, Rice T, Teran-Garcia M, Rao DC, Bouchard C. FTO Genotype Is Associated With Exercise Training–induced Changes in Body Composition. *Obesity*. 2010;18: 322–326.

181. Scott RA, Bailey MES, Moran CN, Wilson RH, Fuku N, Tanaka M, et al. FTO genotype and adiposity in children: Physical activity levels influence the effect of the risk genotype in adolescent males. *Eur J Hum Genet*. 2010;18: 1339–1343.

182. Park SL, Cheng I, Pendergrass SA, Kucharska-Newton AM, Lim U, Ambite JL, et al. Association of the FTO obesity risk variant rs8050136 with percentage of energy intake from fat in multiple racial/ethnic populations. *Am J Epidemiol*. 2013;178: 780–790.

183. Wardle J, Carnell S, Haworth CMA, Farooqi IS, O'Rahilly S, Plomin R. Obesity associated genetic variation in FTO is associated with diminished satiety. *J Clin Endocrinol Metab*. 2008;93: 3640–3643.

184. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26: 2867–2873.

185. Bradbury KE, Young HJ, Guo W, Key TJ. Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire. *J Nutr Sci*. 2018;7: 1–11.

186. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars NJ, Aguirre M, Venkataraman GR, et al. Genetics of 38 blood and urine biomarkers in the UK Biobank. bioRxiv. bioRxiv; 2019. p. 660506.

187. Maddison R, Ni Mhurchu C, Jiang Y, Vander Hoorn S, Rodgers A, Lawes CMM, et al. International physical activity questionnaire (IPAQ) and New Zealand physical activity questionnaire (NZPAQ): A doubly labelled water validation. *Int J Behav Nutr Phys Act*. 2007;4: 62.

188. Grueneberg A, de los Campos G. BGData - A suite of R packages for genomic analysis with big data. *G3 Genes, Genomes, Genet*. 2019;9: 1377–1383.

189. Roux B Le, Rouanet H. Geometric data analysis: From correspondence analysis to structured data analysis. Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis. Springer Netherlands; 2005.

190. Gonzalez-Reymundez A, Grueneberg A, Vazquez AI. MOSS: Multi-Omic Integration via Sparse Singular Value Decomposition. *CRAN R-project*. 2021 [cited 19 Jan 2021].

191. Audigier V, Husson F, Josse J. A principal components method to impute missing values for mixed data. *Adv Data Anal Classif*. 2013;10: 5–26.

192. Josse J, Husson F. missMDA: A package for handling missing values in multivariate data analysis. *J Stat Softw*. 2016;70: 1–31.

193. McCaw Z. RNOmni: Rank Normal Transformation Omnibus Test. *CRAN R-project*. 2020 [cited 1 Jul 2021].

194. Aguate FM, Vazquez AI, Merriman TR, de los Campos G. Mapping pleiotropic loci using a fast-sequential testing algorithm. *Eur J Hum Genet*. 2021; 1–12.

195. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*. 2005;21: 3439–3440.

196. Durinck S, Spellman P, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4: 1184–1191.

197. Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31: 3555–3557.

198. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. NIH Public Access; 2013. pp. 580–585.

199. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv.org*. 2018 [cited 14 Jun 2021].

200. Tukey JW. Comparing Individual Means in the Analysis of Variance. *Biometrics*. 1949;5: 99.

201. Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. *J Comput Graph Stat*. 2006;15: 265–286.

202. Bradfield JP, Vogelezang S, Felix JF, Chesi A, Helgeland Ø, Horikoshi M, et al. A trans-ancestral meta-analysis of genome-wide association studies reveals loci associated with childhood obesity. *Hum Mol Genet*. 2019;28: 3327–3338.

203. Liu Y, Zhang X, Lee J, Smelser D, Cade B, Chen H, et al. Genome-wide association study of neck circumference identifies sex-specific loci independent of generalized adiposity. *Int J Obes*. 2021;45: 1532–1541.

204. Lind L. Genome-Wide Association Study of the Metabolic Syndrome in UK Biobank. *Metab Syndr Relat Disord*. 2019;17: 505–511.

205. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet*. 2019;104: 65–75.

206. Graff M, Scott RA, Justice AE, Young KL, Feitosa MF, Barata L, et al. Genome-wide physical activity interactions in adiposity — A meta-analysis of 200,452 adults. *PLoS Genet*. 2017;13: 130.

207. Berndt SI, Gustafsson S, Mägi R, Ganna A, Wheeler E, Feitosa MF, et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet*. 2013;45: 501–512.

208. Trousdale C, Kim K. Retromer: Structure, function, and roles in mammalian disease. European Journal of Cell Biology. Elsevier GmbH; 2015. pp. 513–521.

209. Bentzinger CF, Wang YX, Rudnicki MA. Building muscle: molecular regulation of myogenesis. Cold Spring Harbor perspectives in biology. Cold Spring Harbor Laboratory Press; 2012.

210. Newmire D, Willoughby DS. Wnt and β-Catenin Signaling and Skeletal Muscle Myogenesis in Response to Muscle Damage and Resistance Exercise and Training. *Int J Kinesiol Sport Sci*. 2015;3: 40–49.

211. Wu JHY, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: Results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet*. 2013;6: 171–183.

212. Singla P. Metabolic effects of obesity: A review. *World J Diabetes*. 2010;1: 76.

213. Sherman WM. Metabolism of sugars and physical performance. *Am J Clin Nutr*. 1995;62: 228S-241S.

214. Nagel M, Watanabe K, Stringer S, Posthuma D, Van Der Sluis S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat Commun*. 2018;9.

215. Davies G, Lam M, Harris SE, Trampush JW, Luciano M, Hill WD, et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun*. 2018;9.

216. Hübel C, Gaspar HA, Coleman JRI, Finucane H, Purves KL, Hanscombe KB, et al. Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *Am J Med Genet Part B Neuropsychiatr Genet*. 2019;180: 428–438.

217. Hernandez Cordero AI, Gonzales NM, Parker CC, Sokolof G, Vandenbergh DJ, Cheng R, et al. Genome-wide Associations Reveal Human-Mouse Genetic Convergence and Modifiers of Myogenesis, CPNE1 and STC2. *Am J Hum Genet*. 2019;105: 1222–1236.

218. Nanda V, Wang T, Pjanic M, Liu B, Nguyen T, Matic LP, et al. Functional regulatory mechanism of smooth muscle cell-restricted LMOD1 coronary artery disease locus. *PLoS Genet*. 2018;14: e1007755.

219. Ahmad RS, Imran A, Hussain MB. Nutritional Composition of Meat. Meat Science and Nutrition. InTech; 2018.

220. Goni L, Cuervo M, Milagro FI, Martínez JA. Future perspectives of personalized weight loss interventions based on nutrigenetic, epigenetic, and metagenomic data. *J Nutr*. 2016;146: 905S-912S.

221. Ramos-Lopez O, Milton-Laskibar I, Martínez JA. Precision nutrition based on phenotypical traits and the (epi)genotype: nutrigenetic and nutrigenomic approaches for obesity care. *Curr Opin Clin Nutr Metab Care*. 2021;24: 315–325.

222. Henseler J. On the convergence of the partial least squares path modeling algorithm. *Comput Stat 2009 251*. 2009;25: 107–120.

223. Tipping ME, Bishop CM. Probabilistic Principal Component Analysis. *J R Stat Soc Ser B (Statistical Methodol*. 1999;61: 611–622.

224. Tarantino G, Monica S, Bergenti F. A probabilistic matrix factorization algorithm for approximation of sparse matrices in natural language processing. *ICT Express*. 2018;4: 87–90.

225. Guan Y, Dy JG. Sparse probabilistic principal component analysis. *J Mach Learn Res*. 2009;5: 185–192.

226. Idris SF, Ahmad SS, Scott MA, Vassiliou GS, Hadfield J. The role of high-throughput technologies in clinical cancer genomics. *http://dx.doi.org/101586/erm131*. 2014;13: 167–181.

227. International Cancer Genome Consortium TICG, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. *Nature*. 2010;464: 993–8.

228. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483: 603–7.

229. Heim D, Budczies J, Stenzinger A, Treue D, Hufnagl P, Denkert C, et al. Cancer beyond organ and tissue specificity: Next-generation-sequencing gene mutation data reveal complex genetic similarities across major cancers. *Int J Cancer*. 2014;135: 2362–2369.

230. Yates III JR. A century of mass spectrometry: from atoms to proteomes. *Nat Methods*. 2011;8: 633–637.

231. Wu HM, Goate AM, O'Reilly PF. Heterogeneous effects of genetic risk for Alzheimer's disease on the phenome. *Transl Psychiatry 2021 111*. 2021;11: 1–9.

232. Li X, Meng X, He Y, Spiliopoulou A, Timofeeva M, Wei W-Q, et al. Genetically determined serum urate levels and cardiovascular and other diseases in UK Biobank cohort: A phenome-wide mendelian randomization study. *PLOS Med*. 2019;16: e1002937.

233. Cole JB, Florez JC, Hirschhorn JN. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. *Nat Commun*. 2020;11: 1467.

234. UK Biobank adds the first tranche of data from a study into circulating metabolomic biomarkers to its biomedical database. [cited 10 Aug 2021].
235. Guasch-Ferré M, Bhupathiraju SN, Hu FB. Use of Metabolomics in Improving Assessment of Dietary Intake. *Clin Chem*. 2018;64: 82.

236. The era of massive cancer sequencing projects has reached a turning point. *Nature*. 2020;578: 7–8.