Generative Models of Biological Variations in Bulk and Single-cell RNA-seq

by

Weiguang Mao

Bachelor, Tsinghua University, 2014

Submitted to the Graduate Faculty of
the School of Medicine in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Weiguang Mao

It was defended on

February 21, 2020

and approved by

Maria Chikina, Assistant Professor, Department of Computational and Systems Biology, University of Pittsburgh

Chakra Chennubhotla, Associate Professor, Department of Computational and Systems Biology, University of Pittsburgh

Casey Greene, Associate Professor, Department of Systems Pharmacology and Translational Therapeutics ,University of Pennsylvania

Dennis Kostka, Assistant Professor, Department of Developmental Biology, University of Pittsburgh

Jian Ma, Associate Professor, Computational Biology Department, Carnegie Mellon University

Dissertation Director: Maria Chikina, Assistant Professor, Department of Computational and Systems Biology, University of Pittsburgh

Generative Models of Biological Variations in Bulk and Single-cell RNA-seq

Weiguang Mao, PhD

University of Pittsburgh, 2020

The explosive growth of next-generation sequencing data enhances our ability to understand biological process at an unprecedented resolution. Meanwhile organizing and utilizing this tremendous amount of data becomes a big challenge. High-throughput technology provides us a snapshot of all underlying biological activities, but this kind of extremely high-dimensional data is hard to interpret. Due to the curse of dimensionality, the measurement is sparse and far from enough to shape the actual manifold in the high-dimensional space. On the other hand, the measurements may contain structured noise such as technical or nuisance biological variation which can interfere downstream interpretation. Generative modeling is a powerful tool to make sense of the data and generate compact representations summarizing the embedded biological information. This thesis introduces three generative models that help amplifying biological signals buried in the noisy bulk and single-cell RNA-seq data.

In Chapter 2, we propose a semi-supervised deconvolution framework called PLIER which can identify regulations in cell-type proportions and specific pathways that control gene expression. PLIER has inspired the development of MultiPLIER [154] and has been used to infer context-specific genotype effects in the brain [121].

In Chapter 3, we construct a supervised transformation named DataRemix to normalize bulk gene expression profiles in order to maximize the biological findings with respect to a variety of downstream tasks. By reweighing the contribution of hidden factors, we are able to reveal the hidden biological signals without any external dataset-specific knowledge. We apply DataRemix to the ROSMAP dataset and report the first replicable *trans*-eQTL effect in human brain.

In Chapter 4, we focus on scRNA-seq and introduce NIFA which is an unsupervised decomposition framework that combines the desired properties of PCA, ICA and NMF. It simultaneously models uni- and multi-modal factors isolating discrete cell-type identity and continuous pathway-level variations into separate components.

The work presented in Chapter 2 has been published as a journal article [109]. The work in Chapter 3 and Chapter 4 are under submission and they are available as preprints on bioRxiv [107, 108].

Table of Contents

List of Tables

List of Figures

Preface

First and foremost I would like to thank my advisor Maria Chikina. When I first started my graduate school, I wasn't quite clear about my research direction. Maria gave me a lot of freedom to learn stuff, try different topics and propose my own project. During this slow transition to become well-equipped, she is always supportive. There are ups and downs in this six-year journey. I am always grateful for her patience and encouragement. Maria is always energetic, actively interact with people and share her brilliant ideas. She is definitely a role model for myself.

I would like to acknowledge members of my thesis committee - Dr. Chakra Chennubhotla, Dr. Casey Greene, Dr. Dennis Kostka and Dr. Jian Ma, for their input to my dissertation. Also I would like to thank Dr. Takis Benos and Dr. Wei Chen for being members of my proposal committee.

I have participated joint group meetings with Dr. Takis Benos and Dr. Nathan Clark for years. I learned quite a lot and got a lot of chances to practice my presentation skill. Moreover, it leads to a fruitful collaboration. I want to thank Dr. Takis Benos and Dr. Nathan Clark and all members of Benos Lab and Clark Lab.

I want to acknowledge my colleagues - the Class of 2019 of CPCB program: Dr. Sabrina Rashid, Dr. Xiongtao Ruan, Dr. Raghavendran Partha and Dr. Sanjana Gupta (soon) and Dr. Jocelyn Sunseri (soon). I want to express my gratitude to Ms. Kelly Gentille and Mr. Gengkon Lum. They have done amazing work to make sure everything is on track. I really appreciate their patience and kindness.

I would like to thank all my local friends in Pittsburgh - Fan Tong, Xiao Zhang, Chiyu Dong, Han Zhao, Guannan He, Ruochen Xu, Jin Hu, Mingda Zhang, Qi Guo, Qian Liu, Danyang Li, Han Lai and Zhuo Chen . They have formed a steady network of support. Especially I would like to acknowledge Susu Xu and Rui Liu. They are not only my personal friends but also my great collaborators. They give me chance to try stuff outside my own expertise.

Last but not least, I would like to thank my parents who are 6785.86 miles away from me. Because of the long-standing visa issue, I didn't quite spend a lot of time with them during my graduate school. Although they don't quite know what I have been doing for these years, they always try their best to figure out ways to help me. I also want to thank my girlfriend Shuai Ye for her unconditional love, support and accompany.

1.0 Introduction

1.1 Introduction to RNA-seq

There have been comprehensive reviews on the pipelines of generating and analyzing bulk RNA-seq data [38, 32, 130, 79]. The overall generative model for the final quantified gene expression is visualized in Fig. 1, which mainly contains three components.

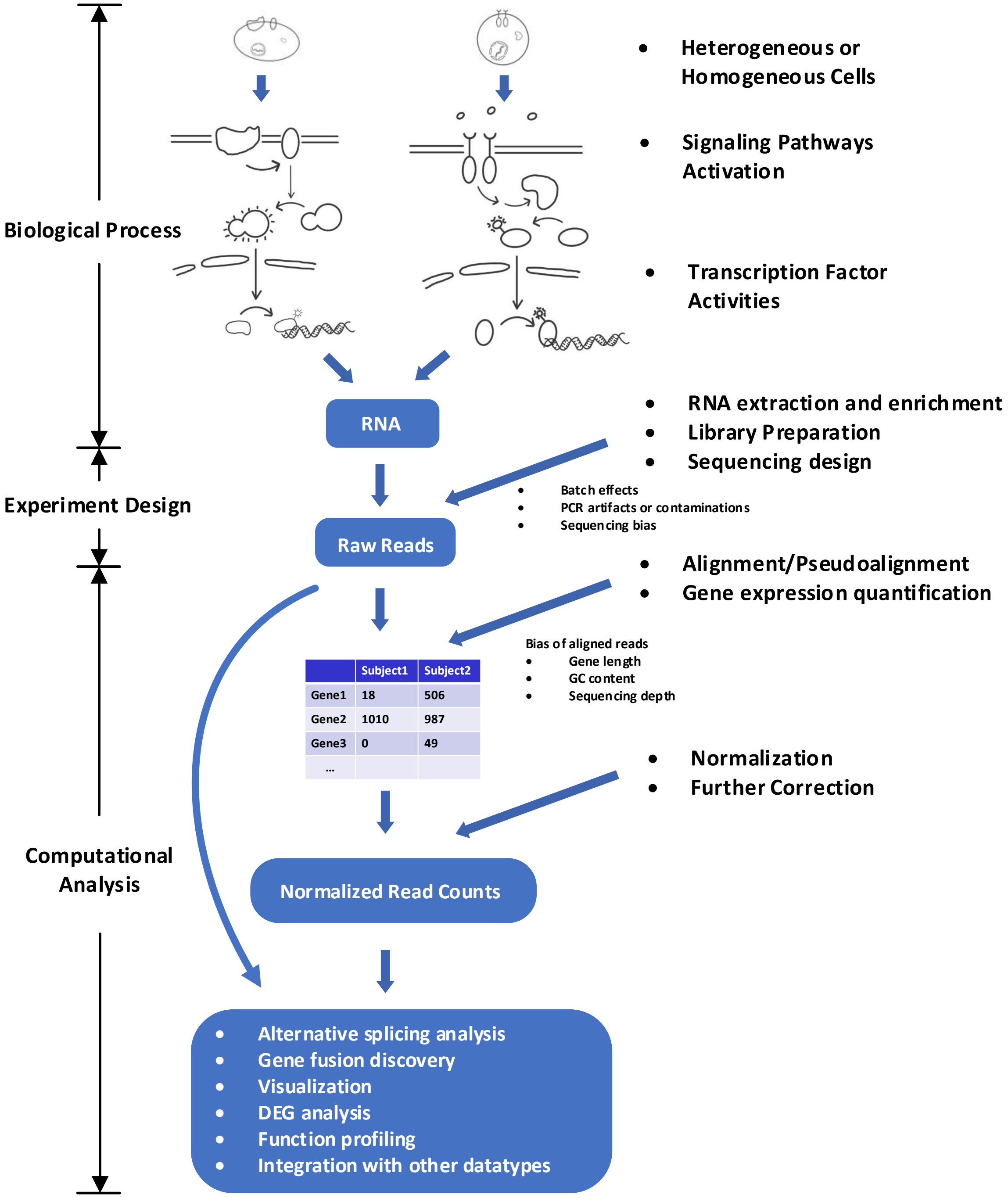
Biological Process This part corresponds to the exact underlying biological mechanisms that control the quantity of RNAs in the sample. The RNAs produced by a single cell are governed by its epigenetically programmed cell type and its response to external stimuli, or cell state. For bulk RNA-seq, the final RNA pool is made up of RNAs from all cells contained in the sample as a snapshot of different cell types and cell states.

Experiment Design This refers to all the necessary experimental steps to turn the original RNA pool into sequencing reads. This includes RNA extraction&enrichment, library preparation and sequencing itself. Each of these steps introduces variations from different sources.

Computational Analysis The computational analysis includes the process of converting raw sequencing output into RNA quantities. Meanwhile it is possible to detect transcripts, alternative splicing events or gene fusions. However the most common output is a gene-level quantification matrix with respect to a predefined transcriptome.

As shown in Fig. 1, noise and nuisance variations are introduced to the RNA-seq readouts at various steps. For example, GC content, gene body coverage evenness, base error rate and nucleotide composition are factors leading to the bias of aligned reads in the alignment step [96]. Variations that are introduced by experiment setup and computational analysis are undesirable because these technical biases imply that the final RNA quantification doesn't accurately reflect the real RNA quantities in the RNA pool.

We consider data normalization to be the data manipulations aimed at removing **all unwanted variation** which may be technical or biological. This is crucial for all downstream analysis relying on precise RNA quantities. In the following section 1.2, we present a review of strategies that are actively utilized to estimate and remove unwanted variations.



	Subject1	Subject2
Gene1	18	506
Gene2	1010	987
Gene3	0	49
...		

Figure 1: An end-to-end pipeline for RNA-seq. It contains three modules which are Biological Process, Experiment Design and Computational Analysis.

1.2 Variation Modeling

A series of reviews [125, 36, 95, 41, 24, 158] have been published focusing on the development of normalization methods. Some of the normalization methods are naturally adapted from microarray analysis, e.g., voom [82] while others focus on methods that are RNA-seq specific, e.g., RUVseq[126]. We can consider all of these available approaches from a universal perspective of *variation modeling*. The variation in each dataset is influenced by various factors such as batch effects, experimental variables, various sources of biological variation, some of which are desirable while others are not. The hierarchy of variation modeling is depicted in Fig. 2.

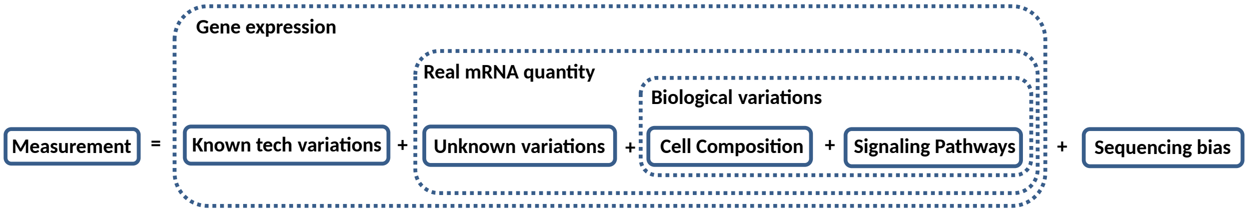


Figure 2: Different variations are additive. The annotations in this figure are consistent with Fig. 4, 5, 6 and 7.

There are two categories of variation modeling in general which are generative modeling approach and heuristic approach. In Fig. 3, we classify different modeling approaches along two dimensions: the formulation, and the type of variations they try to model. Heuristic approach is to remove the unwanted variations one at a time, in a stepwise manner (Section 1.2.1), while the generative modeling approach is to model variations with generative models (Section 1.2.2). There are two subcategories of generative modeling: explicit modeling, and implicit modeling. Explicit models attempt to extract factors that are individually interpretable. For example Cibersort estimates the abundance of specific cell types. Implicit models attempt to only capture a subspace that corresponds to a type of variations without individual sources. For that reason variables extracted by explicit methods can be both analyzed directly and regressed out for data normalization while implicit methods are used for the normalization only. The orthogonal dimension in Fig. 3 represents the type of variations that are of interest and there are mainly two groups: biological variations, and technical variations.

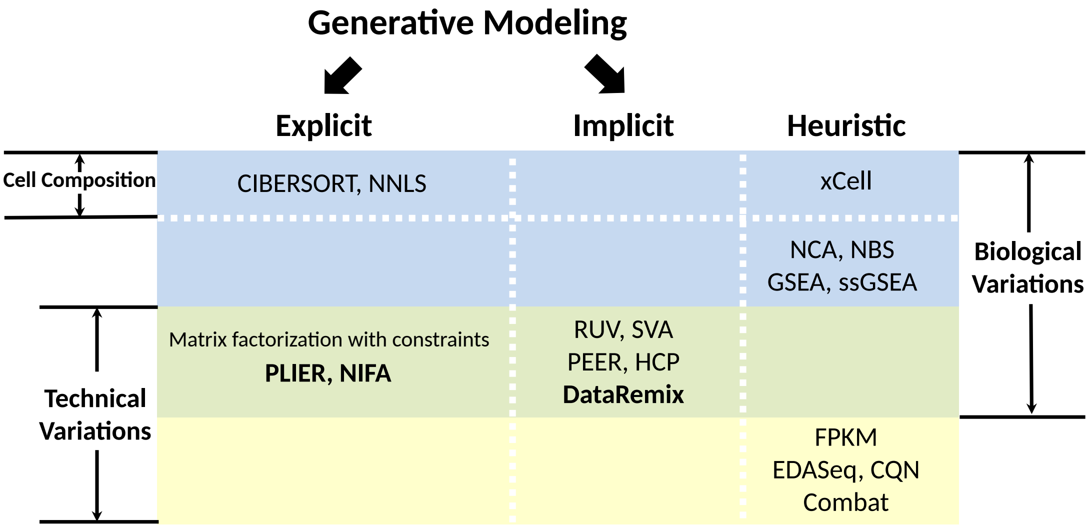


Figure 3: Classification of variation modeling approaches. Generally there are two approaches: generative modeling and heuristic approach. Within the category of generative modeling, there are explicit methods and implicit methods. On the orthogonal dimension, we categories modeling methods by the type of variations they are trying to model: biological variations and technical variations. We also highlight cell composition estimation as a special subcategory. For each subcategory, we only list one or two representative methods and detailed discussions can be found in Sec. 1.2.1.1, 1.2.1.2, 1.2.1.3, 1.2.1.4 and 1.2.2.

1.2.1 Model (and remove) unwanted variations in a heuristic manner

1.2.1.1 Correct sequencing bias Sequencing bias is contributed by various factors, such as gene length, GC-content effects, sequencing depth, etc., and it has to be corrected in order to get more accurate estimation of RNA quantities (Fig. 4).

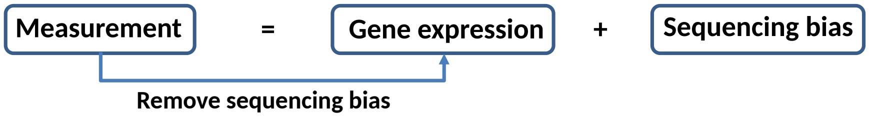


Figure 4: Correct sequencing bias from RNA-seq measurements. The annotations are consistent with those of Fig. 2.

Table. 1 lists normalization metrics that corrects length bias and other experimental noise in order to retrieve gene expression abundance from noisy measurements. EDASeq [135] and CQN [58] are two other methods that are capable to correct GC-content effects.

Methods	Abbreviation	Initially Proposed for
Total count	TC/RC (raw count)	RNA-seq
Upper Quartile	UQ [24]	RNA-seq
Median	Med	RNA-seq
DESeq	DESeq [7]	RNA-seq
Trimmed Mean of M-values	TMM [136]	RNA-seq
Quantile	Q [143]	Microarray
Reads Per Kilobase per Million mapped reads	RPKM [114]	RNA-seq
Effective Reads Per Kilobase per Million mapped reads	ERPKM [132, 95]	RNA-seq
Transcripts Per Million	TPM [92]	RNA-seq
Fragments Per Kilobase per Million mapped reads	FPKM [157]	RNA-seq
transformed $\log_2$(FPKM)	zFPKM [59]	RNA-seq

Table 1: Normalization methods correcting sequencing bias.

1.2.1.2 Regress out known technical covariates

Besides sequencing bias, there are other unwanted variations because of experimental design (Fig. 5). The most prominent factors is the batch effect [83, 51], which can dramatically affect the conclusions of downstream analysis. Leek et al. [87] presents a comprehensive review of batch effects and emphasize how crucial to adjust for batch effects. It is shown that ComBat [69] outperforms other correction methods generally [28].

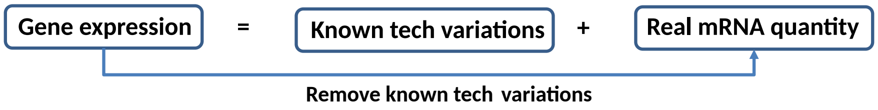


Figure 5: Regress out known technical covariates. The annotations are consistent with those of Fig. 2.

1.2.1.3 Estimate or eliminate unknown variations

Other implicit variations are all included into a complement category of ”unknown variations”. Depending on the source and the goals of the study, **unknown variations are not unwanted all the time**. But in most cases, these implicit variations are contributed by noise or artifacts that are indeed unwanted. We would like to estimate these nuisance variations and regress them out subsequently.

The general strategy is to estimate hidden covariate without removing significant biological variations. As reference spike-in may not work, a series of work have been formulated based on the idea using negative control genes to estimate unwanted variations. It includes RUV-2 [48], RUV-4 [47], RUV [133], RUVseq [126] and RUVnormalize [67]. Similarly sva [86, 88] and svaseq [85] first identify genes that are affected by unknown variations and then perform decomposition to estimate the unknown artifacts. Other methods directly utilize matrix factorization models to accomplish this task. For example [128] uses Principal Component Analysis (PCA) to estimate the unwanted variation and use the estimation as covariates.

1.2.1.4 Estimate cell abundance and pathway activities One source of biological variation that is often of interest is cell-type abundance. There have been continuous efforts to estimate cell proportions from transcriptomics data based on heterogeneous samples [152]. Cell-type variation can be both either a nuisance factor or a factor of interest depending on the question asked. For example in the case of tumor profiling where we may wish to study a pure tumor sample, cell composition is actually considered a nuisance variable. However, with a different scientific question, such variation can be of biological interest. Tumor impurity is in part driven by immune infiltration which is highly predictive of the outcome [155]. Table. 2 provides a brief list of available methods designed specifically for different heterogeneous tissues. These methods provide estimates of the cell-type composition of the sample and these estimates can be analyzed directly, regressed out or both. A key feature that of all of these methods is that they rely on prior knowledge of the gene expression state of the pure cell types in the mixture, either a set of cell-type specific gene expression basis or simply a list of marker genes.

The other source of biological variation is pathway activity (Fig. 6). It's not trivial to estimate pathway effects directly from RNA-seq and it is more common to test whether any a priori defined gene sets show statistically significant difference between sample groups. GSEA [153] first introduced this problem to the field and a series of methods have been

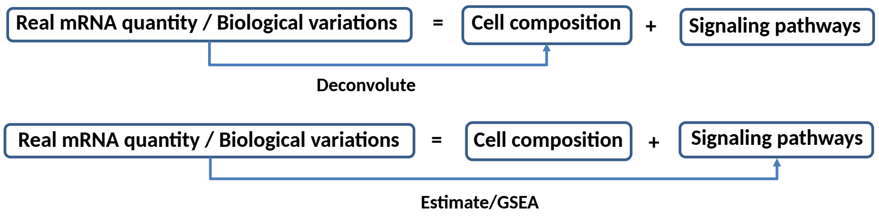


Figure 6: Estimate cell abundance and pathway activities. The annotations are consistent with those of Fig. 2.

developed subsequently[65, 140]. Other related attempts are about how to incorporate the pathway information as constraints into matrix factorization frameworks in order to get more useful representations. NCA [97] incorporates the bipartite structure between transcription factors and regulated genes to capture the dynamics of transcription factor regulations during cell cycle. NBS [62] encodes the global gene-gene interaction network structure as additional constraints into the non-negative matrix decomposition framework in order to get more consistent assignment of tumor subtypes.

Tools	Tissue
DeconRNASeq [53, 54]	Blood
NNLS [1]	Blood
CelLCODE [30]	Blood
xCell [9]	Blood
CIBERSORT [119]	Blood
CIBERSORT-X [120]	Blood
ESTIMATE [168]	Tumor
TIMER [93]	Tumor
MCP-counter [17]	Tumor

Table 2: Normalization Methods estimating cell type proportions.

1.2.2 Model different variations simultaneously

Compared with the stepwise strategy, the other attempt is to model all possible variations at the same time with the help of generative modeling (Fig. 7). Generative modeling is not a well-defined concept and the terminology is not consistent across literatures. Generally in the context of unsupervised learning, the objective of generative modeling is to model the probability distribution of the data and thus make it possible to generate new samples from the distribution. Factor analysis is a special case of generative modeling [145] which assumes the observed data X is generated from a set of unobserved latent factors S through the transformation $X = AS+$noise. A is denoted as the coefficient matrix which defines the linear transformation. I will use the term factor analysis and matrix factorization/decomposition interchangeably through the thesis.

Matrix factorization has been proved as a powerful tool to model the variations [150], and it can summarize the variations embedded in the high-dimensional RNA-seq readouts as compact representations. Matrix factorization also makes it possible to extract variations with certain patterns by including corresponding regularization. The very first application of matrix decomposition can be traced back to [5] using Singular Value Decomposition (SVD) to analyze microarray data. Other general matrix decomposition frameworks include Non-negative Matrix Factorization (NMF) [21], Independent Component Analysis (ICA) [76, 40], Penalized Matrix Decomposition (PMD) [165] and tensor decomposition [63]. More recent efforts include GPseq [146], PEER [149] and HCP [116].

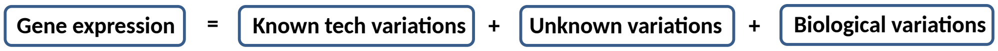


Figure 7: Model different variations simultaneously. The annotations are consistent with those of Fig. 2.

1.2.3 Quality of variation modeling

The final output from normalization is either a high-dimensional matrix with the same dimension of the original RNA-seq count matrix or a compact representation summarizing the input data. There are multiple perspectives to evaluate the representation as listed in Table. 3.

Computationally	Biologically
• A good representation can make it easier to build a classifier or predictor, in which the subsequent learning, hypothesis testing or causality inference task becomes easy to accomplish.	• The biological utility of the data should be optimized with respect to the representations.
• A good representation is able to maintain the majority of the embedded variations and capture the hidden structures.	• A good representation can make the downstream analysis not affected by technical noise and other unwanted variations.
• A good representation should be reproducible and can be generalized across related tasks.	• The representation should be consistent and reproducible across related cohorts.
• A good representation should be interpretable.	• We can annotate the representations with domain knowledge.

Table 3: Different perspectives to assess quality of normalized representations.

1.3 Single Cell RNA-seq

Compared with bulk RNA-seq technology, single-cell RNA-seq makes it possible to query the complexity of transcription at single cell level. The most different step in the pipeline is to isolate each cell from the others before extracting RNAs from the cell. Some of the popular protocols for cell isolation are SMARTer, Smart-seq2, inDrop and 10x genomics drop-seq.

The analysis pipeline dealing with scRNA-seq is mostly adapted from that of bulk RNA-seq. A comprehensive review of current best practice in scRNA-seq analysis is included in [105]. Enormous number of tools have been developed to leverage biological insight at single-cell level. An extensive list [170] of all available tools can be viewed via the following link https://www.scrna-tools.org/tools. Specifically for normalization step [160], some of the popular choices are ZIFA [127], ZINB-WaVE [134], sclvm [22], fsclvm [23], scVI [104] and pagoda [42].

1.4 Main Contributions

In this dissertation, we propose three different generative models to amplify biological signals in bulk and single-cell RNA-seq data. PLER and NIFA are two different matrix decomposition methods and DataRemix is built upon SVD which is one of the most popular matrix factorization methods.

- **PLIER** Pathway-level Information Extractor (PLIER) is a semi-supervised framework that identifies active pathways regulating gene expression and estimates surrogates for cell-type proportions. PLIER has inspired the development of MultiPLIER [154] and has been used to infer context-specific genotype effects in the Brain [121].
- **DataRemix** DataRemix is a supervised model that reweighs contributions of latent factors to reveal the hidden biological signals. We apply DataRemix to the ROSMAP dataset and we are able to report the first replicable *trans*-eQTL effect in human brain.
- **NIFA** (Non-negative Independent Factor Analysis) NIFA, an unsupervised framework, combines the desired properties of PCA, ICA and NMF. It models uni- and multi-modal factors corresponding to cell-type identity and other pathway-level variations simultaneously.

2.0 PLIER - Pathway-Level Information Extractor for gene expression data

2.1 Introduction

One salient feature of high dimensional molecular data structure is the presence of groups of correlated measurements. Gene expression measurements are highly correlated and this correlation structure often reflects coordinated transcriptional regulation or, in studies of heterogeneous tissues, variation in cell-type proportion. This data structure is exploited implicitly any time a clustering is performed, as is often done with cancer datasets in order to define molecularly distinct subtypes [118, 137]. Importantly, correlated expression patterns may also be the result of various technical factors, often referred to as "batch effects" (see [87] for review). It is crucial to identify the mechanisms underlying coordinated gene expression changes while reducing any negative effects of technical noise.

Likewise it is possible to analyze the structure explicitly by projecting the thousands of gene-specific measurements into a smaller dimensional space that captures much of the observed variation. Principal Component Analysis (PCA), which utilizes singular value decomposition(SVD) to project the data onto orthogonal principal components (PCs) of maximal variance, is commonly applied to gene expression datasets. PCA and its higher dimensional analogs have been successfully applied to gain biological insight from complex datasets [5, 63]. However SVD decompositions have several limitations. By construction PCA/SVD produces components that are orthogonal and are dense combinations of the original variables. The orthogonality implies that the components will not always correspond to specific biological variables (which are often not-orthogonal) and the loading density makes interpretation difficult.

Various alternative decomposition methods that seek to improve the interpretability by imposing additional constraints have been proposed. For example, non-negative matrix factorization (NMF) has been applied to cancer gene expression decomposition yielding more intuitive results [21]. Likewise, methods to introduce sparsity into the matrix decomposition have been proposed [173, 165]. However they do not make use of known biological information in their mathematically driven decompositions. We reasoned that the efficient extraction of biological insight contained in the correlated structure of the data requires using the vast information contained in biological genesets during the decomposition. To solve this problem, we have developed **P**athway-**L**evel **I**nformation **E**xtracto**R** (PLIER). PLIER performs a semi-supervised data structure deconvolution and mapping to external knowledge, reducing noise and identifying regulation in cell-type proportions or pathways.

2.2 Methods and Materials

Given a gene expression profile $Y \in \mathbb{R}^{n \times p}$, where n is the number of genes and p is the number of samples, we state the original PCA as a matrix approximation problem. Suppose $n > k$, $p > k$. We wish to find Z and B minimizing

$$||Y - ZB||_F^2 \tag{2.1}$$

$$\text{subject to} \quad \text{rank}(Z) = k, \, \text{rank}(B) = k.$$

Since gene expression measurements are highly correlated, it is reasonable to expect that the data Y can be efficiently represented in this low dimensional space. Without imposing additional constraints on Z and B, an optimal solution can be obtained from the singular value decomposition (SVD) of Y. In a SVD-based decomposition, rows of B are referred to as principal components (PCs). Since PCs are necessarily orthogonal, which our method do not require, we will use a more general term latent variables (LVs).

In order to improve the interpretability of the low dimensional representation in the context of known biology, we impose additional constraints on the matrix Z. Our aim is to encourage the loadings (columns of Z) to align as much as possible with existing prior knowledge. In the most general case such prior knowledge can be expressed as a series of genesets representing biological pathways, sets of tissue- or cell-type specific markers, and coordinated transcriptional responses observed in genome-wide experiments.

Given n genes and m genesets, we represent the prior knowledge as a matrix $C \in \{0,1\}^{n \times m}$, so that $C_{ij} = 1$ indicates that gene i is part of the j_{th} geneset. Using the same notation as above, we define the revised decomposition problem based on the original formulation. We wish to find U, Z, B minimizing

$$||Y - ZB||_F^2 + \lambda_1 ||Z - CU||_F^2 + \lambda_2 ||B||_F^2 + \lambda_3 ||U||_{L^1} \tag{2.2}$$

$$\text{subject to} \quad U > 0, \quad Z > 0.$$

The first term of the optimization is the same as equation (1) and minimizes the overall reconstruction error. The second term specifies that Z should be "close to" sparse combinations of genesets represented by C. The third term introduces an L_2 penalty on B, while the fourth term is an L_1 penalty on U (applied column-wise), which ensures that only a small number of genesets represent each LV.

The parameter λ_1 keeps a balance between the proportion of prior knowledge we include and the degree to which we reconstruct the gene expression profile. We also restrict U and Z to be positive, which enforces that genes belonging to a single geneset are positively correlated with each other and the loadings are positively correlated with the prior information.

We solve the optimization problem by using block coordinate minimization, which iteratively minimizes the error on Z, U, and B. The complete method starts by initializing Z and B from the SVD decomposition and repeats the following steps until B converges.

```
while stopping criterion has not been reached
  $Z^{(l+1)} \leftarrow (YB^{(l)^T} + \lambda_1 CU^{(l+1)})(B^{(l)}B^{(l)^T} + \lambda_1 I)^{-1}$
    Set the negative part of $Z^{(l+1)}$ to be zero
  Solve the convex problem
    $U^{(l+1)} \leftarrow \text{argmin}_U ||Z^{(l)} - CU||_F^2 + \lambda_3 ||U||_{L^1}$
    Subject to $U > 0$
  $B^{(l+1)} \leftarrow (Z^{(l)^T} Z^{(l)} + \lambda_2 I)^{-1} Z^{(l)^T} Y$
```

The stopping criterion is defined as a relative change in $B < 5 \times 10^{-6}$, or a leveling off in the decrease of the relative change in B. While there are no convergence guarantees, in practice this algorithm converges in under a few hundred iterations.

2.2.1 Optimization constants

The optimization has 4 free parameters λ_1, λ_2, λ_3, and k and internal cross validation cannot be used to optimize them as the reconstruction error $||Y - ZB||_F^2$ is always minimized when $\lambda_1 = 0$. However, based on extensive testing with simulations and real data, we have set default parameters that perform well in a range of situations. For example, we find that a reasonable starting value for k can be inferred from the the number of statistically significant PCs which can be determined via permutation by the approach proposed in [88] or the simple "elbow" approach (`num.pc` in our package implements both). However, it is logical that the number of constrained latent variables needed to explain the data is higher, and we suggest increasing the initial k by a factor of 2. Importantly, the method is not sensitive to the exact value of k. LVs found at lower k's persist when k is increased. It is also possible to optimze k with respect to the number of LVs with prior information above some AUC and FDR threshold, but this requires multiple runs.

A good choice for λ_1 and λ_2 can be derived from the observation that if we consider the SVD decomposition of Y as UDV^T we should have that $Z \approx UD^{1/2}$ and $B \approx D^{1/2}V^T$. Therefore the diagonal elements of Z^TZ and BB^T are well approximated by D which thus gives the correct range for the relevant constants. By default we set $\lambda_2 = d_k$ and $\lambda_1 = d_k/2$ with the factor of 2 coming from the positivity thresholding on Z. We find that our method is robust to these choices (Fig. 14). It is also possible to optimize λ_1 along with λ_2 around its default value relative to some external validation source. For example, we can check how well the LVs recovered in B correlate with an independent dataset such as clinical variables, genotype, or another set of molecular measurements.

The correct value of constant λ_3 that controls the sparsity of U is highly dataset dependent as it ultimately depends on how well the available prior information explains the data structure. We have devised an adaptive approach that works well for datasets of diverse characteristics. Specifically, we can specify the fraction of latent variables that we wish to be associated with prior information, 0.7 by default. The λ_3 constant is then periodically adjusted by binary search to meet this goal. Even though this adaptive procedure keeps the number of positive entries in U constant regardless of prior information relevance, the significance of pathway association for each LV is ultimately tested by gene-holdout cross-validation (see below).

For our dataset with matched CyTOF proportions we used default PLIER parameters. For the DGN dataset we used all default parameters except that k was optimized to maximize the number of LVs with significant pathway association.

2.2.2 Gene-holdout Cross Validation

It is natural to ask to what extent the non-zero coefficients of U represent non-random associations between loadings (columns of Z) and prior information. In order to quantify this we design a cross-validation procedure that proceeds as follows. For each pathway included in the entire prior-information compendium a random 1/5th of the positive genes are set to 0 and this new prior information matrix is used to run PLIER. Afterwards, we can test how well the gene loadings in the PLIER output matrix Z are able to recover these held-out genes. Specifically, for each LV-pathway correspondence represented as a positive value in U we compute the AUC and p-value for the recovery of that pathway in the loadings of Z using the held-out set of genes as positive labels and genes not annotated to this pathway as negative labels. We find that the cross-validation procedure produces correct AUC estimates as p-values computed from a gene-level permuted prior information geneset (which preserves dependencies among pathways) are uniformly distributed (Fig. 13).

While this procedure necessarily discards some data and may adversely affect the ability to detect small pathways, we find that the benefit of having accurate statistical estimates outweighs these concerns. PLIER will run in cross-validation mode by default but we allow for cross-validation to be turned off in which case all genes belonging to each geneset are used.

2.2.3 Details of methods comparisons

For validation we compare the Spearman rank correlation of CyTOF-based proportion measurements with estimates obtained from different methods. P-values thresholds indicated on the plot are for the single tailed test. We compare performance on the validation dataset against 4 alternative approaches. Two of the approaches are matrix decomposition methods that are commonly applied to gene expression data: Sparse Principal Component Analysis (SPC) and Non-negative Matrix Factorization (MNF). The other two approaches are reference-based methods that are specifically designed to estimate human blood cell-type proportions. The methods are Non-negative Least-squale regressionm, NNLS (originally applied to cell-type deconvolutioni [1]) and Cibersort [119]. We found that quantile normalization improved the deconvolution performance of matrix decomposition methods (SPC, NMF, and PLIER) but as previously noted reference-based methods (NNLS and Cibersort) performed best with raw (not log transformed) FPKMs.

For NMF we used the default algorithm and matrix norm as implemented in the NMF R package [21]. Since NMF requires a positive matrix we used quantile-normalized log counts which achieved best performance. SPC has no restrictions on the input and in our experiments performed best on z-scored data (z-scored data are also used for PLIER). We used the SPC implementation provided in the PMA package [165]. We used the positivity constraint on the loadings matrix, which improved the results. The sparsity hyperparameters for SPC were set with cross-validation separately for each component as described in the original paper [165]. Since SPC and NMF do not assign a biological cause to the inferred latent variables, for the purpose of evaluation we report the maximum correlation for each cell type. The number of components for SPC and NMF was set to 30 which is the same number that was used for PLIER.

For NNLS and Cibersort we used raw FPKM values which is the preferred data transformation for Cibersort and also performed best in our evaluations. Since NNLS and Ciberosrt are both reference-based methods and can be used with any reference/basis matrix we tried both approaches with two different references, one from [1] and LM22 from the original Cibersort publication. We found that each method performed best with its own original reference. To account for the fact that our cell-type classes are slightly different from those encoded in LM22 or the [1] reference we allowed various combinations of the estimates, for example we created an "all-Bcell" estimate by adding naive and memory B cells and picked the best correlated estimate out of the three. A similar approach was taken for other cell types.

2.2.4 *trans*-eQTLs

For the purpose of all our analysis we define *valid trans associations* as gene-SNP pairs where the gene and all of its homologs (as defined by Ensembl database, [171]) are on a different chromosome from that of the SNP.

2.2.4.1 Gene-centric eQTLs

- Compute p-value for all valid *trans* associations using rank correlation.
- Compute Benjamini-Hochberg false discovery rate on the total number of valid *trans* association test.

2.2.4.2 Pathway-centric eQTLs Since pathway LVs are composed of multiple genes from different chromosomes, all LV-SNP associations are potentially valid *trans* associations. The steps for computing pathway-centric eQTLs are bellow

- Step 1: Perform rank correlation tests on all LV-SNP pairs.
- Step 2: Compute Benjamini-Hochberg FDR on the entire set of pathway-level test (num. of LVs)x(num. of SNPs). Association with FDR>0.05 are not considered further.
- Step 3: Compute gene-level support for pathway-level eQTLs. Perform all valid *trans* association test on the subset of SNPs that passed FDR<0.05 for at least on LV in Step 2.
- Step 4: Compute Benjamini-Hochberg FDR for tests in Step 3 correcting for the total number of tests performed overall (number of tests in Step 1 plus number of tests in Step 3). We note that the p-value threshold for FDR=.2 in the PLIER-centric analysis is higher than the gene-centric analysis (4.1e-07 versus 7.7e-08). It is possible that these PLIER-centric FDRs are overly permissive due to the hierarchical nature of the tests in Step1 and Step 3, however we emphasize that we do not rely on these values for any conclusions in our analysis. They are only used to define the upper limit for associations that are checked for replication.
- Step 5: Filter pathway-level effects with low gene-level support. We defined low gene-level support as 0 genes-SNP associations that pass a gene-centric FDR of <0.2. That is any

pathway-level association has to be supported by at least one gene in gene-centric analysis at a permissive FDR. This step is designed to get rid of any spurious *trans* associations discovered in Step 1 that could arise due to *cis* genes or *cis* homologs contributing to the pathway-level estimate.

2.2.5 Validation Data

2.2.5.1 Sample Processing We used anonymized discard samples that the have the determination of non-human research. Blood was drawn into Tempus tubes (AB scientific) for RNA and into EDTA tubes for Cyto analysis respectively. RNA was extracted using the MagMAX^{TM} for Stabilized Blood Tubes RNA Isolation Kit (Fisher) following the manufacturer's protocol. Libraries were constructed using the TruSeq Stranded mRNA kit (Illumina) at the Epigenetic core at the Weil Cornell medical college.

2.2.5.2 CyTOF Sample Processing CyTOF antibodies were either purchased pre-conjugated from Fluidigm (formerly DVS Sciences) or purchased purified and conjugated in-house using MaxPar X8 Polymer Kits (Fluidigm) according to the manufacturer's instructions. Whole blood samples were processed within 4hrs of collection and stained by additional of a titrated panel of antibodies (table X) directly to 400uL of whole blood. After 20 minutes of incubation at room temperature, the samples were treated with 4mL of BD FACSLyse and incubated for a further 10mins. The samples were then washed and incubated in 0.125nM Ir intercalator (Fluidigm) diluted in PBS containing 2% formaldehyde, and stored at 4oC until acquisition.

Immediately prior to acquisition, samples were washed once with PBS, once with deionized water and then resuspended at a concentration of 1 million cells/ml in deionized water containing a 1/20 dilution of EQ 4 Element Beads (Fluidigm). The samples were acquired on a CyTOF2 (Fluidigm) at an event rate of <500 events/second.

2.2.5.3 CyTOF Data Analysis After acquisition, the data were normalized using a bead-based normalization in the CyTOF software and uploaded to Cytobank for initial data processing. The data were gated to exclude residual normalization beads, debris, and doublets, and exported for subsequent clustering and high dimensional analyses.

Individual samples were first clustered using Phenograph [90], an agnostic clustering method that utilizes the graph-based Louvain algorithm for community detection and identifies a hierarchical structure of distinct phenotypic communities. The communities were then meta-clustered using Phenograph to group analogous populations across patients. These meta-clustered populations were then manually annotated based on similar canonical marker expression patterns consistent with known immune cell populations. These annotations are also used to generate a consistent cluster hierarchy and structure across all samples in the dataset.

2.2.5.4 RNA-seq methods The samples were sequenced SE100 to an average depth of 48.8 million reads. Quality assessment was done with FastQC [19]. Alignment to GenCode hg38 was done using STAR [37]. Transcript counts are assigned using the FeatureCounts tool (subread package [98]). The final counts were filtered for genes that had 0 counts in all samples. The data was transformed to RPKM. We also created a quantile normalized count dataset by filtering all genes that had <3 counts in any samples and quantile normalizing the log transformed counts. This more stringent filtering was performed to avoid data artifacts caused by quantile normalization of low count genes. As RNA samples were processed in two separate batched both final datasets were corrected for batch difference.

2.2.6 Public Data

2.2.6.1 DGN dataset The Depression Gene Networks (DGN) dataset is not available for public release but can be requested from National Institute of Mental Health (NIMH) following instructions in the original publication [115]. The NIMH database contains several normalized versions of this data and for our study we used "trans" normalized data as described in [16]. This data is already normalized for genotype principal components and all known technical factors and no further normalizations were performed.

2.2.6.2 NESDA dataset The NESDA (Netherlands Study of Depression and Anxiety) dataset [166] was obtained from dbGAP (phs000486.v1). Following suggestions from study authors the NESDA dataset was normalized for known technical factors and the first 3 genotype PCs using linear regression.

2.2.6.3 Prior information genesets The generic blood cell-type marker dataset was derived from the IRIS (Immune Response In Silico) [1] and DMAP (Differentiation Map) datasets [122] datasets. Many canonical marker genes (such as CD19, CD3E, CD8A) have a multimodal distribution with on high expressor group and one or more low/medium expressor ones. The highest expression group typically does not overlap with lower expression distributions and we base our marker selection metric on this observation. Genes were considered to be markers if they could be partitioned into high and medium/low expression so that the difference between minimum and maximum values respectively (the gap between these distributions) exceeds a threshold (we used 2 for IRIS and 0.7 for DMAP). This procedure results in highly overlapping sets of markers for related cell types however our method is flexible and can easily handle redundancy. The marker sets derived from the IRIS and DMAP datasets are included in the PLIER R package. For the purpose of analyzing DGN we also included cell-type markers from a recent publication [119] which covers fewer cell types but with highly optimized marker sets. The complete prior information dataset used for DGN analysis includes cell-type markers, "canonical pathways" and "chemical and genetic perturbation" genesets from mSigDB, and a set of transcriptional signatures relevant to immune signaling described in [43].

2.2.6.4 Replication To assess replication in the NESDA dataset SNPs were matched based on LD using the LDlink tool with CEU population [106]. Specifically, if the exact SNP was not present in the NESDA dataset we selected the SNP with the highest LD, and if multiple SNPs had the same LD, we took the one closest in genomic coordinates. We only considered a match if the best LD was above 0.5. We asses the relationship between the NESDA replication π_1 and the p-value obtained in DGN in two different ways. One uses a consistent cutoff of $\lambda = 0.05$ so that the π_1 estimate is simply computed as 1 minus the fraction of p-values above 0.05 divided by 0.95. We also evaluate π_1 using the method implemented in the "qvalue" Bioconductor package [34]. This methods selects the optimal λ for each π_1 estimate. We find that the typical value is around 0.8 though a different value may be selected at each threshold resulting in more noise in the π_1 curve.

2.2.6.5 Platelet phenotypes The sentinel SNPs and their relevant phenotypes (MPV, PLT, or both) are supplied as the supplementary table in [52]. Proxy SNPs were defined as above.

2.2.7 Data Availability

Processed gene expression and cell proportion measurements generated for this study are available through the PLIER package. The raw data can be accessed through Gene Expression Ominibus (GSE130824). The Depression Susceptibility Genes and Networks (DGN) dataset can be obtained from NIHM following instructions provided in the original publication [115]. The NESDA dataset can be obtained from dbGAP (identifier: phs000486.v1).

2.3 Results

PLIER approximates the expression pattern of every gene as a linear combination of eigengene-like latent variables (LVs). In constructing LVs, PLIER surveys a large compendium of prior knowledge (genesets) and produces a dataset deconvolution that optimizes alignment of LVs to a relevant subset of the available genesets. The method automatically finds these relevant genesets among the hundreds to thousands considered (see Fig. 8A). Technical noise reduction is also achieved during the deconvolution as technical factors are preferentially segregated into LVs that do not associate with prior information (see Fig. 9 and 10).

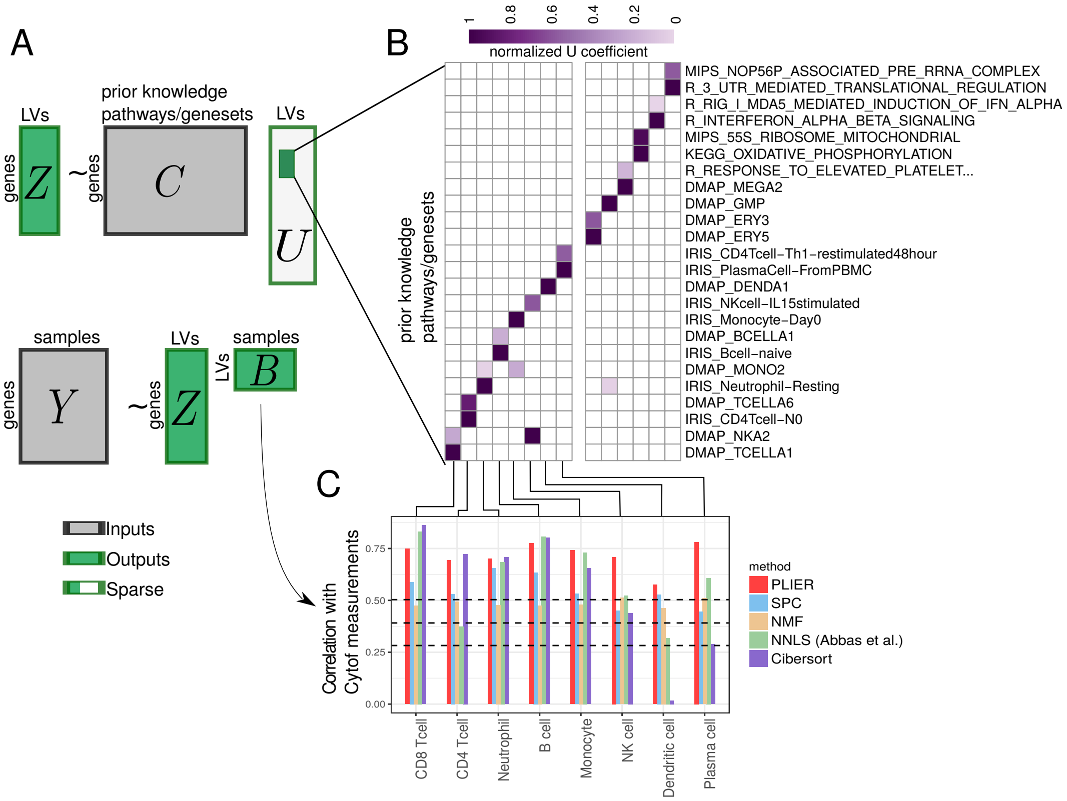


Figure 8: **PLIER overview.** PLIER is a matrix factorization approach that decomposes gene expression data into a product of a small number of latent variables and their corresponding gene associations or loadings, while constraining the loadings to align with the most relevant automatically selected subset of prior knowledge. **A.** Given two inputs, the gene expression matrix Y and the prior knowledge (represented as binary geneset membership in matrix C), the method returns the latent variables (B), their loadings (Z), and an additional sparse matrix (U) that specifies which (if any) prior information genesets and pathways are used for each latent variable. The light gray area of U indicates the large number of zero elements of the matrix. We apply our method to a whole blood human gene expression dataset. **B.** The positive entries of the resulting U matrix are visualized as a heatmap, facilitating the identification of the correspondence between specific latent variables and prior biological knowledge. Since the absolute scale of the U matrix is arbitrary each column is normalized to a maximum of 1. **C.** We validate the latent variables mapped to specific leukocyte cell types by comparing PLIER estimated relative cell-type proportions with direct measurements by Mass Cytometry. Dashed lines represent 0.05, 0.01, and 0.001 significance levels for Spearman rank correlation (single-tailed test). We find that the PLIER estimates are highly accurate, outperforming other matrix decomposition methods. Moreover, PLIER estimates are competitive and in 4 cases outperform both of the dedicated blood mixture deconvolution method NNLS [1] and Cibersort [119].

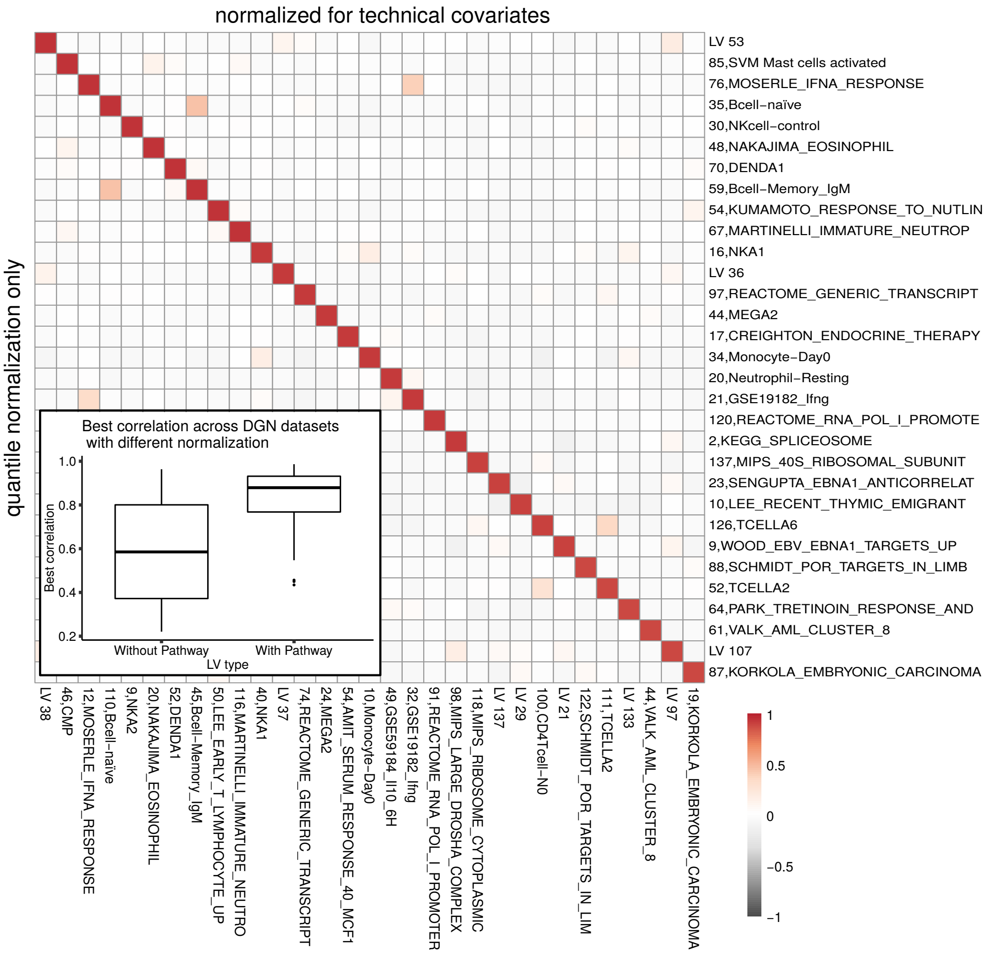


Figure 9: PLIER decompositions are robust to normalization procedure. We compare the PLIER decompositions obtained from two different versions of the DGN dataset: one normalized with all known technical factors and the other normalized only by quantile normalization. (Main panel) Heatmap of the Spearman rank correlations (computed across 922 samples) between LVs from the two decompositions. All pairwise correlations for the top best matched LVs (correlation >0.9) are shown. LVs are named with their corresponding top prior-information geneset (if any). Note that the prior information used is almost identical across the two decompositions. (Inset) The distribution of Spearman rank correlation values across all best reciprocal match pairs of LVs. LVs that use prior information (LVs with any non-zero U coefficients) are more robust to normalization procedure as they are more likely to have a high-correlation match across the two datasets. Boxplot displays the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest.

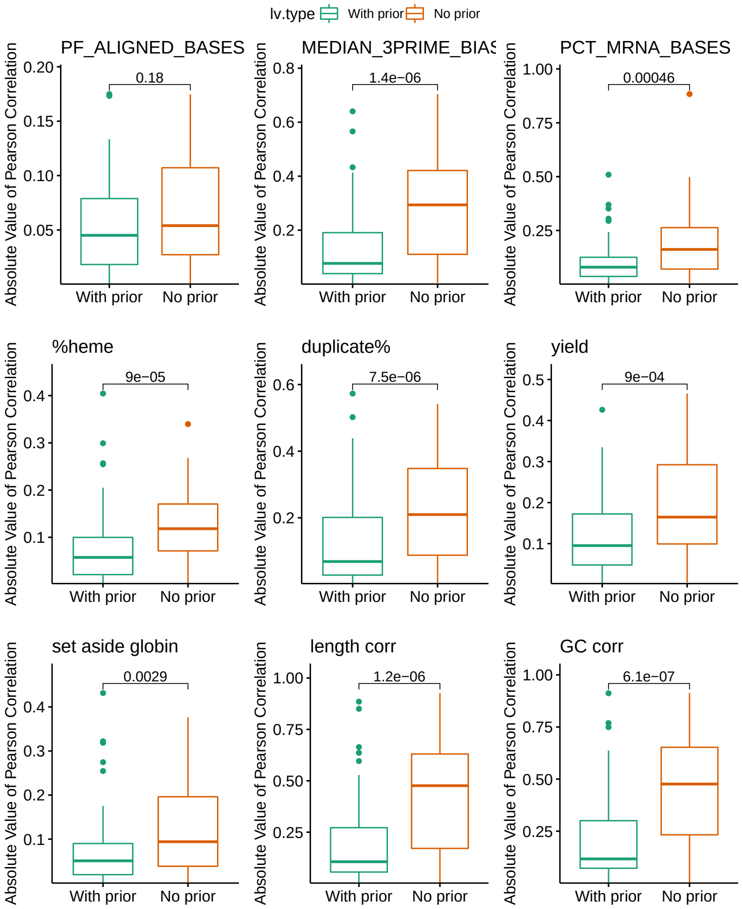


Figure 10: PLIER decompositions isolate technical and biological variation in different components. Using a PLIER decomposition obtained from the qunatile normalized DGN dataset, which has not been corrected for known technical factors, we plot the distribution of absolute values of Pearson correlation (across 922 samples) between technical factors and LVs. Since technical factors are correlated, we select a non-redundant set for our evaluation (PF_ALIGNED_BASES, MEDIAN_3PRIME_BIAS, PCT_MRNA_BASES, %heme, duplicate%, yield, set aside globin, length corr, GC corr) where the last two are per-sample correlations between quantified gene expression vectors and transcript GC content and length respectively. LVs associated with prior information (LVs with non-zero U coefficients) and LVs without prior information (LVs with zero U coefficients) are contrasted using two-sided Wilcoxon rank-sum test. The correlation for LVs with prior information is significantly lower for all but one technical factor (PF_ALIGNED_BASES). Boxplots display the 25th, 50th and 75th percentile, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest.

2.3.1 Simulation

Since in a real dataset the true data-generating model is unknown and is likely more complex than what can be captured with a dimensionality-reducing matrix decomposition, we use a simulation to evaluate the operating characteristics of our method. We hypothesize that our method is able to more accurately recover the “correct” LVs by rotating the matrix decomposition to align with prior knowledge.

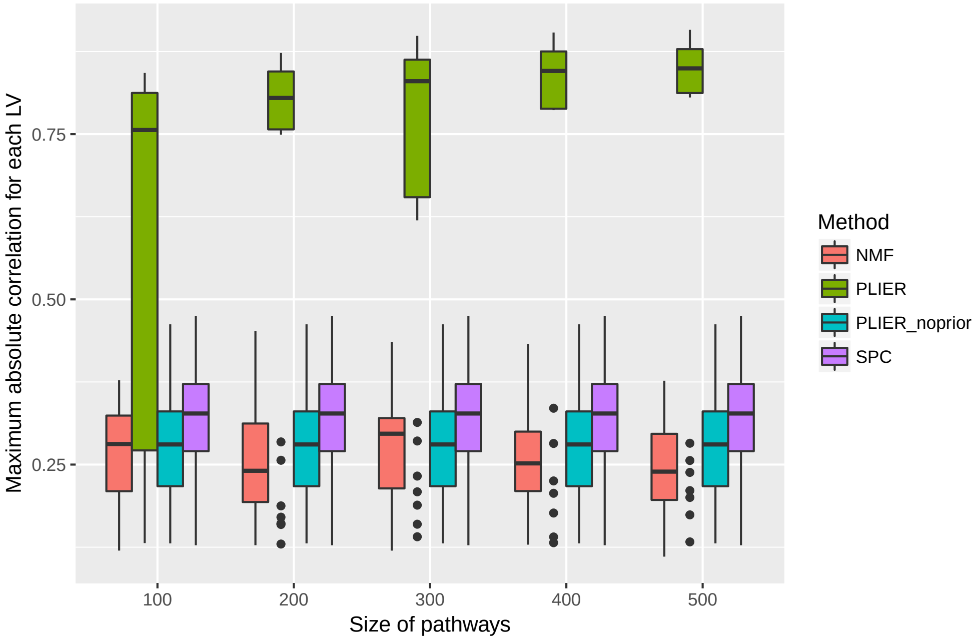


Figure 11: Boxplot of the correlation between simulated LVs and those recovered by various decomposition methods. We compare PLIER against two other methods, NMF [21] and SPC [165], as well as PLIER run without using any prior information. In this simulation we provide PLIER with 1000 pathways of which only 30 are correct and vary the size of the prior-information pathways provided to PLIER. We find that the best performance is achieved by PLIER specifically when prior information is used with a notable improvement when prior-information pathways are larger. Statistics were computed using Pearson correlation across 300 samples. Boxplot displays the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest.

We simulate data with 5000 genes, 300 samples, and 30 latent variable according to the NMF model.

$$Y = ZB + E. \tag{2.3}$$

With both Z and $B > 0$. Each row of B is drawn from Beta distribution with a mean drawn uniformly at random and a variance of 0.1. Each column of B is normalized to sum to one. The columns of Z are drawn from Gamma distribution $\Gamma(5, 1)$. The matrix $E \in \mathcal{N}(0, 1)$ represents random noise. We also generate a prior knowledge matrix C. For each column of Z, we randomly pick up a threshold value on the percentage of genes which belong to a hypothetical prior knowledge geneset. The threshold value varies from 0.01 to 0.1 with a step size 0.01, which is in consistent with that of real biological genesets. With the threshold value, we select the corresponding fraction of genes which come with top values in the column of Z to construct the prior knowledge geneset. Also we generate additional uninformative genesets by randomly picking genes. For the purpose of applying PLIER and SPC, the final data is z-scored.

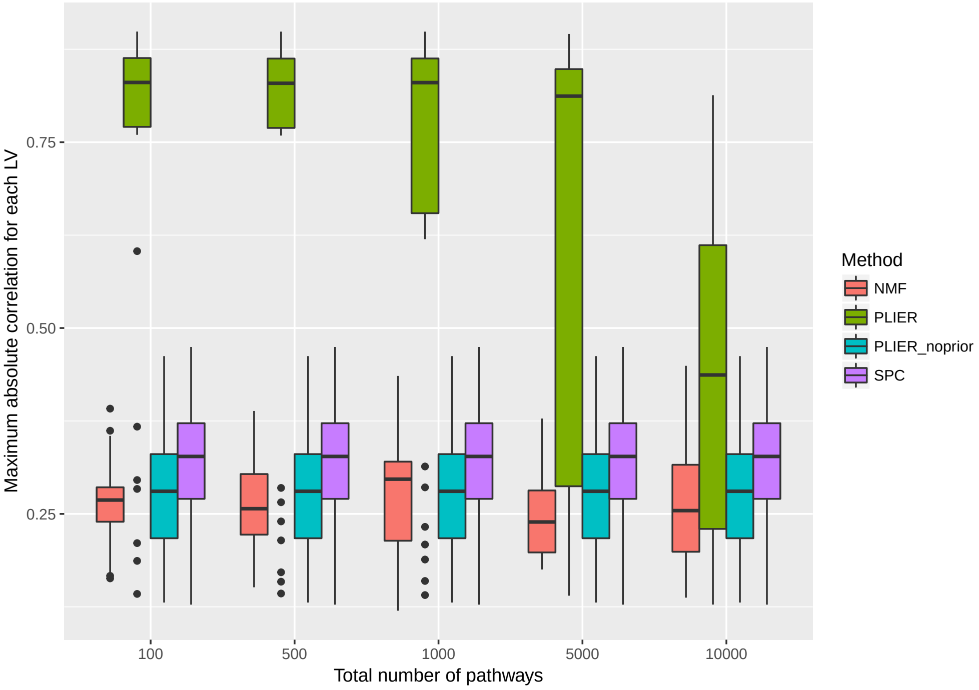


Figure 12: Data is simulated as in Fig. 11 except that the number of genes per pathway is kept at 300 and the number of uninformative pathways is varied. As the prior information gets noisy, PLIER's performance approaches those of others. Statistics were computed using Pearson correlation across 300 samples. Boxplot displays the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest.

Our basic evaluation strategy is based on computing the maximal correlations between simulated and recovered latent variables, and for the purpose of comparing with other methods, we use the absolute value so as to allow factors with reversed sign. Fig. 11 depicts the results of multiple simulation runs processed with four decomposition methods: PLIER, PLIER with no prior information (which can be accomplished by setting λ_3 to a high value), NMF [21] and SPC [165]. NMF is a popular decomposition method that is free of hyper-parameters (though different matrix norms can be used), however it requires positive data as input. SPC is another popular method that can enforce sparsity and positivity, and it has one hyperparameter that we set by cross-validation for each component as described in the original paper [165]. Among these methods only PLIER is able to reliably produce high correlations with the simulated latent variables and only when using prior information. Importantly, we emphasize that the simulation is not based on a PLIER model where we assume that loadings of genes in the pathway and outside the pathway differ by a constant factor but is rather based on the NMF model. Nevertheless the PLIER approach is effective even in the case where the model design differs from the underlying assumptions.

We also investigate how adding noise to the prior information affects performance, hypothesizing that as more irrelevant geneset are included in our prior knowledge matrix C, the advantage of using prior information will be reduced. Repeating the experiment above with varying sets of non-informative pathways we find that the performance indeed drops off as the total number of pathways is increased to 10,000. Though even at that level of prior-information noise, PLIER outperforms other methods (Fig. 12).

2.3.2 Pathway recovery significance

We estimate the significance of LV-pathway association by removing a random 1/5 of the genes annotated to each pathway prior to running PLIER. For each LV-pathway correspondence represented as a positive value in U, we compute the AUC and p-value (Wilcoxon rank-sum test) for the recovery of that pathway in the loadings of Z using the held-out set of genes as positive labels and genes not annotated to this pathway as negative labels. We verify that this procedure produces correct estimates by running PLIER with the geneset collection used for the DGN dataset but randomly permuted gene labels. Gene-level permutation preserves the pathway size distribution and dependency structure but should not have any non-random associations with the structure of the gene expression dataset. We find that in the permuted setting our cross-validation procedure produces uniformly distributed p-values (Fig. 13).

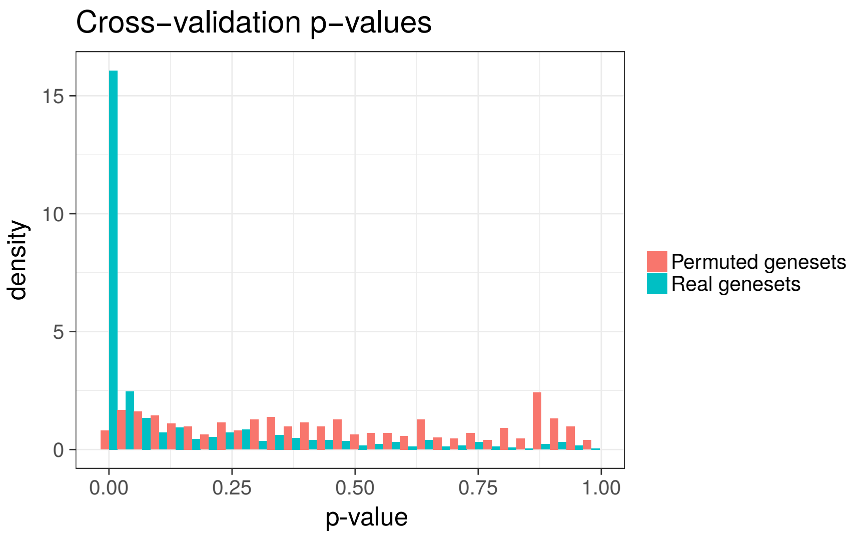


Figure 13: Histogram comparisons of pathway association p-values produced with real genesets and a single run with gene-label permuted genesets. Statistics were calculated the held-out set of genes and genes which are not annotated to the pathway. P-values are calculated with a two-sided Wilcoxon rank-sum test. Uncorrected p-values are plotted in the histogram.

2.3.3 Parameter robustness

The PLIER framework contains 4 free parameters. While we have a procedure for selecting these parameters automatically, it is natural to ask to what extent these affect the results. Using our becnhmarking dataset we systematically evaluate the robustness of LVs recovered at different parameter settings. Our evaluations is two fold: Firstly, we evaluate how well we recover the known cell-type proportions (LV vs. ground truth) for the LVs that are associated with proportion variables. Secondly we evaluate the stability of the LVs themselves with different parameter settings. The results are depicted in Fig. 14(a).

We find that many LVs are recovered with near-perfect correlation across a wide range of parameters. However, even in cases where the LVs themselves are variable (as is the case with the Dendritic cell LV), the actual correlation with known proportions is quite stable. While the results are stable across a parameter range around the default values, we find that increasing the L1 and L2 parameters beyond the stable range drastically alters the result (Fig. 14(a), left panel, bottom rows) and produces non-informative LVs (Fig. 14(a), right panel, bottom rows).

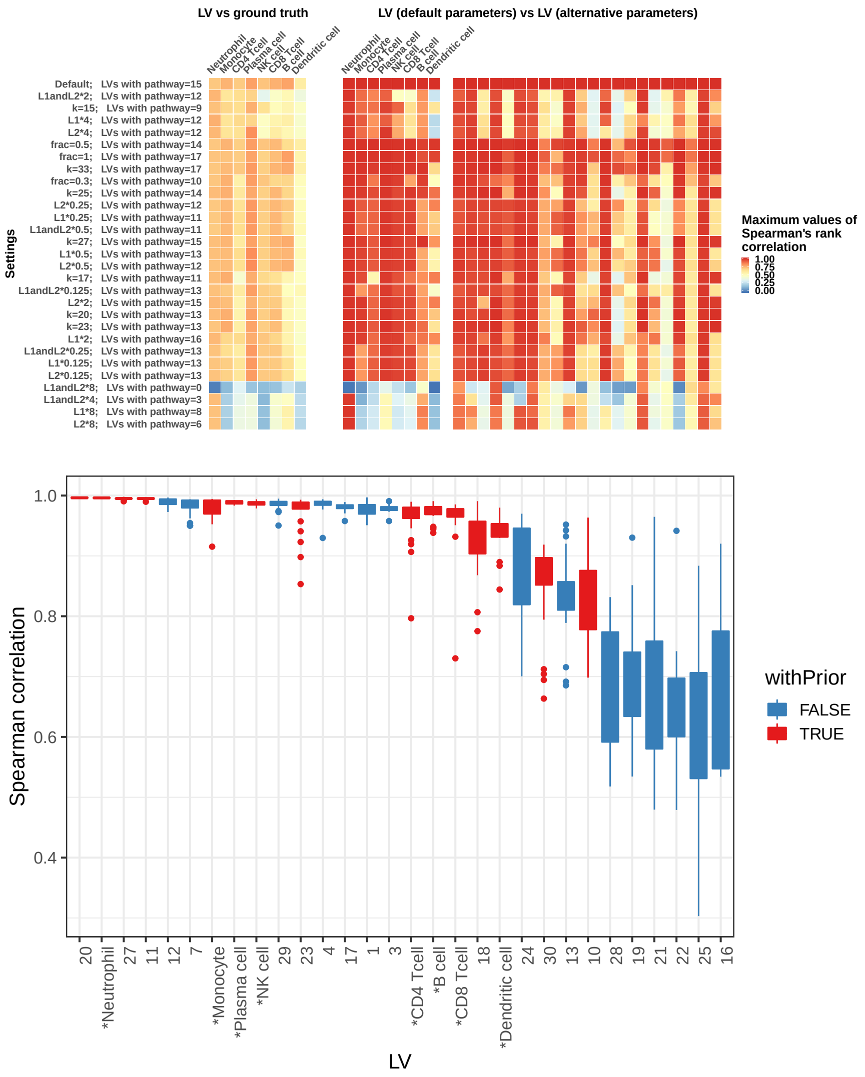


Figure 14: **A.** Robustness of LVs with respect to different parameter choices. Row labels indicate the parameter settings and the number of significant pathway associations. The L1 and L2 parameters are reported relative to the default. (Left panel) Maximum rank correlation of LVs with the ground truth cell-proportion measurements at different parameter settings. Statistics were computed across 35 subjects. (Right panel) Each column corresponds to one of the 30 LVs recovered at the default setting. The heatmap colors indicate the best correlation between the default LVs and those extracted from other parameter settings. First eight columns correspond to LVs that are related to cell type based on correlation with the ground truth. Statistics were computed across 35 subjects. **B.** Robustness of LVs with respect to random initialization. Statistics were computed using Spearman rank correlation across 35 subjects.

The PLIER problem is not convex and thus different initializations will produce different results. While the default initialization is to use SVD, we investigate to what extent the same LV structure can be robustly recovered using random intializations (Fig. 14(b)).

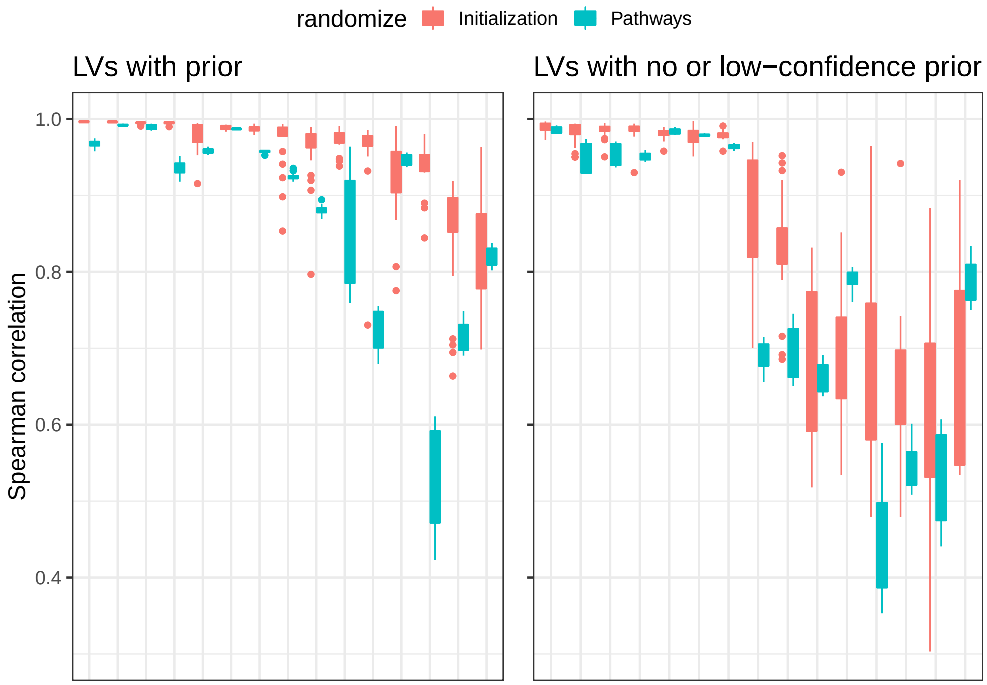


Figure 15: Robustness of LVs with respect to pathway randomization compared to robustness of LVs with respect to random initialization. Statistics were computed using Spearman rank correlation across 35 subjects.

Overall we find that almost all LVs with credible prior information association (FDR<0.05, red boxes) were recovered consistently. In particular LVs correlated with the known cell-type measurements (indicated by *) are highly consistent. LVs that are not linked with prior information (LVs with zero U coefficients) are less likely to be consistently recovered.

We can also test how much the final LVs depend on the pathway input by randomizing gene-pathway assignments. The results of this randomization are plotted in Fig. 15. We find that as expected randomizing pathways indeed has a greater effect on the results than randomizing the starting point, indicating that the prior information provides a considerable constraint.

2.3.4 Technical variation invariance

A key motivation for PLIER is to tease apart technical and biological variation. Specifically, the hypothesis is that LVs that use prior information are indeed of biological origin. If that is the case, we expect that PLIER results are relatively insensitive to normalization for technical factors and we test this hypothesis by applying PLIER to differently normalized versions of data. The DGN datasets [16] used in this study has been normalized for technical variables which reflected information about data collection and RNAseq quality control. We can also apply PLIER to the "naive-normalized" version of the same data represented by log-transformed counts normalized by quantile normalization. Obtaining two different decompositions, we find that many LVs can be matched in one-to-one correspondence based on rank correlations of the loadings. Correlations for top matched pairs are show in Fig. 9. Moreover, the matching LVs use prior information genesets that are either the same or closely related (see row/column names in Fig. 9).

Furthermore, when we compare the entire distribution of best matched correlations for LVs with- and without- prior information, as expected, LVs with prior information (LVs with non-zero U coefficients) produce best matches with higher correlations supporting the hypothesis that these captured biological variations are therefore relatively *normalization invariant* (Fig. 9, Inset).

2.3.5 Distributions of PLIER loadings

We plot loading statistics from our analysis of the DGN dataset in Fig. 16. The PLIER model doesn't assume pathway-level sparsity but rather that the loading values for pathway-associated genes are higher than those of others. Consequently, PLIER doesn't produce strict pathway-level sparsity but rather loadings with many values close to 0 and a long tail (panel B). We found that for this already regularized model including additional group-level of gene-level sparsity was not helpful when validated against known ground truth. Thus, genes not associated with the pathways can still get non-zero loadings, however we view this as a feature because it can provide useful "pathway-completion" information. We exploit this fact to compute properly calibrated p-values for LV-pathway associations using cross-validation (see Sec. 2.3.2)

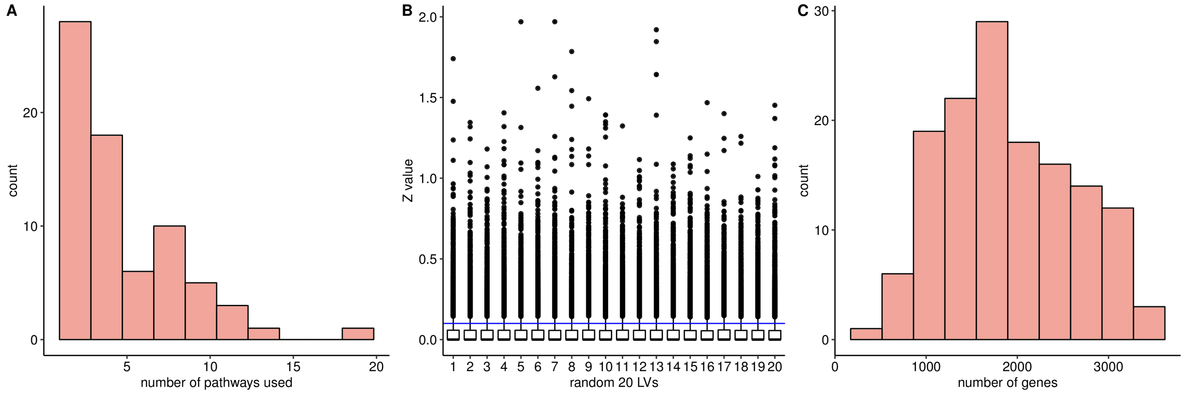


Figure 16: **A.** The distribution of number of LV-associated pathways per LV. **B.** The boxplot of loading value corresponding to 20 random LVs. Boxplot displays the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest. **C.** The distribution of number of genes with loading values above 0.1.

2.3.5.1 LV interpretation and naming The top genes contributing to each genotype-associated LV are depicted in Fig. 17. In many cases the identity of the genes and the corresponding PLIER pathway utilization (see U matrix visualized in Fig. 22) points to a clear cell-type effect (LV44, LV133, LV56) or a canonical pathway (LV21, LV40, LV97, LV120). In these cases the LVs can be interpreted as estimating the specific cell-type proportion or pathway-level effect and are named accordingly.

In some cases the pathway utilization did not allow for unambiguous interpretations. For example, the top pathway for LV16 is ”NKA1”, which is a NK-cell marker gene list. However the top genes in the LV loadings do not correspond to ”canonical” NK-cell markers. This pattern is instead observed for LV30 which also makes use of NK pathways. Thus, LV16 cannot be interpreted as NK cell proportion though its pathway utilization suggests some relationship to NK cell biology. We also note that two of the LVs that have some of the strongest genotype associations do not use any pathway information. We hypothesize that collectively these LVs most likely represent transcription pathways that are not well annotated in our prior information though they may correlate with some prior information genesets.

Nevertheless, these transcriptional pathway potentially have some cell-type origin and we investigate this by checking the bias in cell-type expression in a large independent dataset

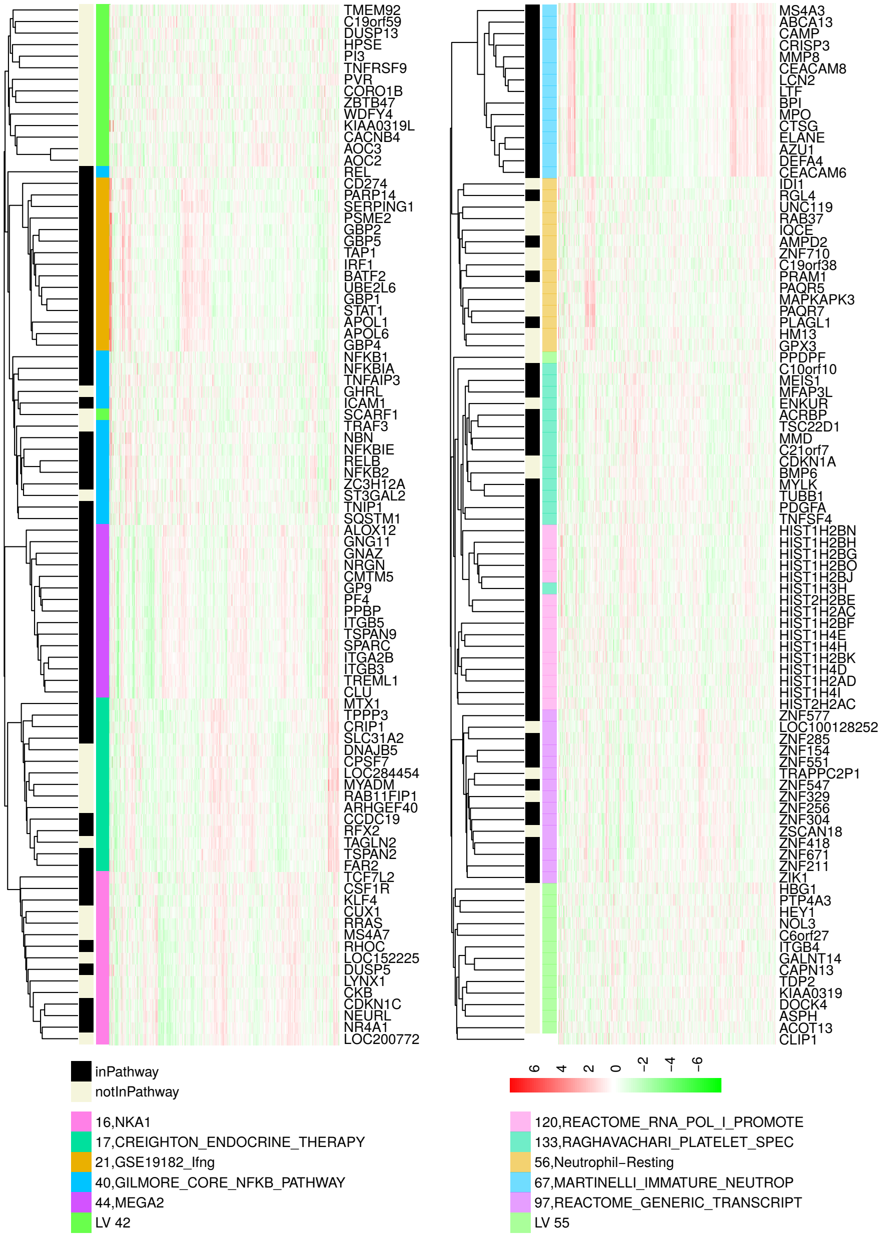


Figure 17: Top genes for all genotype-associated LVs. We plot the top 15 genes for all LVs that had a significant genotype association. Data is plotted as z-scores across 922 subjects.

of immune cell types, ImmGen [60]. The results are visualized in Fig. 17. We find that the top genes for LV16 are biased towards higher expression in myeloid and ILC cells which is consistent with being related to NK-type expression signature. LVs 17, 42 and 56 are likewise biased towards myeloid cell types. This is highly consistent with the effects of the putative *cis* drivers (NEK6, PLAGL1 and IKZF1 respectively, see Table. 4) on proportions of various myeloid cell types as determined in a large GWAS study of blood cell-type composition [14]. LV55 has no identifiable signature in ImmGen data, however it is biased for genes expressed in the erythroid lineage based on DMAP (Differentiation Map) dataset [122]. Top genes include HGB1 (rank 5) and HGB2 (rank 16) – fetal hemoglobins that are expressed but not made into protein. Moreover the putative *cis* driver for LV55 eQTL is NFE2 which is a transcription factor known to be involved in erythrocyte and megakaryocyte development.

2.3.6 Cell-type proportion inference

We validate the method using cell-type proportion inference because it is an important objective. Other methods are available for comparative benchmarking, and predictions can be tested against a direct measurement as gold standard. For this purpose, we generated a validation dataset comprising 35 human whole-blood samples assayed both by RNA-seq and direct CyTOF measurement of cell type proportion. We applied PLIER to the validation dataset using 605 pathways which included 60 cell-type markers and 545 canonical pathways from MSigDB [153]. We produced a decomposition with 14 latent variables annotated with high confidence (AUC >0.7, FDR < 0.05, see Methods for cross-validation procedure) to one or more genesets, of which 8 represented cell types also measured by the CyTOF panel. The correlation between the cell-type PLIER LVs and CyTOF measurements in these 35 samples had a mean of 0.71 (range 0.58-0.78) (Fig. 8).

We compared PLIER against the current established methods for mixture decomposition inference. These methods either rely on low-rank matrix decomposition or reference-based approaches that fit gene expression values to cell-type specific signatures. We include the most widely used constrained matrix decomposition approaches: Non-negative Matrix Factorization (NMF), and Sparse Principal Component Analysis (SPC) (see Methods for details). For a reference-based approach, we tested Cibersort [119] and NNLS [1]. Both of these approaches combine a regression algorithm with a dedicated cell-type specific refernce matrix that is explicitly optimized for human blood deconvolution.

PLIER performed considerably better than other constrained matrix decomposition methods and surprisingly outperformed the refernece-based supervised approaches on 4 out of the 8 cell-types. The excellent performance of the essentially unsupervised and general PLIER method is in part due to the capacity of PLIER to sort through many candidate genesets and find the ones most informative for the specific dataset. PLIER can be supplied with multiple and even discordant markers sets for the same cell-type and will automatically pick the one that models the data.

2.3.7 Genotype-quantitative trait association

While PLIER shows excellent performance when benchmarked for cell type deconvolution, it is not specifically designed for this task. Instead, it is a general method for estimating pathway activity that it is applicable to a wide variety of gene expression interpretation problems.

As an example, we evaluated the usefulness of PLIER for the difficult task of genotype-quantitative trait association. Two groups of eQTLs are typically distinguished: locally acting *cis*-eQTLs that affect a nearby gene, and *trans*-eQTLs that are commonly mediated at the pathway level [16]. Many *trans*-eQTLs exert their effect by altering the activity of a regulatory protein, which in turn affects the expression of many downstream genes [163]. *Trans*-eQTLs, which provide important insight into gene regulatory networks, are difficult to detect and are less commonly identified than *cis*-eQTLs due to the multiple hypothesis burden of testing millions of variants by tens of thousands of genes.

We analyze the recently published DGN dataset [16], which contains whole blood RNA-seq and genotype measurements from 922 individuals, to demonstrate how the PLIER framework extracts a broad spectrum of pathway effects and enables network-level eQTL discovery and interpretation. For the candidate prior information, we used a comprehensive collection of 4,445 genesets comprising biochemical and transcriptional pathways (”canonical pathways” and ”chemical and genetic perturbations” from MSigDB [153]), cell-type markers from multiple sources [119, 1, 122] and cytokine signatures [43]. The PLIER decomposition produced 86 LVs that have at least one matched pathway with an FDR < 0.05, and were associated overall with 318 of the 4,444 pathway genesets evaluated. The decomposition captured cell-type variation with a high degree of specificity, differentiating naive and memory B-cells, plasmacytoid and myeloid dendritic cells, and multiple subtypes of CD8 T-cells. PLIER also captured variation in non-leukocyte cell types such as megakaryocytes and erythrocytes, and transcriptional pathways such as Type I and Type II interferon signaling, and NKFB pathway. Overall, we find that 29 LVs were unambiguously related to cell type, canonical pathways or cytokine signaling (see Fig. 18 for U matrix visualization and Table. 10 for a complete list of LV-geneset associations).

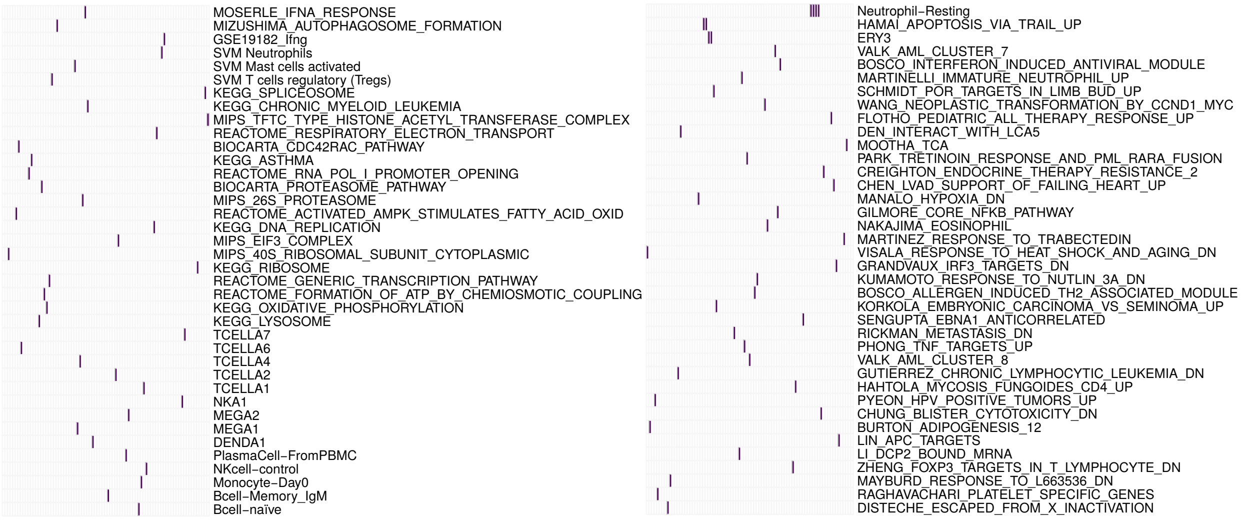


Figure 18: Pathway associations for the DGN dataset. We plot the U matrix for the complete set of LVs that are associated with at least one prior information pathways with and FDR <0.05. LVs correspond to columns and pathways correspond to rows. For ease of visualization, only the top pathway (largest U coefficient) for each LV is retained and U matrix is split into two columns. We note that pathways for abundant cell types such as neutrophils and erythrocytes are top hits for multiple LVs.

In order to perform eQTL analysis, we treated the PLIER LVs as quantitative traits (see Methods for details), and identified 12 LVs showing significant associations with genotypes (Table 4, see Methods for details). In contrast to gene level *trans*-eQTLs, the PLIER eQTLs are pathway-level effects that capture the concerted behavior of multiple genes (Fig. 22A). The gold-standard for eQTL discovery is reproducibility in an independent dataset. As each pathway-level eQTL effect is supported by a number of gene-level effects we can directly compare the gene-level replication rates of standard (gene-centric) *trans*-eQTLs and pathway-centric analysis which only considers gene-level eQTLs if they also correspond to pathway-level eQTLs (see Methods for details). Using an independent dataset of human blood expression data assayed with Affymetrix microarray [166] we compared the true-positive rate, π_1, (see Methods) for gene-centric and pathway-centric eQTLs and find that the pathway-centric eQTLs are more reproducible at every p-value threshold. For example, at a cutoff that corresponds to gene-level FDR of 0.2 the gene-centric π_1 is ≈ 0.2 while for pathway-centric eQTLs it is ≈ 0.6 (see Fig. 19 for replication across a range of cutoffs).

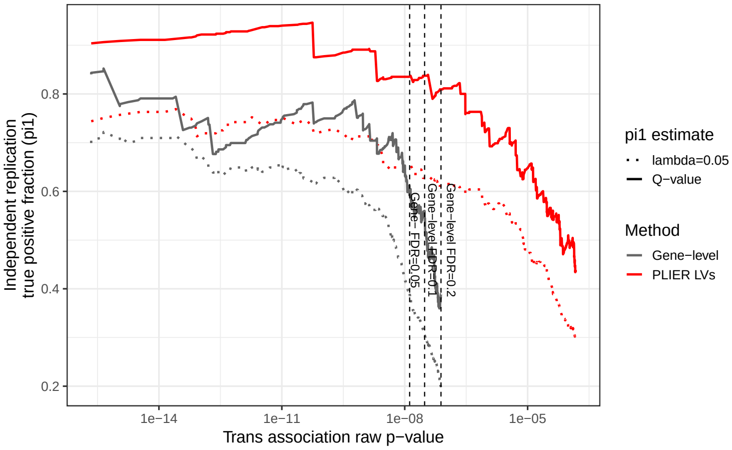


Figure 19: Pathway-centric PLIER eQTLs have a higher replication rate. We compare trans-eQTL discovered with the standard gene-centric approach to those discovered by PLIER with respect to independent replication in the NESDA dataset. While gene-centric approach considers all possible SNP-gene associations that satisfy valid *trans*-eQTL criteria (see Methods), the pathway-centric approach only considers those gene-level effects that are associated with pathway-level eQTLs. We find that at the same raw p-value threshold, pathway-centric eQTLs have a notably higher replication rate. Since pathway-centric associations are by construction linked to a pathway-level effect, they are more likely to represent real and replicable indirect associations. Statistics were computed using Spearman rank correlation across 922 subjects with a two-sided test. P-values indicated on the x-axis are uncorrected.

Besides improving the accuracy of *trans*-eQTL discovery, the PLIER decomposition identifies the pathway(s) associated with the LV-eQTL, which can provide precise biological interpretation of the genetically regulated processes. For example, PLIER shows that SNP rs1354034 (located within gene ARHGEF3) is associated with two LVs, LV44 and LV133, that are related to megakaryocyte/platelet lineage based on their pathway association (Fig. 22A, B). In the published gene level analysis of the DGN dataset, this SNP yields the largest number of significant *trans*-eQTLs, however no biological interpretation was inferred [16]. Using PLIER, we find two of the associated LVs are annotated to platelet pathway processes, which is consistent with a known effect of this SNP on platelet number (PLT) and platelet volume (MPV) [52]. However, our analysis further shows that the two

LV id	LV name	snps	cis-Gene(s)	Benjamini-Hochberg FDR
44	Mega/platelet 1	rs1354034	ARHGEF3	1.707e-41
133	Mega/platelet 2	rs1354034	ARHGEF3	0.03095
120	Histones	rs1354034	ARHGEF3	0.0336191
97	Zinc fingers, pseudogenes	rs1471738	SENP7	4.011e-13
56	PLAGL1 associated, myeloid	rs9321957	PLAGL1	0.0001421
42*	IKZF1 associated, myeloid	rs10251980	IKZF1	3.39e5-61
17	NEK6 associated, myeloid	rs16927294	NEK6	0.008223
67	Neutrophils	rs13289095	PKN3,SET,ZDHHC12	0.03361
55*	NFE2 associated, erythrocyte	rs35979828	NFE2	3.538e-10
21	Interferon-gamma	rs3184504	SH2B3	0.0002198
40	NFKB/TNF	rs12100841	PPP2R3C	0.005094
16	Myeloid/ILC	rs1138358	BCL2A1,MTHFS,ST20	0.0008103

Table 4: Summary table of all pathway-level effects found in the DGN dataset. Statistics were computed using Spearman rank correlation across 922 subjects with a two-sided test. False discovery rates are computed using the Benjamini-Hochberg procedure on the total number of tests (number of LVs × number of SNPs). SNP-LV associations that passed FDR<0.05 were further filtered to account for potential *cis* genes or mismapped *cis* homologs contributing to the LV estimtate (see Methods for details). In most cases pathways were named based on their geneset association captured in the U matrix. Some pathways are named based on further analysis of the expression patterns of top gene in an independent dataset of mouse immune cells, ImmGen [60] (see Fig. 20) and/or a the presence of a putative *cis* eQTL transcriptional mediator. The complete pathway utilization for these LVs can be seen in Fig. 22. The expression patterns for top 15 genes driving each latent variable are plotted in Fig. 17. Latent variables with no pathway association in PLIER decomposition (that is no positive entries in U) are starred.

LVs linked to this SNP are supported by different genes that show distinct expression patterns (Fig. 22B). These results suggest that the two LV-eQTLS may distinguish two different processes of platelet/megakaryocyte biology. A recent hematopoetic lineage report supports this formulation. This single cell study shows that genes associated with the two LVs express at different developmental time points [124]. Specifically, mouse orthologs of MEIS1 and TSC22D1 (from LV133) are expressed in all megakaryocyte precursors, while ITGA2B (from LV44) is megakaryocyte specific, suggesting that these two LVs capture processes that are active at different times in megakaryocyte development.

LV133 and LV44 are positively correlated with each other in the DGN dataset. Notably, the effects of the rs1354034 alleles on LV133 and on LV44 go in opposite directions (Fig.

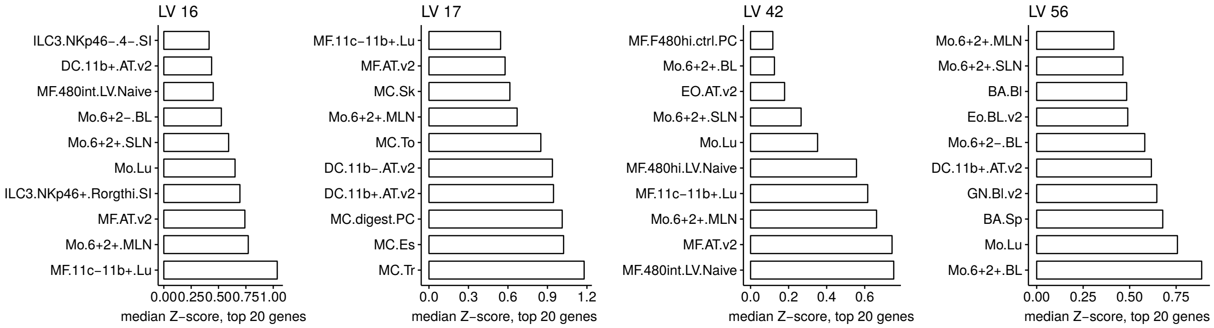


Figure 20: Top scoring Immgen cell types for genotype-associated LVs with no or ambiguous PLIER pathway annotations. In order to add further interpretation to the pathway-level eQTLs that had either no or ambiguous pathway associations, we investigate the gene expression of top loading genes in a comprehensive database of mouse immune cells, Immgen [60]. This database was not used in the PLIER decomposition, so it provides an independent source of immune gene expression patterns. Z-scores were computed across top 20 loading genes.

22C). Furthermore, we find that using partial correlation analysis, whereby the LVs are corrected for each other, dramatically improved the eQTL statistics (Fig. 21). These results strongly argue that the LV44 and LV133 effects are independent.

phenotype	reported SNP	Close gene	LV44 p-value	LV 133 p-value	proxy SNP
MPV	rs10876550	COPZ1	**1.1847e-05**	0.69933	rs10876550
PLT	rs2911132	ERAP2	0.13817361	**2.4417e-05**	rs2549803

Table 5: Summary table of the associations between the two mega/platelet LVs and SNPs known to affect only one platelet phenotype. Statistics were computed using Spearman rank correlation across 922 subjects with a two-sided test. Raw *p*-values are reported. A total of 80 SNPs with known platelet phenotypes were tested [52]. While no SNPs outside the ARGHEF3 locus achieved genome-wide significance, some associations were significant at FDR<0.05 when we consider only the 160 (80 SNPs × 2 LVs) hypotheses that are tested (significant *p*-values are in bold). We find that the associations of the two mega/platelet LVs with other loci known to affect platelet biology are distinct. Our analysis suggests that the early mega/platelet LV (LV133) is more closely related to the process controlling platelet number (PLT) while the late mega/platelet LV (LV44) is related to the process controlling platelet volume (MPV).

We speculate that the independent regulation of the two LV-eQTLs by the same locus results from an effect on different regulators that are modulated at different periods of megakaryocyte development. The rs1354034 SNP is known to be pleiotropic as it is linked to both MPV and PLT phenotypes, which are affected independently by other genetic vari-

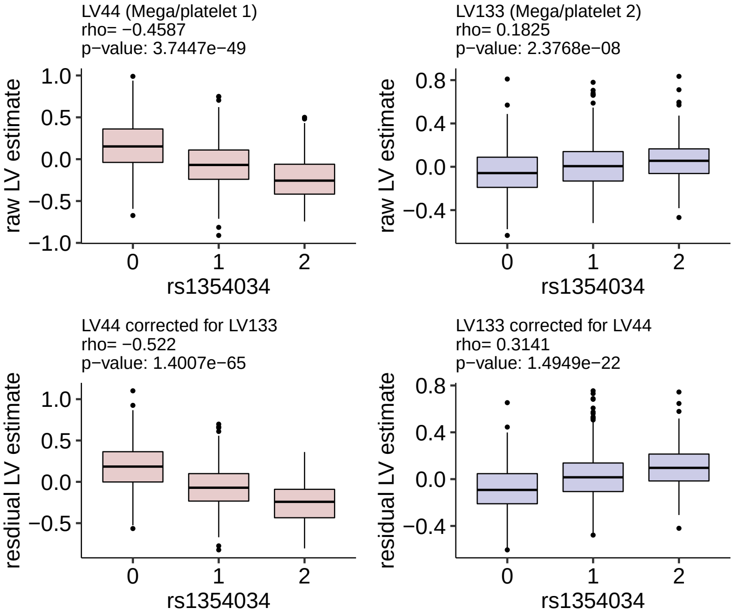


Figure 21: The effects of rs1354034 of LVs 44 and 133 are independent. We plot the relationship between two platelet/megakaryocyte LVs and minor allele counts of rs1354034 using raw LV estimates (first row) or corrected estimates – residuals from the linear regression fit on the other LV (second row). We find that while the estimates for these two LVs are positively correlated, the eQTL effects are substantially improved when regressing one LV on the other and using the residuals for eQTL testing. Statistics were computed using Spearman rank correlation across 922 subjects with a two-sided test and uncorrected *p*-values are reported. Boxplots display the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest.

ation [52]. We hypothesize that the effects of rs1354034 on multiple LVs is reflective of its pleiotropic function. Indeed, correlation of the two LVs with SNPs known to be specifically linked to MPV or PLT alone shows divergent patterns. In addition to the association with rs1354034, the developmentally early LV133 is most strongly associated with a SNP linked to platelet number, whereas the later LV44 is most strongly associated with a SNP linked to platelet volume (Table 5). This analysis supports a model where ARGHEF3 exerts its pleiotropic affects on platelet volume and number at different developmental time points. These results demonstrate how PLIER can leverage dataset structure and external knowledge to resolve fine-grained mechanistic insight underlying complex biological processes. Additional demonstrations of how PLIER can be applied to single-cell RNA-seq or cross-study concordance analysis are presented in Sec. 2.3.9 and 2.3.10.

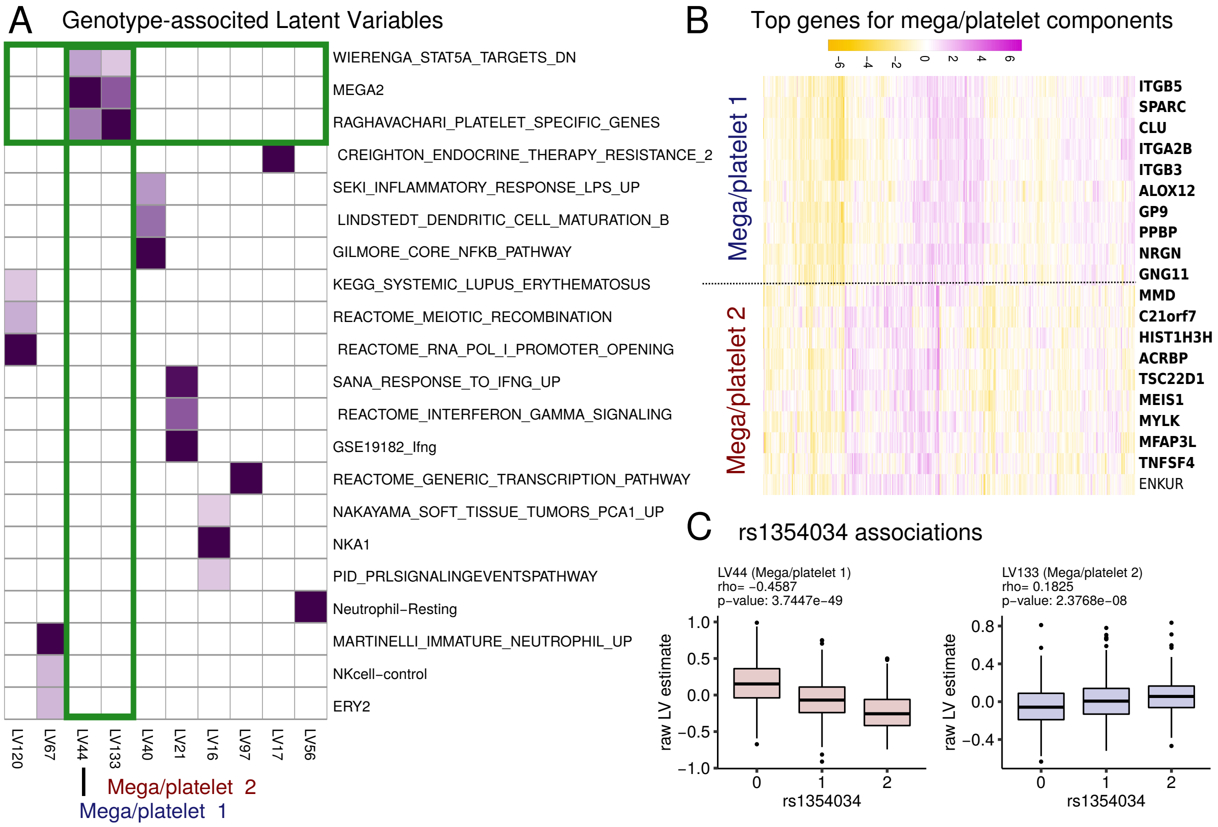


Figure 22: **A.** A heatmap of a subset of the U matrix corresponding to LVs with a genotype effect (LV-eQTLs). Only pathways with a cross-validation FDR of < 0.05 are shown. We find that two latent variables (LV44 and LV133) share pathway annotations (albeit with different coefficient) that suggest a relationship with megakaryocyte and platelet biology. **B.** Heatmap of the top genes in the loading for LV44 and LV133. Genes that are annotated to the pathways shown in panel A are in bold. **C.** Boxplots of the association of LV44 and LV133 with SNP rs1354034 (n=344, 429, 149 for 0, 1, 2 respectively) While the LV estimates are positively correlated, the effects of rs1354034 are opposite. These results indicate that the pathways captured by the expression patterns of LV44 and LV133 are independently regulated by the rs1354034 locus. Boxplot displays the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest. P-values indicate unocrrected two-tailed Spearman rank correlation test.

2.3.8 Comparison of methods for pathway-level eQTL discovery

We compared PLIER to other methods in its ability to recover pathway-level eQTLs. PLIER pathway-level eQTLs are deemed significant at Benjamini-Hochberg FDR < 0.05 (correcting for the total number of tests). The same raw p-value threshold is used for all other methods (even though the FDR at this threshold for alternative methods is higher). We consider only the best SNP for each latent variable and display the results of all eQTLs discovered as well as those filtered for gene-level support (see Methods). We find that PLIER indeed is able to find more associations while PLIER (no prior) and SPC perform comparably. NMF performed worse than SVD on this datasets and is thus omitted (Fig. 23). For this analysis, rather than using cross validation the SPC sparsity parameter was explicitly optimized to maximize the eQTL discovery objective.

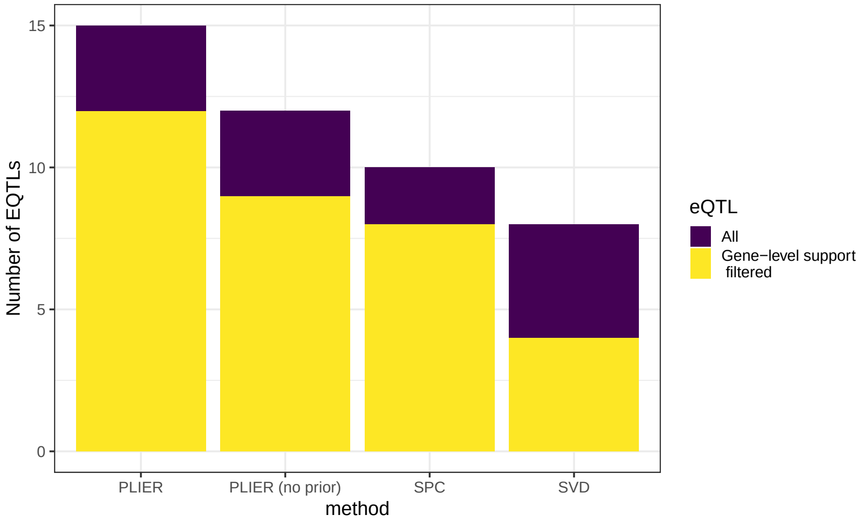


Figure 23: Comparison of eQTL discovery results from different decomposition methods.

2.3.9 Analysis of single-cell RNA sequencing dataset

While our approach doesn't specifically address the unique features of scRNA-seq data, it can already be applied out-of-the-box to single-cell data. We have applied PLIER to scRNA-seq data from mouse sensory neurons [159]. Despite the fact that the prior information database does not contain any genesets derived from sensory neuron sub-types, we find several latent variables that are associated with prior genesets with high confidence. Consistent with expectation, the pathways involved are related to neurological tissues and cell-type identity (Fig. 24A). Moreover, our approach also finds pathways that are independent of the major cell types (such as LV10). Because cell type is the dominant signal in the dataset, this pathway-level effect is not easily observed in raw gene expression data (Fig. 24B) but stands out clearly when correcting for other sources of variation (Fig. 24C). Thus, PLIER is able to both reveal additional heterogeneity in this complex dataset and associate it with prior information in a single computational step.

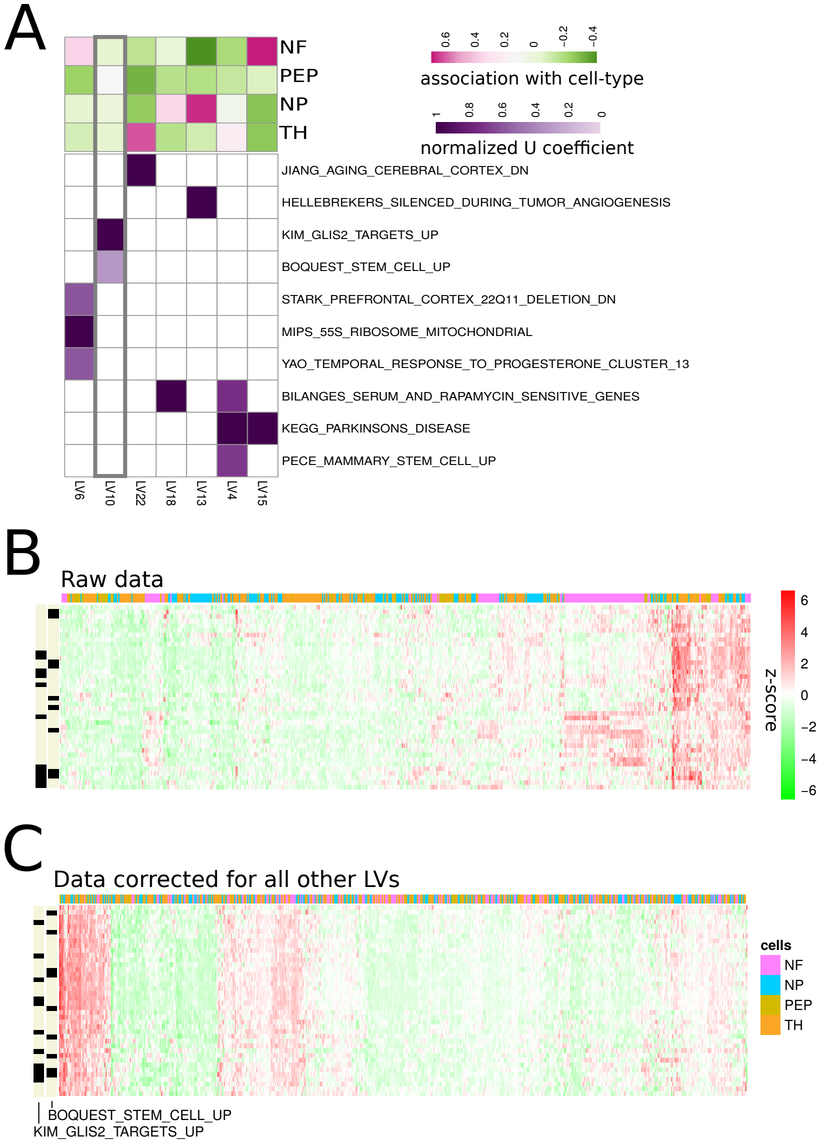


Figure 24: **A.** The subset of the U matrix with the highest-confidence (AUC>0.75, FDR<0.01) pathway associations. Spearman rank correlations with cell types (139, 169, 81 and 233 for NF, NP, PEP, and TH respectively) defined in the original paper are displayed above. While many of the LVs are correlated with cell-type identity, we find some pathways that are not strongly associated with cell types, such as LV10 (highlighted in grey). **B.** Gene-expression z-scores for the top 40 genes in LV10 across all cells are displayed in a heatmap with red indicating high expression. Pathway membership of individual genes is indicated with row annotations (black indicated annotation to the pathway) and cell types are indicated with column annotations. We find that when viewed in raw data space the top genes associated with LV10 show several patterns of expression and cluster according to cell type. **C.** Same data as in B corrected for all LVs except for LV10. The genes now show a single consistent pattern and no longer cluster by cell types.

2.3.10 PLIER models are transferable across datasets and can be used to improve concordance

One key feature of PLIER is that it extracts latent variables that correlate with prior information (LVs with non-zero U coefficients). PLIER LVs are thus less likely to depend on individual gene measurement and are more likely to reflect effects that are common across different studies. To illustrate this property, we have compared PLIER decompositions of the DGN dataset with that of the NESDA dataset. The NESDA dataset is also whole-blood but uses the Affymetrix platform and has considerably lower signal to noise ratio. Nevertheless, we find that applying PLIER decomposition to the two datasets yields surprisingly consistent results. In particular, many LVs can be matched across datasets based on gene-loading correlation and this matching is often one-to-one (Fig. 25A and B). Moreover, the matched LVs often use either the same or highly related prior information (Fig. 25B). Considering LVs that are best reciprocal hits as matched pathway-level estimates, we find that differential expression with respect to three demographic variables is more concordant in LV space than gene space (Fig. 25C).

2.4 Discussion

2.4.1 On the use of PLIER for mixture proportion estimation

We show that PLIER is competitive with the best available reference-based method (Cibersort) on mixture proportion estimation. Cibersort relies on known quantitative cell-type signatures. While SVM-based framework is robust to outliers and discrepancies, it is likely that the hard-coded Cibersort signature is not a good fit for our dataset. Even though the cell-type marker genesets used by PLIER are in part produced from the same source data [1, 122], there are two important distinctions. PLIER is considerably more tolerant of errors in marker genes since the model simply stipulates that we wish to find latent variables such that the loading values corresponding to the marker genes are higher *on-average* than the background, without specifying a target value. Moreover, since PLIER automatically

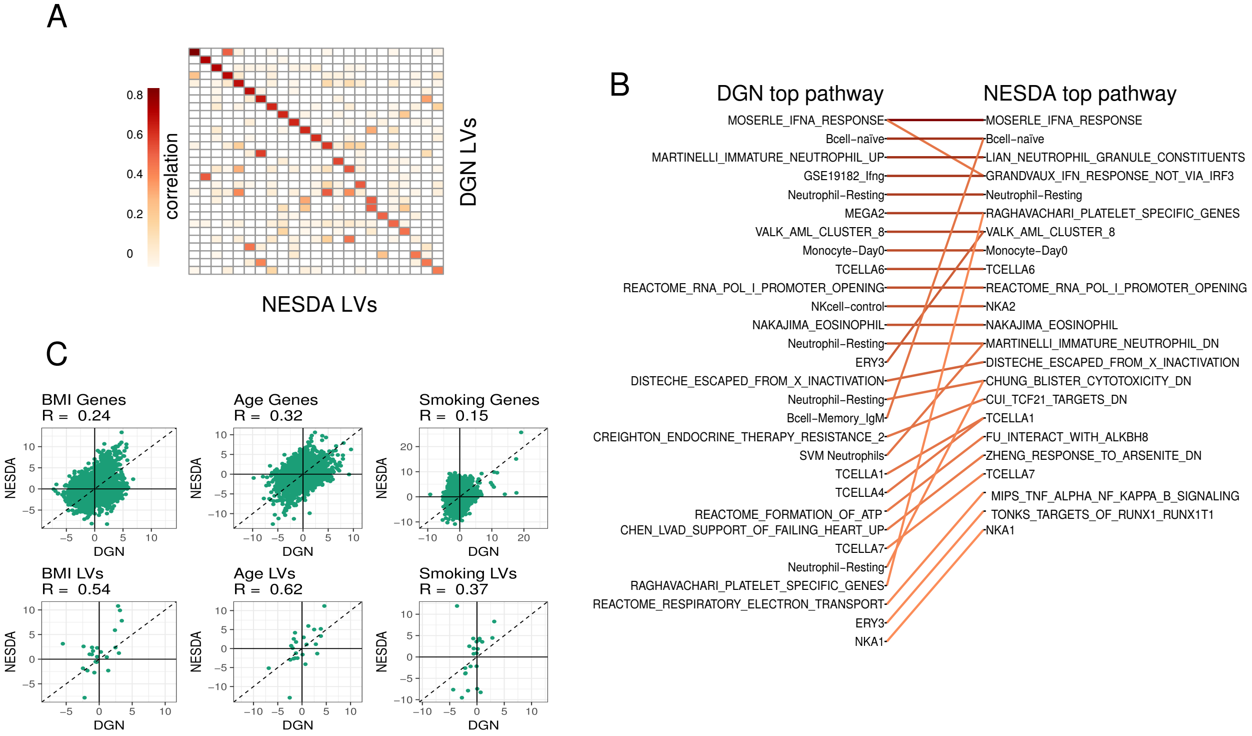


Figure 25: **LV-based meta-analysis increases cross-dataset concordance** Two whole blood datasets DGN (RNAseq) and NESDA (array) were independently decomposed using PLIER. We assessed the correspondence between the resulting LVs by comparing their loadings on the common set of genes. **A.** All pairwise loading correlations (across 10,550 common genes) among LVs that have at least one cross-dataset match with a correlation >0.5. We observe a strong "sparse" pattern with few LV pairs achieving a high correlation. Statistics were computed using Spearman rank correlation. **B.** Pairs of LVs that have a correlation of >0.3 are depicted as a bipartite graph. Each LV is automatically named by the top pathway that supports it. The LV order corresponds to panel A (top to bottom for DGN and left to right for NESDA). We note that many LVs are in one-to-one correspondence though some LVs that are distinct in one dataset collapse to a single related LV in the other. For example, naive and memory B cells are resolved in DGN but correspond to a single B cell LV in NESDA. This is also the case with the two platelet-related pathways (MEGA2 and RAGHAVACHARI_PLATELET_SPECIFIC_GENES). Overall, while the two datasets are decomposed independently, the resulting decompositions align well and the aligned LVs often have either identical or highly related top pathways. **C.** We define a one-to-one LV mapping by only using pairs in B that are best reciprocal hits. This allows us to align the two datasets in LV space analogously to alignment by gene identity. Given aligned representations we investigate the differential expression concordance with respect to three demographic variables. Each sub-panel depicts a scatter-plot of gene or LV T-statistics (922 and 1,848 indidividuals for DGN and NESDA respectively) for the variable of interest. We find that the concordance of differential expression (as measured by Pearson correlation of the T-statistic) is dramatically increased in LV space.

selects a few relevant pathways out of hundreds or thousands of available ones, it can be supplied with multiple and possibly discordant marker sets for the same cell type.

It is important to note that the purpose of PLIER is general pathway-activity estimation.

We do not expect that PLIER will substitute reference-based methods for the explicit task of mixture component inference where reference-based methods have several conceptual advantages. For example, PLIER operates best on z-scored data and thus by default discards valuable information about total transcript abundance. Moreover, PLIER is only applicable to relatively large datasets. In particular the number of major variance components, that can not be greater than the number of samples (and is typically much less), must be at least the number of mixture components we would like to estimate. Thus, PLIER cannot be applied for mixture component estimation in datasets with just a few samples, where reference-based methods should have a clear advantage. Importantly, performance of reference-based methods is highly dependent on the basis signatures (pure cell expression states) which may vary according to assay platform and processing pipeline. A basis signature optimized for a particular data acquisition framework will provide the optimal performance.

2.4.2 Alternative approaches

There are several methods that can take prior information about genesets into account in order to learn a biologically meaningful low-dimensional representation, for example, Bayesian Factor Analysis [25] that extracts pathway-level latent variables and our previously proposed method CellCODE [30] that estimates cell-proportion variation from cell-type marker genesets. However, these methods require that the genesets are specified *a priori* and that genes can be partitioned into these sets (though some overlap is allowed). In contrast, in our method the pathways themselves are subject to optimization and our method is designed to effectively choose just a few relevant genesets from thousands of available ones.

As our goal is to force gene loading to be represented by biologically coherent genesets, it is natural to seek a solution based on group lasso regularization, which can perform variable selection at the group level. However, given that the biological genesets are highly redundant and overlapping, group lasso, which requires non-overlapping groups, is unsuitable. While it is possible to define more complex norms that accommodate group overlaps, there are some drawbacks. For example, a related method termed structured sparse PCA [68] has been developed for image analysis. This method implements a direct optimization of the column support, but can only constrain the support to be the complement of a union of predefined groups, which corresponds to rectangle-bounded regions for images, but is not interpretable for genesets. Another related method that considers biological genesets explicitly is the Overlap Group Lasso which employs an alternative norm that enforces the biologically desirable union-of-groups support [123]. However, the implementation is computationally expensive on large numbers of groups and its native form does not explicitly deal with the issue of geneset/pathway incompleteness.

2.4.3 Future developments

Despite the promising results there are a number of areas for potential improvement and our future work will center on improving the recovery of LVs with only a few supporting genes as well as improving performance on very large geneset collections. For example, even on simulated data we find that increasing the amount of irrelevant prior information degrades the method's performance. On the other hand, the available prior information represented in geneset databases such as mSigDB is constantly increasing which makes robustness to large prior information collections a top development priority.

3.0 DataRemix - a universal data transformation for optimal inference from gene expression datasets

3.1 Introduction

Genome-wide gene expression studies have become a staple of large-scale systems biology and clinical projects. However, while gene expression is the most prevalent high-throughput technology, technical challenges remain. Raw gene expression values must be normalized for any technical and nuisance biological variation and the normalization strategy can have dramatic effects on the results of downstream analysis. This is especially true in cases where the sought-after gene expression effects are likely to be small in magnitude, such as expression quantitative trait loci (eQTLs). Increasingly sophisticated normalization methods have been proposed and many are computational intensive and/or can have multiple free parameters that must be optimized [88, 148, 101, 70, 116]. Moreover, it is not uncommon for one dataset to yield multiple normalized versions that maximize performance in a particular setting (such as the discovery of *cis*- and *trans*-eQTLs [16]), highlighting the complexity of the normalization problem.

Singular value decomposition (SVD) is one of the most widely used gene expression analysis tools [5, 6] that can also be used for data normalization. Using the SVD we can simply remove the first few principal components that are presumed to represent technical factors such as batch effects or other nuisance variation. In some cases this dramatically improves downstream performance, for example in the case of eQTL analysis [116]. The drawback of this method is that the exact number of components to remove must be determined empirically and some meaningful biological signals may be lost in the process.

More sophisticated approaches attempt to partition data structure into true biological and nuisance variation and remove only the latter [88, 148, 101, 70, 116]. These can improve on the naive SVD-based normalization but require additional input such as technical covariates, or the study design. The success of these methods ultimately depends on the availability and quality of such meta data and some methods still rely on parameter optimization to maximize performance. These widely used normalization approaches all have a common theme that they rely in part on the intrinsic data structure. One key property that contributes to the success of these approaches is that for many biological questions of interest, nuisance variation (of technical or biological origin) is larger in magnitude than true biological variation. Our proposed method, DataRemix, explicitly formalizes this view of the data normalization problem.

In this work we demonstrate that biological utility of gene expression datasets can be dramatically improved with a simple three-parameter transformation, DataRemix. Our method does not require any dataset-specific knowledge but rather optimizes the transformation with respect to some independent *objective* of data quality, such as the quality of the gene-correlation network or the number of *trans*-eQTL discoveries. Because our method requires only the gene expression data and biological validity objective, it can be applied to any publicly available dataset. We focus our study on gene expression data for which methods for quantifying biological validity are well established, but our approach can be readily applied to any high-throughput molecular data for which similar quality metrics can be defined. We show that this strategy can outperform methods that make explicit use of dataset-specific factors, and can further improve datasets that have been extensively normalized via an optimized, parameter-rich model. We also show how the optimal parameters of DataRemix can be found efficiently by Thompson Sampling with a dual learning setup, making the approach feasible for computationally expensive objectives such as eQTL analysis.

3.2 Methods and Materials

3.2.1 The DataRemix framework

We formulate DataRemix as a simple parametrized version of SVD which can be directly optimized to improve the biological utility of gene expression data. Given a gene-by-sample matrix X, SVD decomposition can be thought of as a solution to the low-rank matrix approximations problem defined as:

$$\min_{U_k,\Sigma_k,V_k} \|X - U_k\Sigma_k V_k^T\|_F^2 \tag{3.1}$$

where U and V are unitary matrices. With the SVD decomposition $U\Sigma V^T$, the product of k-truncated matricies $U_k\Sigma_k V_k^T$ gives the rank-k reconstruction of X. We introduce two additional parameters p and μ to define a new reconstruction:

$$\text{DataRemix}_{\{k,p,\mu\}}(X) = U_k\Sigma_k^p V_k^T + \mu(X - U_k\Sigma_k V_k^T) \tag{3.2}$$

Here, k is the number of principal components of SVD and $p \in [-1, 1]$ is a real number which alters the scaling of each singular value. For $p = 1$, this approach reduces to the original SVD-based reconstruction . For $p = 0$, the transformation gives the frequently used whitening operation [45]. As depicted in Fig. 26, generally, different choices of p reweigh the contribution of each variance component, possibly making some low-variance biological signals visible while down-weighting technical and other systematic noise. The parameter μ is a non-negative weight that adds the residual back to the reconstruction in order to make the transformation *lossless*.

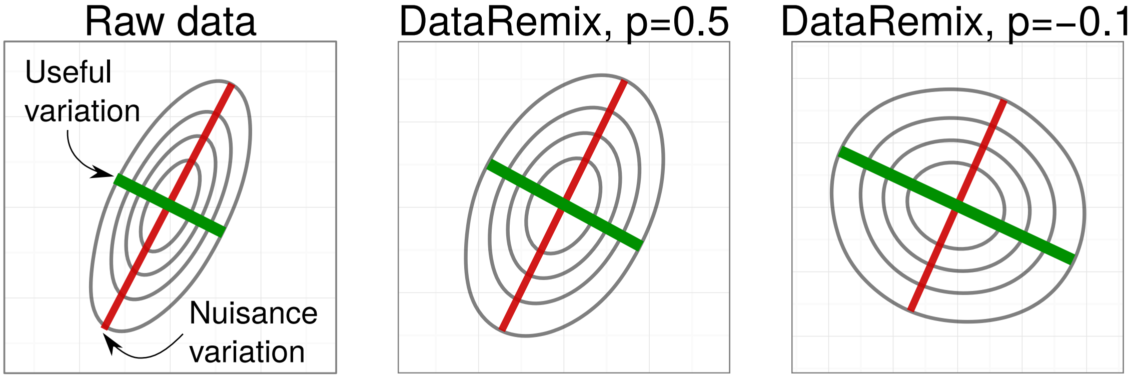


Figure 26: Visual representation of DataRemix transformation. We simulate a 2-dimensional dataset where the nuisance variation contributes more variance than true biological variation. Different power parameters p reweigh the contributions of the two variance axes, making the true biological variation more "visible".

Intuitively, we expect this approach to succeed because sophisticated normalization methods that use both data structure and some external variables, such as technical covariates, can be thought of as implicit regularizations on the naive SVD-based normalization (which simply removes the first k components), and this formulation simply makes this explicit.

The general workflow of DataRemix is shown in Fig. 27. The downstream biological objective depends on your study. For example, if you focus on *trans*-eQTL analysis,

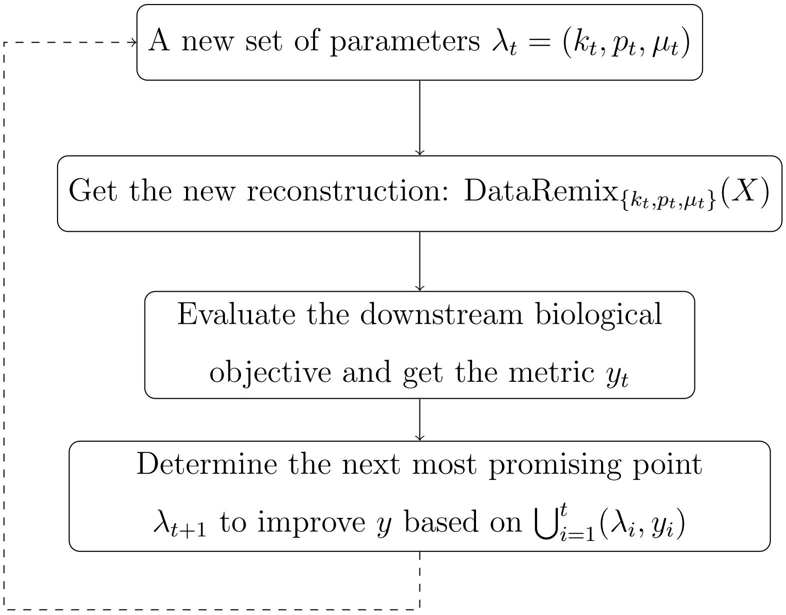

A new set of parameters $\lambda_t = (k_t, p_t, \mu_t)$
Get the new reconstruction: $\text{DataRemix}_{\{k_t, p_t, \mu_t\}}(X)$
Evaluate the downstream biological objective and get the metric y_t
Determine the next most promising point λ_{t+1} to improve y based on $\bigcup_{i=1}^{t}(\lambda_i, y_i)$

Figure 27: The workflow of DataRemix.

the biological objective will be to increase the number of *trans*-eQTLs detected from the DataRemix-normalized gene expression profile and the metric y will be the number of *trans*-eQTLs deemed significant. The parameter optimization step which determines the next point to check is detailed in the Methods section.

3.2.2 Parameter Optimization

The parameters $\lambda = (k, p, \mu)$ need to be optimized with respect to a particular biological objective. Grid search and random search [18] are among the most popular strategies, but these methods have low efficiency. Most of the search steps are wasted and the optimality of parameters is highly constrained by the step size and available computing power. In order to utilize the search history and keep a good balance between exploration and exploitation, we can formulate parameter search as a dual learning task.

We define a general performance measure $y = L(\lambda, \mathcal{D})$, with λ representing the parameter tuple (k, p, μ), $\mathcal{D}$ as the data, L as the evaluating process and y as the biological objective. Ideally we can determine the optimal point $\text{argmax}_\lambda L$ easily by gradient descent based method, but usually L is derivative-free and it is also time intensive. Thus we introduce a surrogate model $f(\lambda)$ which can directly predict $L(\lambda, \mathcal{D})$ only given λ, and there are two conditions on f: $\operatorname{argmax}_\lambda f$ should be easy to solve and f should have enough capacity.

With these two properties, we can sequentially update f with (λ_t, y_t) and propose to evaluate L at $\lambda_{t+1} = \operatorname{argmax}_\lambda f$ in the next step. By gradually updating f with newly evaluated samples (λ, y), $\operatorname{argmax}_\lambda f$ approaches the true underlying optimal $\operatorname{argmax}_\lambda L$ as f can gradually fit to the underlying mapping function L. This provides a more efficient approach to explore the parameter space by exploiting the search history. In this work, we model f as a sample from a Gaussian Process with mean 0 and kernel $k(\lambda, \lambda')$, where $\lambda = (k, p, \mu)^T$. It is well known that the form of the kernel has considerable effect on performance. After experimentation we settled on the exponential kernel as the most suited for our application. The exponential kernel is defined as below (note the difference from the squared-exponential or RBF kernel).

$$k(\lambda, \lambda') = \exp\left(-\frac{\|\lambda - \lambda'\|_2}{2}\right) \tag{3.3}$$

We observe $y_t = f(\lambda_t) + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$. For Bayesian optimization, one approach for picking the next point to sample is to utilize acquisition functions [144] which are defined such that high acquisitions correspond to potentially improved performance. An alternative approach is the Thompson Sampling approach [15, 3, 61]. After we update the posterior distribution $P(f|\lambda_{1:t}, y_{1:t})$, we draw one *sample* f from this posterior distribution as the optimization target to infer λ_{t+1}. Theoretically it is guaranteed that λ_t converges to the optimal point gradually [2]. With this theoretical guarantee, we focus on Thompson Sampling approach to optimize parameters for DataRemix.

3.2.2.1 Estimation of Hyperparameters First we rely on the maximum likelihood estimation (MLE) to infer the variance of noise σ^2 [131]. Given the marginal likelihood defined by (3.4), it is easy to use any gradient descent method to determine the optimal σ^2

$$\begin{aligned}\log p(\vec{y}|\vec{\lambda}) = &-\frac{1}{2}\vec{y}^T(K + \sigma^2 I)^{-1}\vec{y} - \frac{1}{2}\log|K + \sigma^2 I| \\ &- \frac{t}{2}\log 2\pi\end{aligned} \tag{3.4}$$

where $\vec{y} = y_{1:t} = (y_1, \ldots, y_t)^T$,$\vec{\lambda} = \lambda_{1:t} = (\lambda_1, \ldots, \lambda_t)^T$ and K is the covariance matrix with each entry $K_{ij} = k(\lambda_i, \lambda_j)$.

3.2.2.2 Sampling from the Posterior Distribution

Since Gaussian Process can be viewed as Bayesian linear regression with infinitely many basis functions $\phi_0(\lambda), \phi_1(\lambda), \ldots$ given a certain kernel [131], in order to construct an analytic formulation for the sample f, first we need to construct a certain set of basis functions $\Phi(\lambda) = (\phi_0(\lambda), \phi_1(\lambda), \ldots)$, which is also defined as feature map of the given kernel. Then we can write the kernel $k(\lambda, \lambda')$ as the inner product $\Phi(\lambda)^T\Phi(\lambda')$.

Mercer's theorem guarantees that we can express the kernels in terms of eigenvalues and eigenfunctions, but unfortunately there is no analytic solution given the exponential kernel we used. Instead we make use of the random Fourier features to construct an approximate feature map [129]. First we compute the Fourier transform p of the kernel (see Sec. B.2 for derivation).

$$\begin{aligned} p(\vec{\omega}) &= \frac{1}{(2\pi)^3} \int \exp(-i\vec{\omega}^T\vec{\Delta}) \exp(-\frac{\|\vec{\Delta}\|_2}{2}) d\vec{\Delta} \\ &= \frac{8}{\pi^2(4\|\vec{\omega}\|_2^2 + 1)^2} \end{aligned} \tag{3.5}$$

where $\vec{\omega} = (\omega_1, \omega_2, \omega_3)^T$ and $\vec{\Delta} = \lambda - \lambda'$. Then we draw m_t iid samples $\omega_1, \ldots, \omega_{m_t} \in \mathbb{R}^3$ by rejection sampling with $p(\omega)$ as the probability distribution. Also we draw m_t iid samples $b_1, \ldots, b_{m_t} \in \mathbb{R}$ from the uniform distribution on $[0, 2\pi]$. Then the feature map is defined by the following equation.

$$\Phi(\lambda) = \sqrt{\frac{2}{m_t}} [\cos(\omega_1^T\lambda + b_1), \ldots, \cos(\omega_{m_t}^T\lambda + b_{m_t})]^T \tag{3.6}$$

where the dimension m_t can be chosen to achieve the desired level of accuracy with respect to the difference between true kernel values $k(\lambda, \lambda')$ and the approximation $\Phi(\lambda)^T\Phi(\lambda')$.

3.2.2.3 Thompson Sampling Any sample f from the Gaussian Process can be defined by $f(\lambda) = \Phi(\lambda)^T\theta$, where $\theta \sim N(0, I)$ and $\Phi(\lambda)^T$ is defined by (3.6). In order to draw a posterior sample f, we just need to draw a random sample θ from the posterior distribution $P(\theta|\vec{\lambda}, \vec{y})$.

$$\begin{aligned} P(\theta|\vec{\lambda}, \vec{y}) &\propto P(\vec{y}|\vec{\lambda}, \theta)P(\theta) \\ &\propto N(A^{-1}\Phi(\vec{\lambda})\vec{y}, \sigma^2 A^{-1}) \end{aligned} \tag{3.7}$$

where $A = \Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + \sigma^2 I$ and $\Phi(\vec{\lambda}) = (\Phi(\lambda_1) \cdots \Phi(\lambda_t))$. (see Sec. B.2 for more details). The overall algorithm is summarized as the following pseudo code.

Algorithm 1 Thompson Sampling for Searching λ

```
Extra Parameters
$t_{max}$: the maximum number of iteration steps
$\xi$: a pre-defined probability which ensures the search doesn't get stuck in a local optimum

1. Get a short sequence $\mathcal{D}_1 = (\lambda, y)$ as seeds by random search.
2. Draw $m_t$ iid samples $\omega_1, \ldots, \omega_{m_t} \in \mathbb{R}^3$ and $m_t$ iid samples $b_1, \ldots, b_{m_t} \in \mathbb{R}$ according to (3.5)
3. Iterate from $t = 1$ until $\lambda$ converges or it reaches $t_{max}$
    (1) At step $t$, estimate the hyperparameter $\sigma^2$ given $\mathcal{D}_t$ according to (3.4)
    (2) Draw a sample $f$ given $\mathcal{D}_t$ according to (3.7) with feature map determined by (3.6)
    (3) $\lambda_{t+1} = \begin{cases} \text{argmax}_\lambda f(\lambda) & \text{w.p. } 1 - \xi \\ \text{random search} & \text{w.p. } \xi \end{cases}$
    (4) Evaluate $y_{t+1}$ given $\lambda_{t+1}$
    (5) $\mathcal{D}_{t+1} = \mathcal{D}_t \bigcup (\lambda_{t+1}, y_{t+1})$
```

3.2.3 Correlation network evaluation

We evaluated the quality of the correlation network derived from a particular dataset using guilt-by-association pathway prediction. Specifically, the genes were ranked by their average Pearson correlations to other genes in the pathway (excluding the gene when the gene itself is a pathway member). The resulting ranking was evaluated for performance using AUC or AUPR metric. For pathway ground-truth, we used the "canonical" pathways dataset from MSigDB, comprising 1,330 pathways [153].

3.2.4 eQTL mapping

eQTL association mapping was quantified with Spearman rank correlation. For *cis*-eQTLs, testing was limited to SNPs which locate within 50kb of any of the gene's transcription start sites (Ensembl, version 90). *cis*-eQTl is deemed significant at 10% FDR with Benjamini-Hochberg correction for the total number of tests. For *trans*-eQTLs, the significance cutoff is 20% FDR with Benjamini-Hochberg correction for the total number of tests. Since the Benjamini-Hochberg FDR is a function of the entire *p*-value distribution in order to ensure consistency comparisons, the rejection level was set once based on the *p*-value that corresponded to 10% or 20% FDR in the original *cis*-optimized $D_{\mathrm{HCP-cis}}$ and *trans*-optimized $D_{\mathrm{HCP-trans}}$ dataset respectively. To reduce the computational cost of grid evaluations, all the optimization computations were performed on a set of 100,000 subsampled SNPs.

3.2.5 Public Data

3.2.5.1 GTEx Dataset We downloaded the complete gene-level TPM data (RNASeQCv1.1.8) from the GTEx consortium [103]. These data were quantile normalized to create the raw dataset. We subsequently subjected the dataset to several different normalization approaches (Table. 6) that account for hidden and known technical factors.

The technical covariates selected were those with the median values of the variance they explained across genes that were above 0.01. The 8 variables that met this threshold were: SMTS (Tissue type, area from which the tissue sample was taken), SMTSD (Tissue type, more specific detail of tissue type), SMUBRID (Uberon ID), SMNABTCHT (Type of nucleic acid isolation batch), SMEXNCRT (Exonic Rate: the fraction of reads that map within exons), SMGNSDTC (Genes detected), SMTRSCPT (Transcripts detected) and SMNTRNRT (Intronic Rate: the fraction of reads that map within introns).

DataSet	Description
Remove PC	We keep removing first several (up to 300) principal components (PCs) until the network quality metrics (mean AUC and mean AUPR) no longer improve.
Remove tech	We remove the technical covariates by ridge regression with cross validation.
Remove tech + PC	We remove the technical covariates as above and subsequently remove residual PCs until the network performance metrics no longer improve.
DataRemix	DataRemix normalization is performed with k ranging from 1 to 100. $p \in [-1, 1]$ and $\mu \in [0, 1]$
HCP	HCP normalization is performed with following parameter settings. $k \in [1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$, $\lambda \in [1, 5, 10, 20]$, $\sigma_1 \in [1, 5, 10, 20]$ and $\sigma_2 \in [10, 20]$. We run grid search to pick up the best combination of parameters.

Table 6: Different normalizations of the GTEx dataset.

3.2.5.2 DGN Dataset Depression Gene Networks (DGN) dataset contains whole-blood RNA-seq and genotype data from 922 individuals. The genotype data was filtered for MAF>0.05. The genomic coordinate of each SNP was taken from the Ensembl Variation database (version 90, hg19/GRCh37). SNP identifiers that were not present in that release were excluded. After filtering, there were 649,875 autosomal single nucleotide polymorphisms (SNPs). Data is available upon application through NIMH Center for Collaborative Genomic Studies on Mental Disorders. For gene expression we used the gene-level quantified dataset. The dataset came already filtered for expressed genes and was further filtered for gene symbols that were not present in Ensembl 90 leaving 13,708 genes. The dataset comes in two covariate normalized versions with normalization parameters optimized for *cis*- and *trans*-eQTL discovery separately. To create the naive-normalized dataset, we applied a log transformation, $log(x + 1)$, to the raw counts and quantile normalized the results.

3.2.5.3 ROSMAP dataset The raw data was obtained from Synpase (syn3219045). The data was optimized for the network quality objective using the canonical pathway genesets from MSigDB [153]. The data was corrected for sex, age and 10 genotype principal components. In order to quantify exon-level effects we used the Synapse BAM files to quantify exon-level FPKMs using featureCounts [99].

3.2.5.4 NESDA The NESDA (Netherlands Study of Depression and Anxiety) dataset was obtained from dbGAP (phs000486.v1). Following suggestions from study authors, the NESDA dataset was normalized for sex,age, and the first 10 genotype PCs using linear regression. Genotypes were imputed using Michigan Imputation Server [35] using 1000 Genome Phase 3 (Version 5) as the reference panel. We assesed the replication of DGN eQTLs based on exact gene and SNP matches.

3.3 Results

3.3.1 Simulation Study

In order to evaluate the performance of DataRemix when different variance components align with the true biological signals, we performed a simulation study focusing on three representative cases. The cases are: 1) only high-variance components encode biological signals (high-variance Fig. 28), 2) only low-variance components encode biological signals (low-variance) and 3) both high- and low-variance components correspond to useful variations (general case). We simulated gene expression profile along with ground-truth pathways and evaluated whether DataRemix could improve the recovery of the simulated pathways (AUC and AUPR) using guilt-by-association.

We simulated gene expression profile with 5000 genes, 300 samples and 50 latent factors based on the following linear model.

$$X = WH + E$$

We set W and H to be positive. Each column of W and each row of H was drawn from a Normal distribution with mean equal to zero, and the variance parameters were drawn from Exponential distribution with 1e-3 as rate. In this way, the singular values can decrease gradually as the rank increases and each latent factor can have a non-negligible effect when recovering simulated pathways. The matrix $E \in \mathcal{N}(0, 2)$ represents random noise.

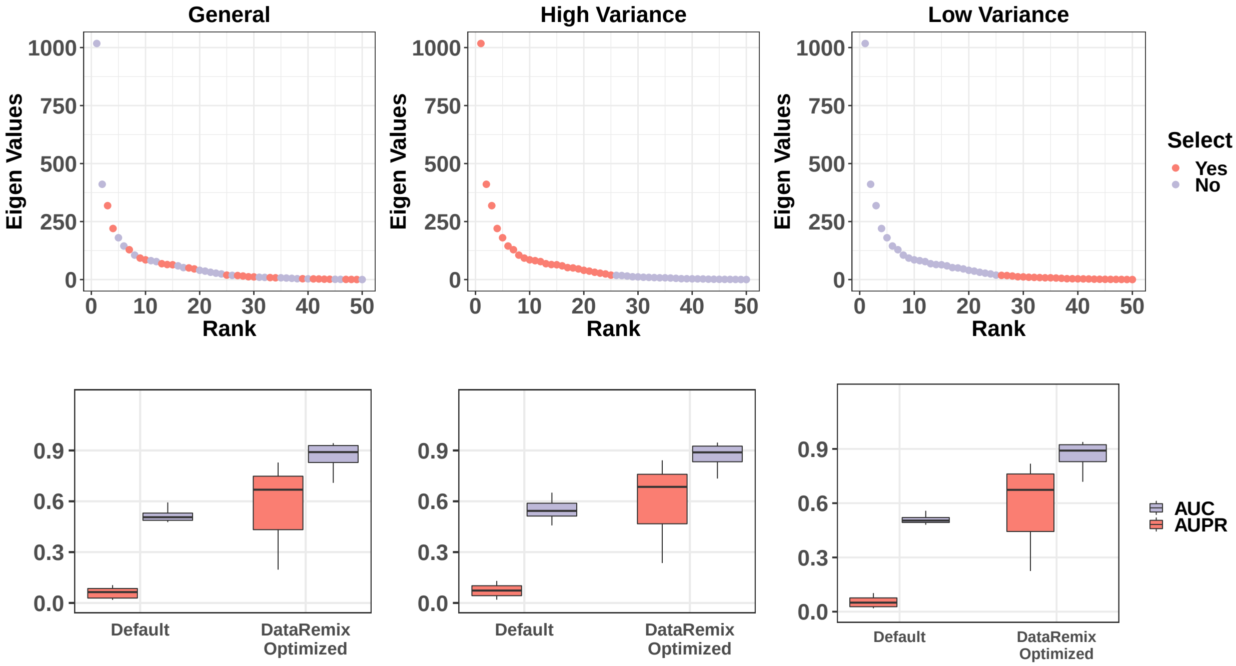


Figure 28: We simulate gene expression data with a low rank approximation so that the component variance distribution approximates that which is typically seen in gene expression data (top row). According to our assumptions only some of the low rank components represent useful biological variation. The left, middle and right panel depict the general, high-variance and low-variance case with the pink points denoting the factors with biological variations. These factors are used to construct the ground-truth pathway membership matrix. In the second row, we compare the AUC and AUPR for recovering the pathway co-membership via guilt-by-association analysis on the correlation network. DataRemix is able to improve this metric by reweighing the contribution of different variance components.

The gene expression profile is consistent across three cases and a different pathway matrix is generated separately according to each assumption. In the high-variance case, we select the top 25 latent factors. In the low-variance case we pick up the last 25 latent factors and randomly sample 25 latent factors for the general case. Then for corresponding columns in W, we randomly select a threshold between 0.01 and 0.1 with 0.01 as the step size. With the threshold value, we pick up the corresponding highest quantile of genes to construct the pseudo geneset as ground truth. The simulated data is used to construct a gene-correlation network which is evaluated according to guilt-by-association recovery of the ground-truth pathways, a commonly accepted network quality metric. We evaluate both the raw data and the optimized Remixed result. In all 3 cases DataRemix was able to substantially improve network quality metrics.

3.3.2 Quality of the correlation network derived from the GTEx gene expression study

The GTEx datasets [103] is comprised of human samples from diverse tissues, many of which were obtained post-mortem and there are many technical factors which have considerable effects on the gene expression measurements. On the other hand this rich dataset provides an unprecedented multi-tissue map of gene regulatory networks and has been extensively analyzed in this context. It is natural to assume that a dataset that is better at recovering known pathways is likely to yield more credible novel predictions. Thus, we use DataRemix to optimize the known pathway recovery task as a function of the correlation network computed on a Remixed dataset.

We formally define the objective as the average AUC across "canonical" mSigDB pathways (which include KEGG, Reactome and PID) [153] using guilt-by-association. Specifically, the genes are ranked by their average Pearson correlations to other genes in the pathway (excluding the gene when the gene itself is a pathway member). Fig. 29A depicts the results of grid search for the parameters k and p (with μ fixed at 0.01) and the contour plot shows a clear region of increased performance. Using the optimal transformation found by grid search, we plot per-pathway AUC improvement in Fig. 29B and find that the AUC is substantially increased for almost every pathway.

In Fig. 30 we systematically evaluate the performance of DataRemix against alternative

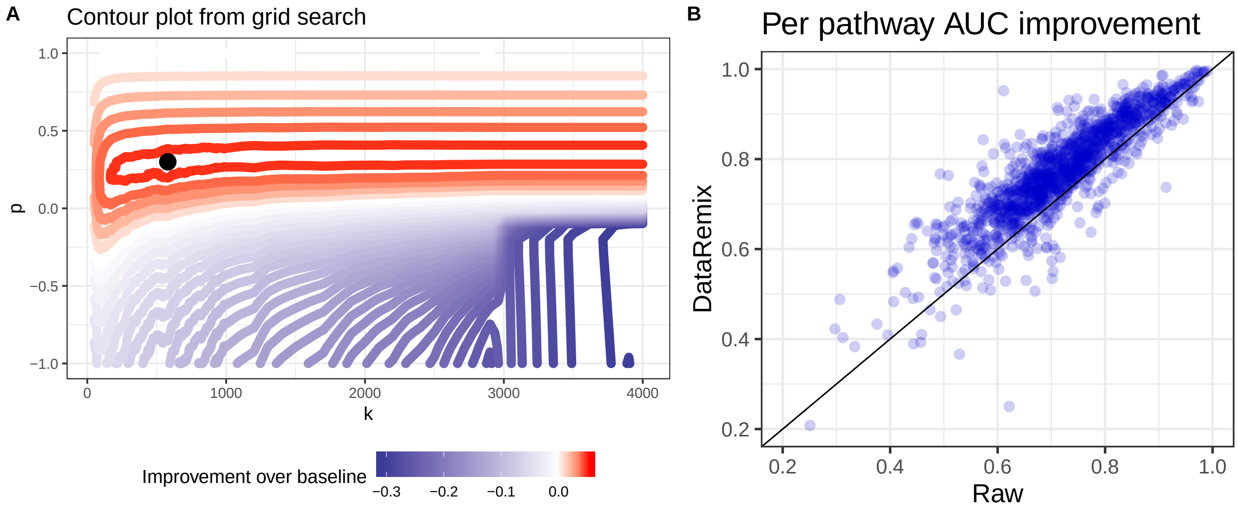


Figure 29: **A.** The improvement in performance of DataRemix transform of the pathway prediction task visualized as a function of k and p parameters (μ is fixed at 0.01). Performance is measured as the mean AUC across all pathways in the "canonical" mSigDB dataset and the red contours indicate improvement over the performance on untransformed data. **B.** Per-pathway performance improvement for the DataRemix transformation corresponding to the optimal point in **A**.

methods. For the purpose of evaluation we include the naive method of simply removing known and hidden factors from the data. We consider removing principal components (Remove PC), removing known technical variables (Remove tech), and a combination of the two (Remove tech and PC). Since the number of hidden factors is not known, we optimize the number of PCs removed to the specific network quality objective (see Methods for further details). We also include a penalized mixed linear model method "Hidden Covariates with Prior" (HCP) which takes known covariates as input. In addition to the number of hidden components, this method has 3 hyperparameters that were optimized to maximize the network quality objective via grid search. HCP has been extensively benchmarked perviously and has been shown to outperform both naive methods and the widely used PEER approach (see [148] for PEER and [116] for HCP including performance comparison). Moreover, HCP is considerably faster than PEER making an extensive hyperparameter search feasible.

We find that on this dataset DataRemix is able to outperform all naive methods including ones that make use of known technical covariates, achieving performance that is comparable to that of HCP. In summary, our DataRemix framework is able to match the performance of

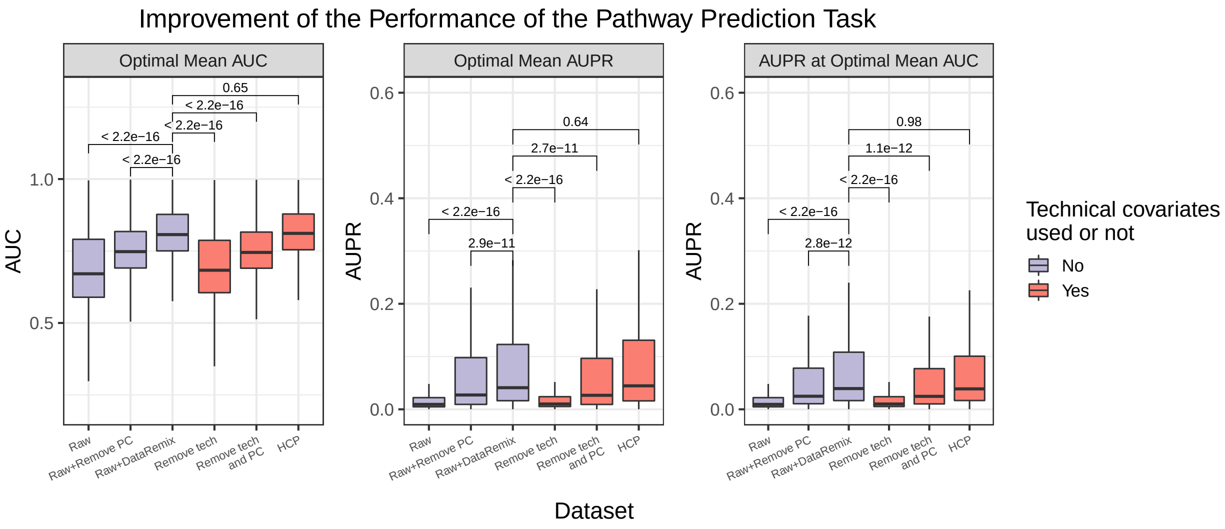


Figure 30: We compare our DataRemix approach to other common normalization strategies with respect to correlation network quality. Here, we consider different normalizations of the GTEx dataset and the details are described in Table. 6. We compute several "naive" normalizations which simply remove known factors (tech), top k principal components where k is optimized for the task (PC) or both (tech+PC). We also consider "Hidden Covariates with Prior" (HCP) which is a mixed linear model that takes known factors into account and has been shown to outperform other methods in various normalization tasks [116] . The four hyperparameters in HCP are optimized by grid search. Each box plot shows the distribution in AUCs or AUPRs across the "canonical" mSigDB pathways. P-values compare the results achieved by DataRemix against others using the Wilcoxon ranksum test. DataRemix's performance surpasses all naive methods and is comparable to HCP while using no technical covariates and considerably less computation (see text for details) .

the best competing method, HCP, *while using no technical covariates.* It is worth pointing out that once a truncated SVD decomposition is computed, a single DataRemix evaluation requires only two matrix multiplications while HCP is an optimization problem which needs to be solved iteratively with two matrix inversions at each step.

3.3.3 eQTL discovery in the DGN dataset.

We also consider the task of discovering *cis*- and *trans*-eQTLs on the Depression Gene Networks (DGN) dataset [16]. In the original analysis this dataset was normalized using the Hidden Covariates with Prior (HCP) [116] with four free parameters that were separately optimized for *cis*- and *trans*-eQTLs. The rationale behind separate *cis* and *trans* optimized normalization can be understood in terms of which variance components represent true biological vs. nuisance variation in the two contexts. Specifically, *cis*-eQTLs represent *direct* effects of genetic variation on the expression of a single gene. On the other hand, *trans*-eQTLs represent network level, *indirect* effects that are mediated by a regulator. Thus, *trans*-eQTLs are reflected in systematic variation in the data which becomes a nuisance factor when only direct effects are of interest. It thus follows that the data should be more aggressively normalized for *cis*-eQTL discovery. The original analysis of this dataset optimized the HCP parameters separately for the *cis* and *trans* tasks yielding two different datasets that we refer to as $D_{\text{HCP}-\text{cis}}$ and $D_{\text{HCP}-\text{trans}}$.

The HCP model takes various technical covariates as input, and 20 of the covariates used in the original study cannot be inferred from the gene-level counts. In order to investigate how much improvement can be achieved via DataRemix in the absence of access to these covariates, we also consider a "naively" normalized dataset, quantile normalization of log-transformed counts, or D_{QN}.

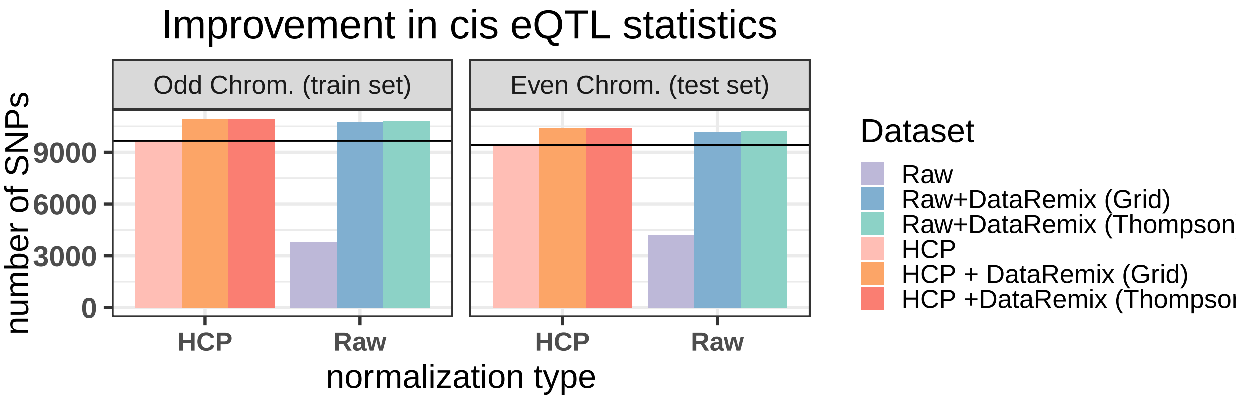


Figure 31: Final results from DataRemix parameter search using a cross-validation framework. *cis*-eQTL statistic is defined to be number of SNP-gene interaction deemed significant at 10% FDR (Benjamini-Hochberg correction for the total number of tests), where the SNP is located within 50kb of the gene's transcription start site. Optimal parameters are determined using the odd chromosome SNPs only and then tested on the even chromosome SNPs. While the raw dataset is considerably worse than HCP, both are improved to a similar level with DataRemix. We find that the DataRemix transform does not overfit the objective as the degree of improvement is similar across the test and train SNP sets. (Note, the starting value of the raw or HCP dataset differ between the test and train SNP set). Moreover, we find that Thompson Sampling is able to match grid search results using only 100 evaluations.

3.3.3.1 *cis*-eQTLs In this task we focus on optimizing the discovery of *cis*-eQTLs. We define *cis*-eQTLs as a SNP-gene interaction where the SNP is located within 50kb of the gene's transcription start site. The interaction is quantified with Spearman rank correlation and deemed significant at 10% FDR (Benjamini-Hochberg correction for the total number of tests).

We perform our analysis in a cross-validation framework, whereby we optimize DataRemix parameters (using grid search or Thompson Sampling) using SNPs on the odd chromosomes and then evaluate the parameters on the, held-out, even chromosome set. Since there are no hyperparameters to optimize the even chromosome validation is performed exactly once.

The final results for both the train and test set are depicted in Fig. 31. As expected, the quantile-normalized dataset D_{QN} performs considerably worse than $D_{\mathrm{HCP-cis}}$, which is specifically optimized for *cis*-eQTL detection. However, the two datasets achieve comparable performance after applying DataRemix. Moreover, the final performance of the Remixed D_{QN} dataset is an improvement on $D_{\mathrm{HCP-cis}}$ demonstrating the near optimal normalization is possible without access to technical covariates. Importantly, we find that the optimal parameters are indeed generalizable as we achieve a similar level of improvement on the train and test chromosomes.

3.3.3.2 *trans*-eQTLs In our second task, we optimize the discovery of *trans*-eQTLs in the same DGN dataset. Ideally, *trans*-eQTLs represent network-level effects and thus give some insight about the regulatory structure of gene expression. However, in practice *trans*-eQTLs are simply defined as SNP-gene associations where the SNP and the gene are located on different chromosomes. While this is a useful heuristic definition, it doesn't guarantee that the association is mediated at the network level. One possible source of bias is mis-mapped RNAseq reads which contaminate the quantification of the apparently *trans*-associated gene with reads from a homologous locus that has *cis* association. Even in the absence of technical artifacts, direct interchromsomal interactions have been observed (see [164] for a comprehensive review). In order to focus on potential indirect effects, we apply an additional filter to *trans*-eQTL discovery. Specifically we require SNPs involved in a *trans* effect to be associated with more than one gene at a FDR of 20% (Benjamini-

Hochberg correction for the total number of tests (approximately 8 × 10^9). We term these SNPs *trans*-SNPs$^+$. In comparison with same chromosome *cis*-eQTLs, inter-chromosome *trans*-eQTLs are rare and *trans*-SNPs$^+$ (as defined above) are more rare still. In fact, using the odd chromosome SNPs subsampled at 20%, we find only 88 such SNPs using $D_{\text{HCP}-\text{trans}}$ dataset and this is the default value we wish to improve.

Here again we find that the dataset specifically optimized for the task of *trans*-eQTL detection, $D_{\text{HCP}-\text{trans}}$, considerably outperforms the raw data D_{QN}, however DataRemix is able to improve both to a similar performance. As is the case with the *cis*-eQTL objective, the cross-validation procedure gives consistent results and no overfitting is observed for either grid search or Thompson Sampling (Fig. 32). We note that Thompson Sampling is able to achieve a better performance than grid search, though the improvement is small in absolute magnitude due to the scarcity of *trans*-eQTLs. In this case, the optimal region for the DataRemix transformation is relatively small (Fig. 33) and thus Thompson Sampling has an advantage since it can search off the grid.

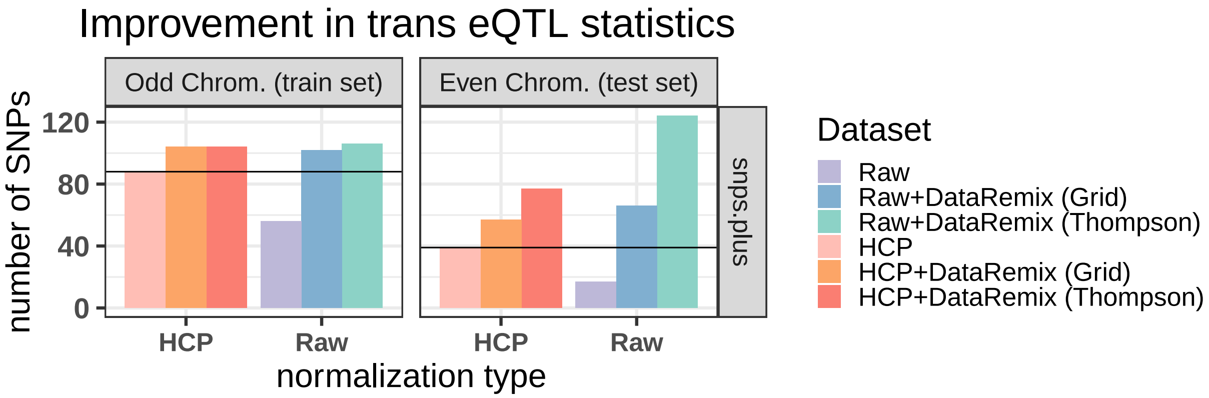


Figure 32: Final values for the eQTL statistics obtained from two versions of datasets. *trans*-eQTL statistic is defined to be number of SNPs involved in a *trans* effect and associated with more than on gene at a FDR of 20% (Benjamini-Hochberg correction for the total number of tests). Here we make a comparison between quantile normalized D_{QN} and HCP normalized $D_{\text{HCP}-\text{trans}}$ with parameters optimized for *trans*-eQTL discovery. We find DataRemix is able to improve upon either of starting datasets and the improvements on both the train and test dataset are comparable which indicates that overfitting is not a problem

3.3.3.3 DataRemix performance transfers across different network objectives

It is well know that for statistical analyses of genomic datasets, more significant associa-

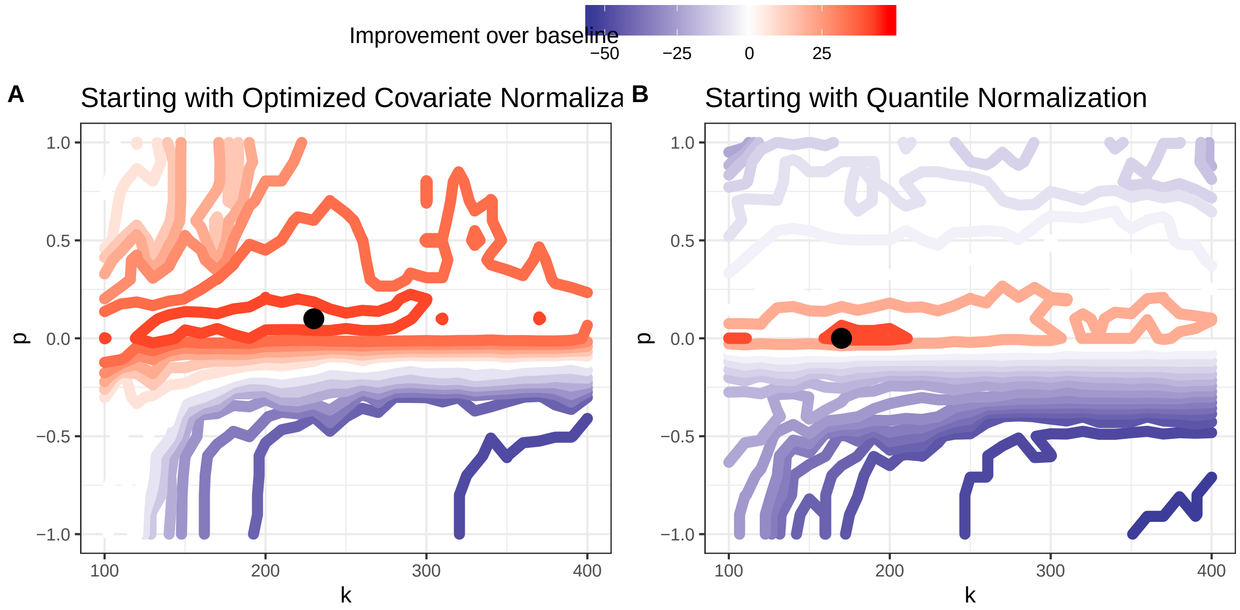


Figure 33: Contour plot representing the effects of the k and p parameters on the performance of DataRemix regarding *trans*-eQTLs discovery on training set. The μ parameter is fixed at 0.01. Red contours represent parameter combinations that increase the number of *trans*-eQTLs beyond what can be achieved using the $D_{HCP-trans}$ dataset. Panel **A** shows the results starting with $D_{HCP-trans}$ while D_{QN} is used for panel B. Improvement can be achieved starting with either dataset. We find that the region of improved performance is smaller than that for *cis*-eQTLs and is particularly concentrated when starting with the D_{QN} (panel **B**) dataset.

tions do not necessarily mean improved biological findings. However, it is generally agreed that improvement in *cis*-eQTL detection cannot be achieved through artificial means but indeed represents improved correction for confounding factors [148, 116]. There is no such consensus for *trans*-eQTLs which are rare, and subject to many artifacts. Consequently, it is important to further corroborate the biological validity of the *trans*-optimized dataset through independent means.

Since *trans*-eQTLs are likely to reflect pathway-level effects, we expect that a dataset that is optimally transformed for *trans*-eQTL discovery should also produce better correlation networks. We thus investigate if optimal DataRemix transform is transferable across these tasks by verifying whether the Remixed dataset optimized with respect to *trans*-eQTL discovery also improves the network quality criterion. Similar to our analysis of the GTEx datasets, we use the correlation network to perform guilt-by-association pathway predictions

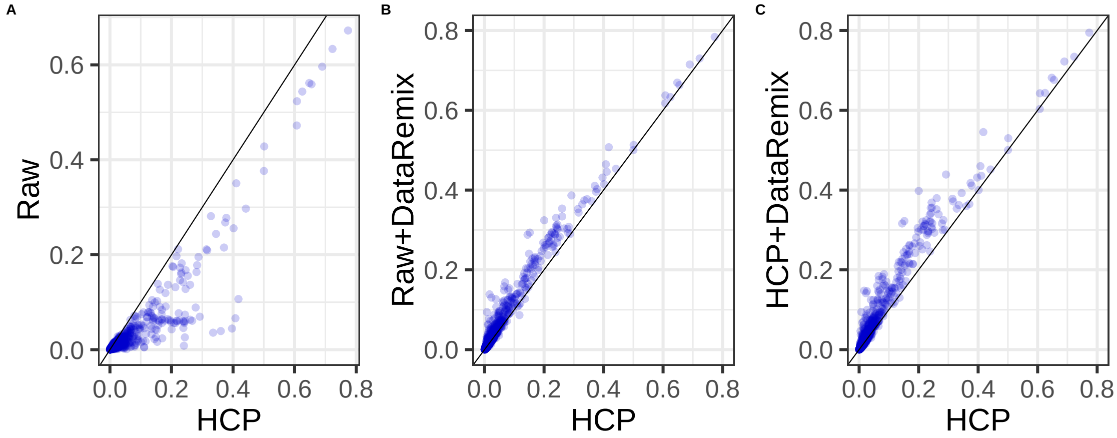


Figure 34: DataRemix-transformed datasets improve the pathway prediction objective which is not explicitly optimized. Each plot is a per-pathway AUPR (area under precision-recall curve) from various datasets (y-axis) contrasted with the results from the optimal covariate-normalized dataset $D_{\mathrm{HCP-trans}}$, which serves as the baseline (x-axis). Panel **A** shows the contrast between $D_{\mathrm{HCP-trans}}$ and D_{QN}. The performance of $D_{\mathrm{HCP-trans}}$ is considerably better. Panel **B** shows the results of the Remixed D_{QN} datasets (optimized for *trans*-eQTL discovery with Thompson Sampling). Even though D_{QN} starts out as considerably worse, the Remixed version is able to outperform $D_{\mathrm{HCP-trans}}$. Panel **C** shows the results of Remixed $D_{\mathrm{HCP-trans}}$. We choose to show AUPR instead of AUC because we find that Remixed version matches but doesn't further improve the AUC performance of $D_{\mathrm{HCP-trans}}$

and evaluate the results over 1,330 MSigDB canonical pathways. Fig. 34 shows scatter plots of per-pathway AUPR (area under precision-recall curve) for several comparisons with respect to the baseline $D_{\mathrm{HCP-trans}}$ dataset. In the first panel we contrast the performance to D_{QN} and observe that, as expected, $D_{\mathrm{HCP-trans}}$ brings a considerable improvement over the quantile normalized dataset. In the second panel we contrast $D_{\mathrm{HCP-trans}}$ with the Remixed version of D_{QN} (optimized for *trans*-eQTL discovery with Thompson Sampling). We find that the pattern becomes opposite and the Remixed D_{QN} dataset performs consistently better that $D_{\mathrm{HCP-trans}}$. The final panel shows the results of Remixing $D_{\mathrm{HCP-trans}}$ itself which also improves the performance. Overall, we find that DataRemix improves multiple criteria of biological validity as optimizing for the *trans*-eQTL objective also results in improved correlation networks.

A major finding of our study is that for the eQTL and pathway prediction tasks, the starting point of normalizing DGN datasets appears to matter relatively little. Even though the quantile-normalized dataset performs considerably worse in the beginning, after Remixing its performance matches that of the optimal covariate-normalized datasets. Of course, if covariates are available, it is preferable to use them and in the case of DGN, slightly further improvement can be achieved. However, our results indicate that in some cases datasets *can* be effectively normalized even in the absence of meta-data about quality control or batch variables. This is an important consideration for many legacy datasets where such information is not available.

3.3.4 Novel Biological Findings

3.3.4.1 New *trans*-eQTLs effects in the DGN dataset At the optimal DateRemix parameters for D_{QN}, we find 3,000 gene-SNP trans associations at a Benjamini-Hochberg FDR of 0.2 where in contrast to 1,691 for $D_{HCP-trans}$. We verified the replication of these associations in an independent dataset, NESDA and find that 1,013 (33%) of the DataRemix associations had a replication FDR of < 0.2 while for the default $D_{HCP-trans}$ dataset the same number was 707 (41%). The replication rate was somewhat smaller on the Remixed dataset, which is expected as the replication was performed on raw NESDA data. However, the *total* number of replicated effects was greater.

We highlight an example of new regulatory module recovered via DataRemix that appears to be biologically credible based on independent replication and the known functions of the genes involved. We find that SNP rs11145917 located near CARD9 gene is associated with three genes in the alpha interferon response (Table. 7). The locus has been associated with Crohn's disease [44] and Ulcerative colitis [8] though to our knowledge no mechanism has been proposed. We find that rs11145917 has a *cis* effect on CARD9 and the *trans* effects are partially mediated by CARD9 expression. In summary, our analysis suggests that CARD9 may affect baseline activity of the alpha interferon pathway, which is a testable prediction with potential clinical importance.

3.3.4.2 Analysis of the Religious Orders Study and Memory and Aging Project (ROSMAP) Study We sought to apply our method to the Religious Orders Study and Memory and Aging Project (ROSMAP) Study dataset which consists of 370 human samples

SNP	Gene	Method	Spearman rho	p-value	FDR(B.H.)
rs11145917	SIGLEC1	DataRemix	-0.1782	5.1052E-08	0.0889
		Raw	-0.1510	4.1326E-06	>0.2
		NESDA replication	-0.0414	8.1499E-02	0.2148
	IEIT1	DataRemix	-0.1783	5.0403E-08	0.0881
		Raw	-0.1627	6.7749E-07	>0.2
		NESDA	-0.07919	8.6260E-04	0.0050
	ISG15	DataRemix	-0.1830	2.1867E-08	0.0451
		Raw	-0.1541	2.5755E-06	>0.2
		NESDA replication	-0.07451	1.7229E-03	0.0088

Table 7: The association of rs11145917 with genes in the alpha interferon pathway is replicated in an independent dataset. We note that the FDRs for the NESDA dataset represent a correction for the total number of replication test performed, that is only gene-SNP pairs which passed a FDR $<$ 0.2 in the DGN dataset. Since the fraction of true positives in the replication scenario is higher, the FDRs are lower than the genome-wide FDRs at the same p-value.

with paired gene expression and genotype information. To our knowledge no *trans*-eQTLs have been reported for human brain and indeed we could not detect any genome-wide significant *trans* effects in the ROSMAP dataset. Since no *trans*-eQTLs can be detected, there is no variance in this objective and thus our method cannot be applied directly. However, using the DGN dataset we have shown that optimizing for *trans*-eQTLs discovery also optimizes the network quality objective demonstrating that these two objectives are related. Thus, for the ROSMAP dataset we can optimize network quality (which is quantitative and thus always has some variance across different DataRemix parameter combinations) and hope to implicitly optimize *trans*-eQTLs discovery. Fig. 35A shows the change in mean AUC and mean AUPR for the network objective after applying DataRemix (see Methods for details). We find that while the mean AUC changes modestly the mean AUPR is nearly doubled. Applying *trans*-eQTLs analysis to the Remixed ROSMAP dataset we detect a single significant effect between CYP2C8 (chr10) and rs10821352 (chr9). This effect was replicated in the CommonMind Consortium dataset [46] with a p-value of 3.1382e-16 (Spearman rank correlation). The interaction passed all quality checks. Specifically, all CYP2C8 30-mers mapped back to CYP2C8 indicating that artifacts from mismapped reads were unlikely and furthermore the eQTL effect was consistent across all 8 exons (Fig. 36). To our knowledge this is the first replicated *trans*-eQTLs reported in human brain data.

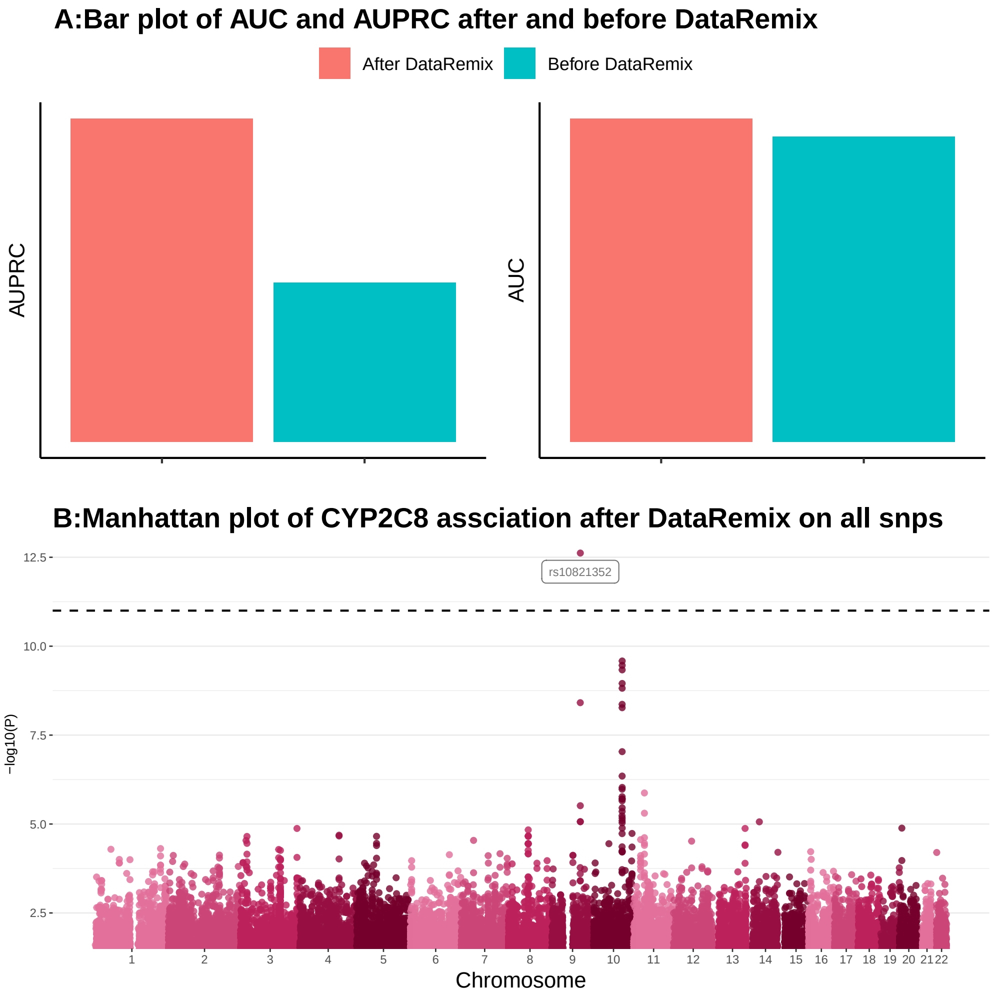


Figure 35: **A.** Improvement in the network quality objective after running DataRemix with Thompson sampling. **B.** Manhattan plot of associations with CYP2C8 expression. The CYP2C8 gene is located on chromosome 10. A single SNP on chromosome 9 shows a strong *trans* effect with a *p*-value that is notably smaller than the group of *cis*-effect SNPs on chromosome 10.

The gene, CYP2C8, is a member of the cytochrome P450 and is thought to be involve in the metabolism of polyunsaturated fatty acid and lipophilic xeonbiotics. The xenobiotic metabolism function is supported by the correlation network around CYP2C8. Among its top neighbors are GSTA4 (rank 1, Spearman ρ =0.68), CES4A (rank 4, Spearman ρ =0.66)

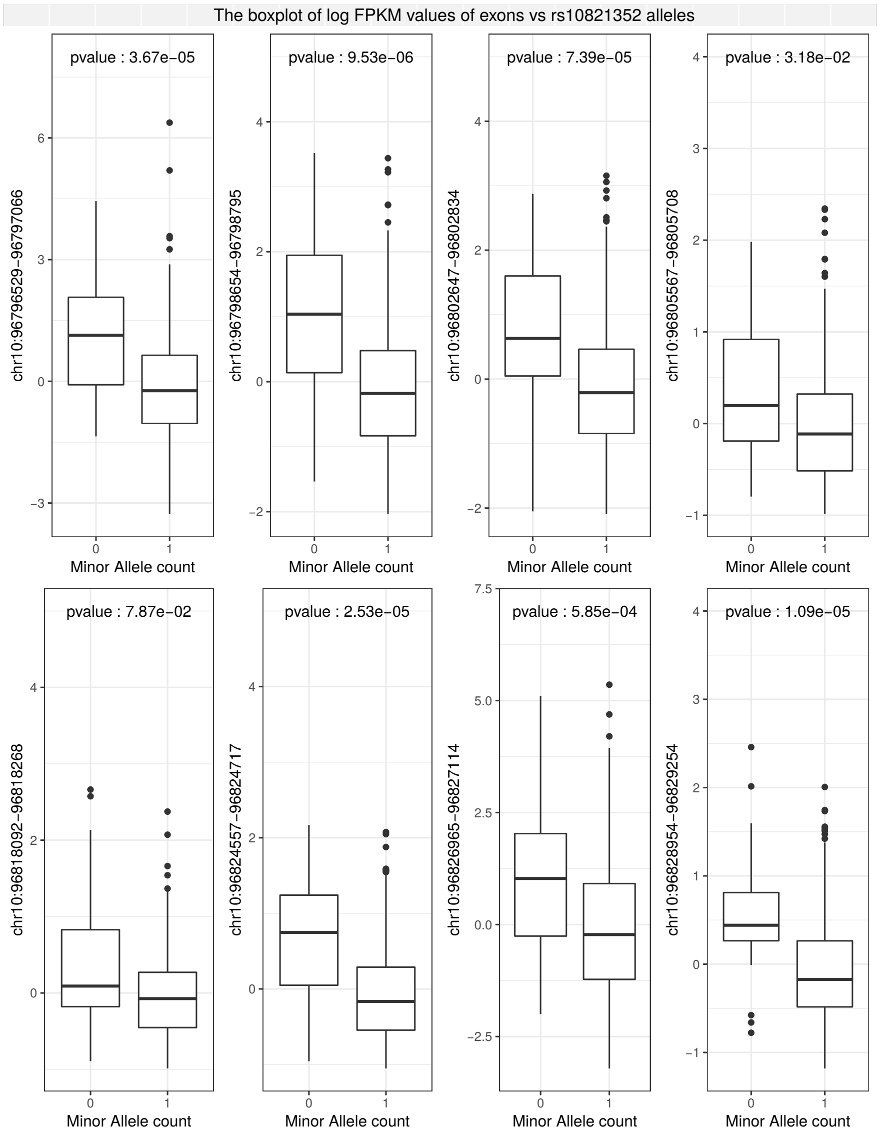


Figure 36: The effect of rs10821352 on CYP2C8 expression is consistent across exons. We used raw FPKM data to quantify the exon-level *trans*-eQTL effects. The effect is consistent across exons further confirming that it is unlikely to be due to homolog mismapping and other technical artifacts.

–two other genes implicated in xenobiotic metabolism. The precise mechanistic nature of how genotype in the rs10821352 locus affects CYP2C8 expression is unclear. No *cis*-eQTLs for rs10821352 could be detected in ROSMAP and none are reported in the GTEx brain data.

3.3.5 Thompson Sampling Performance

We find that Thompson Sampling matches the best grid-search performance within 100 steps giving a 40-fold reduction in the number of evaluations (Fig. 37). We also note that it is possible for the Thompson sampling to surpass the grid-search results since the parameter combinations are not constrained by the choice of grid.

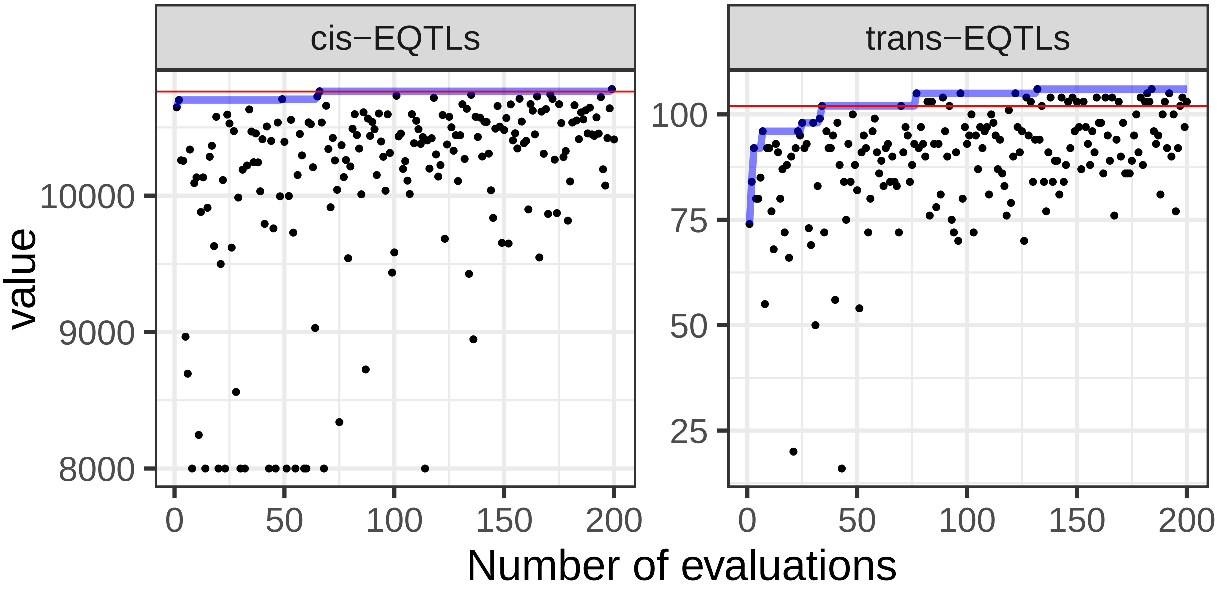


Figure 37: Objective evaluations as a function of iteration number for the *trans*-eQTL and *cis*-eQTL objectives using the quantile normalized D_{QN} dataset. Red lines indicate the maximum value that was obtained by grid search and blue lines indicate the cumulative maximum of Thompson Sampling.

3.4 Discussion

We have proposed DataRemix, a new optimizable transformation for gene expression data. The transformation is able to improve the biological validity of gene expression representations and can be used for effective normalization in the absence of any knowledge of technical covariates. One limitation of the DataRemix approach is that it works best on data that is well approximated by a single Gaussian. However, it is relatively straightforward to adapt the approach to matrix decompositions different from SVD that are more suitable for non-Gaussian data, such as ICA (Independent Component Analysis). We also note that it is possible to introduce additional parameters that specify more complex weighting schemes. However, as the number of parameters is increased, there is a potential for over-optimization of a specific objective above others. We emphasize that in our simple parametrization, we observe that multiple metrics of biological validity improve when only one is explicitly optimized. Specifically we find that optimizing for *trans*-eQTL discovery also improves the correlation network as measured by guilt-by-association pathway prediction. This property is less likely to be preserved as the number of parameters is increased.

4.0 NIFA - Non-negative Independent Factor Analysis for single cell RNA-seq

4.1 Introduction

Single-cell RNA sequencing (scRNA-seq) techniques have allowed researchers to query the complexity of transcription regulation at an unprecedented level of detail. scRNA-seq technologies have the power to reveal both distinct cell types and transcriptional heterogeneity within a defined cell population. However, as individual transcript measurements are noisy and often difficult to interpret in isolation, scRNA-seq analysis methods rely heavily on multivariate techniques.

As the number and size of single-cell datasets increases, it becomes important to develop methods that can quickly summarize the *biological* information embedded in a scRNA-seq dataset as a set of interpretable variables which can be used for downstream analysis. One kind of summary measures is the identity and number of cell types present in a datasets. In recent years there has been a proliferation of clustering methods designed to address this problem [73, 26]. Clustering approaches assume that the data is well described by a discrete set of cell types, but in many cases, questions about continuous biological variation, such as developmental trajectories or levels of pathway activation are also of interest.

Such continuous variables do not conform to the assumptions of clustering algorithms but can be effectively modeled as latent factors. For example, cell-cycle variation has been repeatedly discovered in single-cell data, both using sophisticated latent variable models [22] and simple Principal Component Analysis (PCA) [77].

Of course, cell-type identity can also be thought of as a latent factor and this observation underlies the popularity of Independent Component Analysis (ICA) in single-cell pipelines. Unlike PCA which seeks directions that maximize variance, ICA finds maximally independent or maximally non-Gaussian directions [66].This property is well suited for the analysis of single-cell datasets as directions that maximally separate cell types are multi-modal and thus highly non-Gaussian. For this reason ICA is used as a dimensionality reduction preprocessing step [26]. However, the ICA formulation is not a proper likelihood framework as it has no reconstruction error. A side effect of this is that it requires a loading orthogonalization step to prevent latent variables from collapsing. This rigid formulation restricts the interpretability of individual components a criticism is also valid for PCA/SVD. For the case of PCA/SVD, there are a number of alternative factor analysis methods that produce more interpretable components by relaxing orthogonality and introducing additional constraints, for example, NMF [84] and SPC [165]. It is natural to ask if analogous approaches can be applied to find interpretable multi-modal factors.

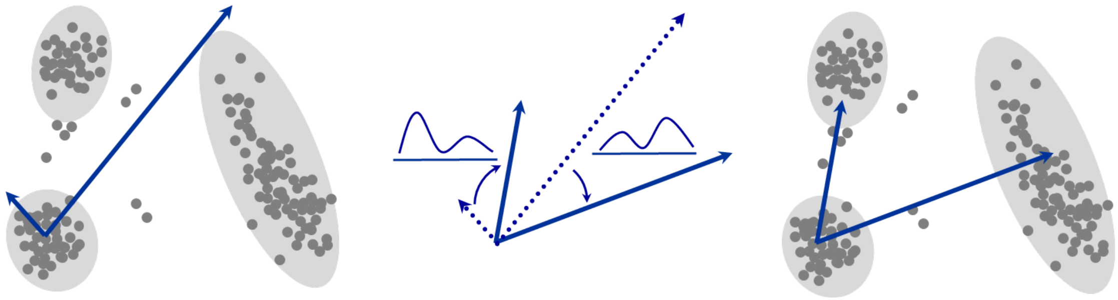


Figure 38: Illustration of the NIFA. Left: We assume there are three hypothetical cell clusters and there are two latent components which point to arbitrary directions. Middle+Right: By imposing multi-modal prior, we force the latent factors to rotate and align with the directions that can best separate the cell-type identity.

We propose **N**on-negative **I**ndependent **F**actor **A**nalysis (NIFA) that combines properties of ICA, PCA and NMF. As illustrated in Fig. 38, our approach simultaneously models uni- and multi-modal factors thus isolating discrete cell-type identity and continuous pathway-level variations into separate components. Furthermore, our model constrains the factor loading to be non-negative providing greater biological interpretability.

4.2 Methods and Materials

4.2.1 The statistical model

X represents a scRNA-seq matrix with dimension P-by-N, where P is the number of genes and N is the number of cells (Fig. 39 and 40). Given X, we want to infer A which denotes loading matrix with dimension P-by-K and S which stands for sources or latent variables with dimension K-by-N. We denote the n_{th} column of X as $X_n = (X_{1n}, X_{2n}, \ldots, X_{Pn})^T$, the j_{th} row of A as $A_j^T = (A_{j1}, A_{j2}, \ldots, A_{jK})$ and the n_{th} column of S as $S_n = (S_{1n}, S_{2n}, \ldots, S_{Kn})^T$ (see Fig. 39). We assume the noise model to be Gaussian with a single precision parameter β.

$$X_n | A, S_n, \beta \sim N(0, \Sigma), \Sigma^{-1} = \text{diag}(\beta)_{P\times P} \tag{4.1}$$

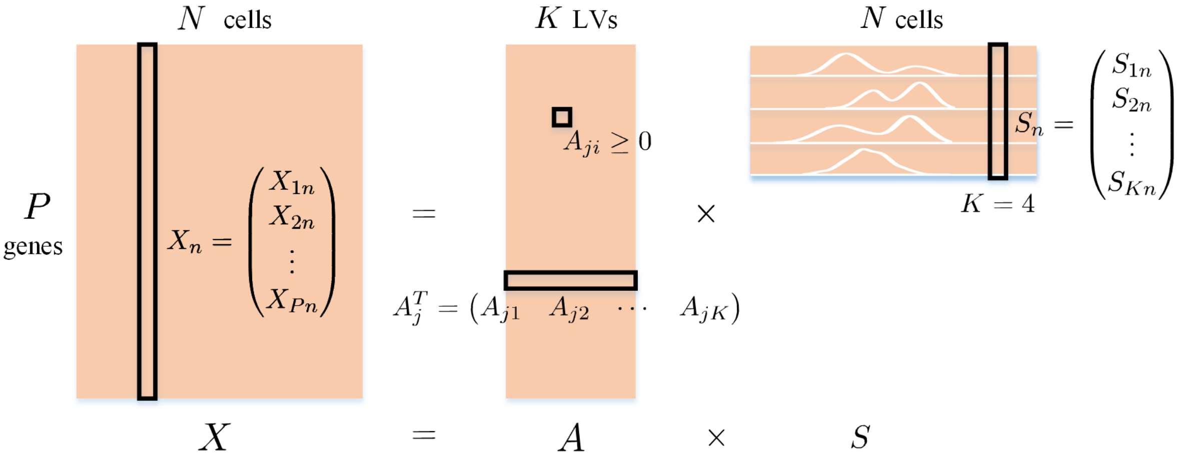

N cells
K LVs
N cells
P
genes
$X_n = \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{Pn} \end{pmatrix}$
$=$
$A_{ji} \geq 0$
$A_j^T = (A_{j1} \quad A_{j2} \quad \cdots \quad A_{jK})$
$\times$
$S_n = \begin{pmatrix} S_{1n} \\ S_{2n} \\ \vdots \\ S_{Kn} \end{pmatrix}$
$K = 4$
X
$=$
A
$\times$
S

Figure 39: A schematic representation of the NIFA model. The gene × cells matrix X is decomposed as a non-negative loading matrix A and a factor matrix S. We impose multi-modal priors on the rows of S, but the exact number of modes is automatically determined and thus can be one.

4.2.1.1 The prior distribution Each latent variable is associated with M component distributions which we assume follows a Gaussian distribution with μ_{im} as the mean and σ_{im}

as the inverse of the variance. ϵ_{imn} is a set of binary latent variables, and $\sum_{m=1}^{M} \epsilon_{imn} = 1$. If S_{in} is generated from component j, then $\epsilon_{imn} = 1$ if $m = j$ and $\epsilon_{imn} = 0$ if $m \neq j$.

$$P(S_n|\epsilon, \mu, \sigma) = \prod_{i=1}^{K} \prod_{m=1}^{M} N(S_{in}|\mu_{im}, \sigma_{im})^{\epsilon_{imn}} \tag{4.2}$$

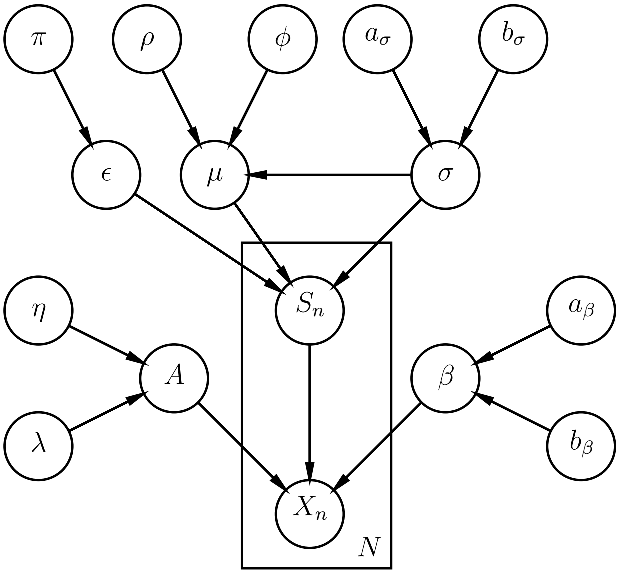


Figure 40: Parameters of the NIFA model are summarized in a directed acyclic graph.

The loading matrix A is modelled with a truncated normal prior where $a = 0$ and $b = \infty$ indicating each entry A_{ji} falls within the interval $[0, \infty)$. η and λ denotes the mean and the inverse of the variance. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

$$\begin{aligned} &P(A_{ji}|\eta_{ji}, \lambda_i) \\ =&(2\pi)^{-\frac{1}{2}}(\lambda_i)^{\frac{1}{2}} \exp(-\frac{1}{2}\lambda_i(A_{ji} - \eta_{ji})^2) \cdot \frac{1(A_{ji} \geq 0)}{1 - \Phi(-\eta_{ji}\lambda_i^{\frac{1}{2}})} \end{aligned} \tag{4.3}$$

The dependency structure of the NIFA model is summarized in Fig. 40. We assume the noise parameter β comes with a Gamma prior with parameter a_β and b_β. The membership indicator ϵ_{imn} comes with a Bernoulli prior with mixing proportion π_{im}. The μ_{im} is assumed to follow a Gaussian distribution with parameters ρ_{im} and ϕ_{im} and the inverse of the variance σ comes with a Gamma distribution with parameters $a_{\sigma_{im}}$ and $b_{\sigma_{im}}$.

4.2.1.2 Parameter Inference The joint likelihood $P(X, A, S, \epsilon, \mu, \sigma, \beta)$ is as follows (Eq. 4.4).

$$\begin{aligned} P(X, A, S, \epsilon, \mu, \sigma, \beta) &= \prod_{n=1}^{N} P(X_n|A, S_n, \beta) P(S_n|\epsilon, \mu, \sigma) \\ \prod_{j=1}^{P}\prod_{i=1}^{K} P(A_{ji}|\eta_{ji}, \lambda_i) &\prod_{i,m,n} P(\epsilon_{imn}) \prod_{i,m} P(\mu_{im}) P(\sigma_{im}) P(\beta) \end{aligned} \tag{4.4}$$

In order to efficiently infer the parameters, we apply variational inference technique, more specifically, mean-field approximation [20]. By assuming each variational parameter is independent of each other, we formulate the joint posterior distribution $Q(S, A, \epsilon, \mu, \sigma, \beta)$ (see Eq. 4.5) for the model and minimize the KL-divergence between Eq.4.4 and Eq.4.5 to derive the expression $q(\cdot)$ for each variational parameter as an approximation of single posterior distribution.

$$\begin{aligned} Q(S, A, \epsilon, \mu, \sigma, \beta) &= \prod_{n=1}^{N} q(S_n) \cdot \prod_{j,i} q(A_{ji}) \cdot \prod_{i,m,n} q(\epsilon_{imn}) \\ &\cdot \prod_{i,m} q(\mu_{im}) q(\sigma_{im}) \cdot q(\beta) \end{aligned} \tag{4.5}$$

The derivations of variational updates for our model are detailed in Sec. B.3.

4.2.1.3 Hyperparameters Our model has a number of hyperparameters, however, most of them are Bayesian priors and have relatively little impact on the results. One of the main hyperparameters of considerable relevance is the number of latent factors K. For the independent factor analysis model, the typical approach is to calculate the likelihood or ELBO (variation lower bound), comparing values directly or with BIC criterion [78] and selecting K corresponding to the optimal values. Since scRNA-seq data often has thousands of cells, the computation for likelihood-based or ELBO-based tuning method is time-intensive and impractical. Instead we can rely on variance-based method SVD with BIC criterion [4] to figure out a conservative estimate of the number of latent factors. Importantly we use this only as a reference value, we perform all our evaluations across a range of K parameters.

We have one more discrete hyperparameter which is the number of Gaussian mixtures M for each latent factor. However, this parameter needs to be just the maximum number of components one can expect to find. Since our model fits the Gaussian mixtures by variational inference, it has the desirable property that the number of mixture components is determined automatically as some of the mixing coefficients go to 0. In experiment we set this hyperparameter to be be 4 and find that for non-developmental datasets, where we expect to find discrete cell types, the final number of modes is usually either one or two. This conforms to the biological intuition that cell types differ from each other by a set of (not necessarily unique) ”marker” genes. Such marker genes typically have a bimodal distribution corresponding to high and low expression. While the distributions may overlap due to technical noise we typically do not observe intermediate expression modes.

4.2.1.4 NIFA initialization NIFA updates are relatively expensive and thus a good initialization can significantly affect running time. We initialize NIFA with a simple matrix decomposition with non-negativity constraints on the loadings. Specifically, we initialize NIFA with a solution to a simpler optimization problem.

$$\min_{A,S} ||X - AS||_F + \lambda_1 ||A||_F + \lambda_2 ||S||_F \quad \text{subject to } A \geq 0 \tag{4.6}$$

This problem can be solved quickly by alternating least squares.

4.2.2 Preprocessing Pipeline

The data preprocessing pipeline is illustrated in Fig. 41. Some pre-processing steps were only applied to certain methods. For example, we employed SVD smoothing (that is reconstructing the input as a truncated SVD with rank=50) because it makes the NIFA constant-variance Gaussian error assumption more valid. However, we found that empirically this had little effect on the results. For NMF we used KL divergence (or equivalently a Poisson error model) which is most appropriate for unsmoothed data. For NIFA we chose to z-score the input by row (gene). The z-scoring operation theoretically makes it easier to pick up on small variance but highly deferentially expressed genes and it produced modest improvement in most (though not all) benchmark datasets. Row z-scoring was not applied to NMF as it produces negative numbers. It was also not applied to ICA as the ICA objective function references only the shape of the factor distribution and thus is invariant under row scaling.

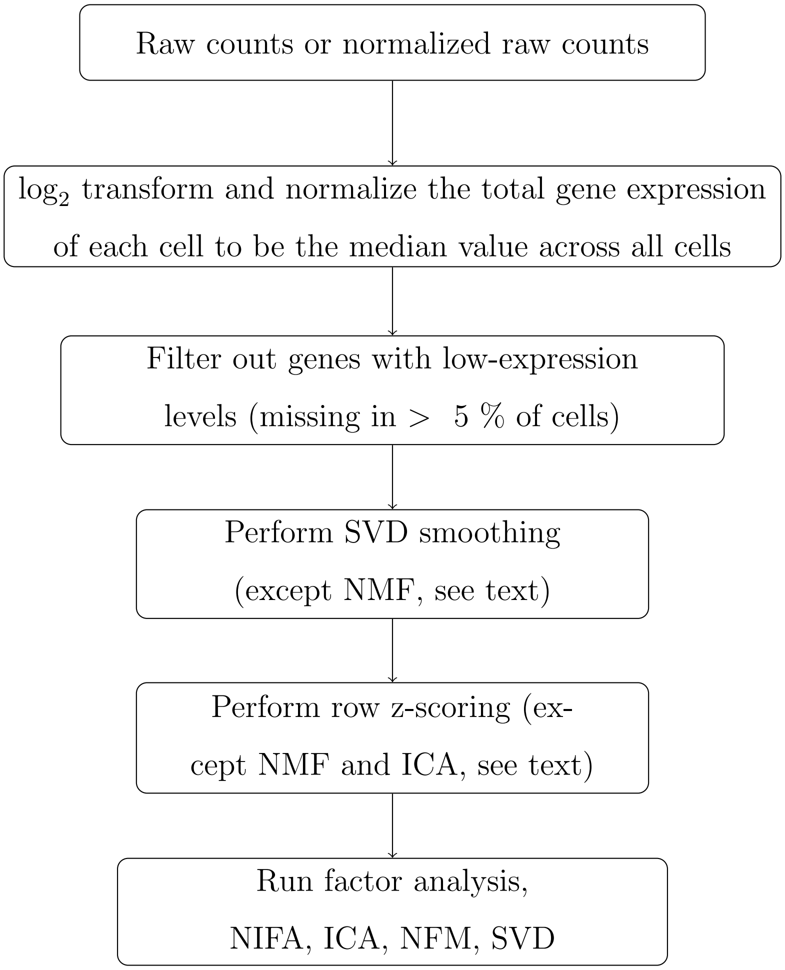

Raw counts or normalized raw counts
$\log_2$ transform and normalize the total gene expression of each cell to be the median value across all cells
Filter out genes with low-expression levels (missing in > 5 % of cells)
Perform SVD smoothing (except NMF, see text)
Perform row z-scoring (except NMF and ICA, see text)
Run factor analysis, NIFA, ICA, NFM, SVD

Figure 41: Preprocessing workflow.

4.2.3 Simulation Details

The dimension of the simulated matrix X is set to be 2000-by-500 (gene-by-sample). There are 6 latent factors, each of which contains 2 Gaussian mixtures. The dimension of loading matrix A is 2000-by-6 with each column corresponding to a single loading and the matrix S is 6-by-500 with each row corresponding to a single latent factor. We draw the first loading vector from Gamma distribution $\Gamma(5, 1)$. Then the subsequent loadings are simulated by adding noise following Gaussian distribution $N(0, 2)$ to the first simulated loading. We take the absolute values of noise to make sure loadings are kept positive. In this way, we can control the collinearity to simulate correlated loadings. Each latent factor is generated hierarchically. First we draw mean and variance parameters for the first Gaussian mixture associated with each latent factor from Gaussian distribution $N(2, 10)$ and Gamma distribution $\Gamma(10, 1)$ correspondingly. Regarding each latent factor the rest of mixtures are generated by adding noise drawn from a uniform distribution $U(2, 4)$ to the mean of the first mixture. Then each entry in the latent factor is assigned to any of the mixtures with probability and the exact value is drawn based on the distribution of assigned mixture. Finally X is generated as the sum of AS and noise drawn from Gaussian distribution $N(0, 0.1)$. In order to generate non-negative NMF input we offset the matrix by a minimum constant $c = \min(C)$ which makes the result $X + c$ non-negative.

4.2.4 Evaluation

We compare NIFA with NMF [49] and ICA (fastICA implementation, [110]). For NMF we used KL loss which we found dramatically outperformed the square loss alternative.

Cell-Type Correspondence We compute the absolute Pearson correlations between each cell type and decomposed factors and we annotate the factor by the corresponding cell type with maximum absolute correlation.

Pathway Enrichment We perform a hypergeometric test on each of the loadings with the top 500 genes as foreground and the rest as background. For SVD and ICA, we use the absolute values of loadings to perform the test. The p values are adjusted with Benjamini-Hochberg procedure and we denote pathways with adjusted p-val < 0.05 as significant enriched pathways. The pathway enrichment is summarized as average fold enrichment across the significant loading-pathway associations.

4.2.5 Datasets included

We test NIFA on several gold or silver standard scRNA-seq dataset [55]. The gold standard dataset contains relatively homogeneous cell lines or the experimental conditions are well controlled. The silver standard dataset defines cell types based on expert knowledge. The data is mostly downloaded through (https://github.com/hemberg-lab/scRNA.seq.datasets) or corresponding GEO repositories. We also include simulated datasets generated by Splatter [169] using Kumar [80] and Zheng [172] as simulation input. The complete sets of datasets are described in Table 8.

Abbreviation	Protocol	Evidence	Type	Tissue	Cells	Cell Types
Camp [27]	SMARTer	Homogeneous cell line	Gold	Human: Liver	777	7
ImmuTherapy [138]	Smart-Seq2	k-means clustering	Silver	Human: Metastatic melanoma	16,291	11
Klein [74]	inDrop	Principal genes identified by PCA	Silver	Mouse: Embryonic Stem Cells	2,717	4
Kolodziejczyk [75]	SMARTer	Homogeneous cell line	Gold	Mouse: Embryonic Stem Cells	704	3
Li [94]	SMARTer	Homogeneous cell line	Gold	Human: Colorectal Tumors	561	9
Liu [102]	10x drop-seq	k-means clustering & specific markers	Silver	Mouse: Spleens or Tumors	1,607	13
Nestorowa [117]	Smart-Seq2	hierachical clustering & specific markers	Silver	Mouse: Hematopoietic Stem Cells	1,656	9
Olsson [124]	SMARTer	Flow cytometry & cell sorting	Gold	Mouse: Hematopoietic Stem Cells	382	4
SimKumar4easy [39]	NA	Simulated	Gold	NA	500	4
SimKumar4hard [39]	NA	Simulated	Gold	NA	499	4
SimKumar8hard [39]	NA	Simulated	Gold	NA	499	8
Zhengmix4eq [39]	NA	Simulated	Gold	NA	3,555	4
Zhengmix4uneq [39]	NA	Simulated	Gold	NA	6,414	4
Zhengmix8eq [39]	NA	Simulated	Gold	NA	3,971	8

Table 8: The complete list of datasets evaluated in this study. Gold datasets are those where cell types are determined due to the experimental design (for example by sorting cells). Silver datasets are those where cell types were assigned from the data using biological prior knowledge.

4.3 Results

4.3.1 Simulation Studies

We simulate data with $P = 2000$, $N = 500$, $K = 6$ (number of factors) and $M = 2$ (number of mixtures associated with each factor). We simulate the latent factor independently but the columns of the loading matrix A are correlated. Simulation details are described in section 4.2.3. For the decomposition, we set $K = 8$ (all methods) and $M = 3$ to see if NIFA can robustly recover the right latent factors given larger K and M, which is often the case in practice. As shown in Fig. 42, none of the methods can recover all latent factors since the loadings are highly correlated. But NIFA is able to accurately recover most of latent factors compared with alternative decomposition methods. NIFA also correctly recovers the number of mixture components as one of the mixing coefficients goes to 0 (not shown).

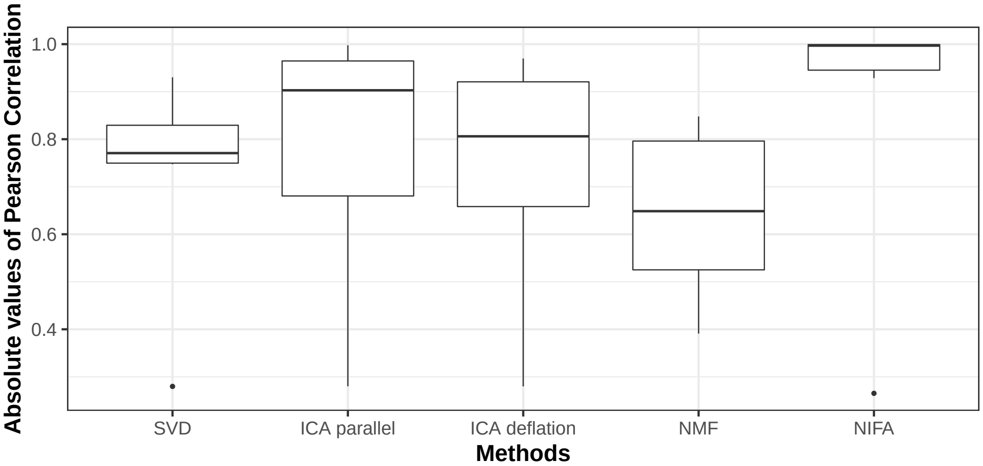


Figure 42: Evaluation on a simulated dataset. Boxplot of the correlation between simulated S and those recovered by SVD, ICA (parallel), ICA (deflation), NMF and NIFA. We find that the best performance is achieved by NIFA compared with all the other common methods.

4.3.2 Evaluation

There is a number of ways to evaluate factor analysis models. One natural evaluation is the reconstruction error (see [91] for example). However, for any decomposition there are infinitely many alternatives with exactly the same reconstruction error, yet these may differ greatly with respect to the individual factors and loadings. Instead of reconstruction error we focus on evaluating the biological utility in several different ways.

4.3.2.1 Cell type identification One of the desirable properties of an interpretable factor analysis is that there is an one-to-one correspondence between the factors and a known data generating variable. In the case of scRNA-seq data the gold or silver standard of cell-type identity is one such variable. In the ideal case each cell type corresponds to a unique factor in the model. In order to evaluate this property we compute maximum one-to-one correlations between factors and cell-type assignments (Fig. 43). We find that on average our NIFA model performs better than NMF and ICA at this cell-type detection task. Importantly, while there can be large differences between NMF and ICA, the performance of NIFA (which combines features of both) always tracks with the best method.

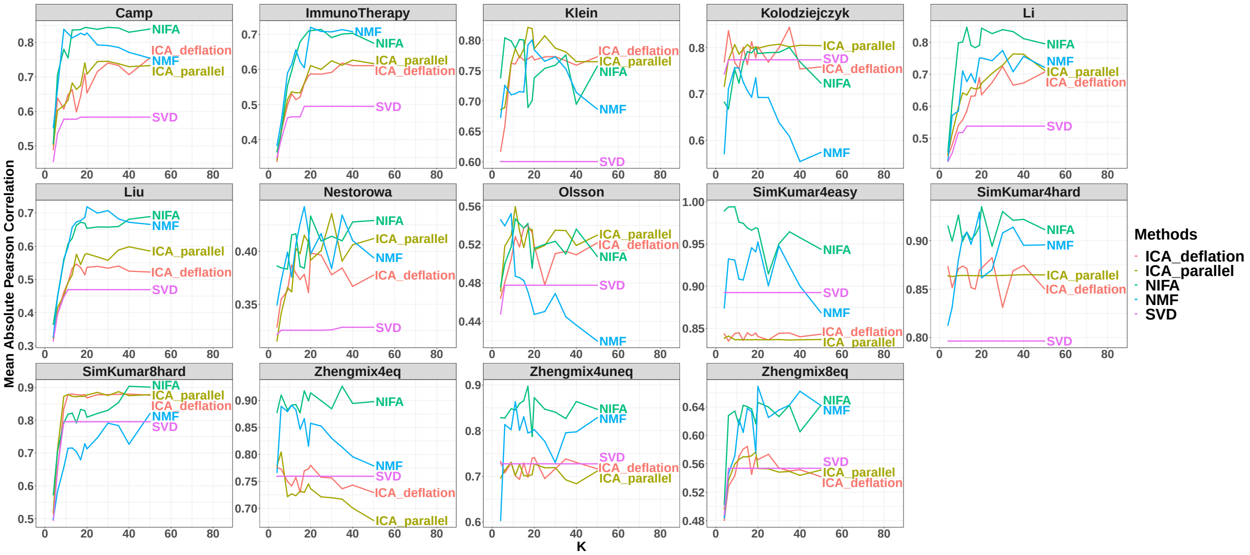


Figure 43: Evaluation of one-to-one correspondence between factors and cell types. Given a set of factors and a set of cell-type labels we evaluate the maximum correlation between each cell type and a factor. For clarity, we plot the mean correlation value across all cell types. In order to account for the possibility that different models may need different number of factors (K), we report the results at varying K. We compare NIFA with ICA, NMF (KL-loss), and SVD as a baseline. We find that on average our NIFA model performs better than NMF and ICA and importantly while there can be large differences between NMF and ICA the performance of NIFA (which combines features of both) always tracks with the best method.

4.3.2.2 Pathway enrichment Of course, one important feature of factor analysis models is that the factors should be interpretable even in the absence of any ground truth knowledge. In such cases the factors are interpreted by inspecting the genes in their loading. The expectation is that for a factor that captures a unique biological variable (which could be binary cell type or continuous pathway activation) the top loading genes are enriched for a few known functional modules. We evaluate this property by computing pathway enrichment for each factor as a hypergeometric test with the top 500 genes as foreground. This evaluation strategy allows us to evaluate the general biological validity of the model, independently of cell-type annotation. In this way the model can be credited for finding factors which capture pathway or cell-type signals even if these do not correspond to an annotated cell type. The pathway databases we use are ”canonical pathways” from MsigDB [100] and a comprehensive set of cell-type markers from xCell [9]. For canonical pathways we excluded pathways that had greater than 20% overlap with ribosomal or mitochondria genesets (defined as ”KEGG_RIBOSOME” and ”KEGG_OXIDATIVE_PHOSPHORYLATION” respectively). We found that these were consistently enriched but provided little biological insight as variations of these pathways were often technical.

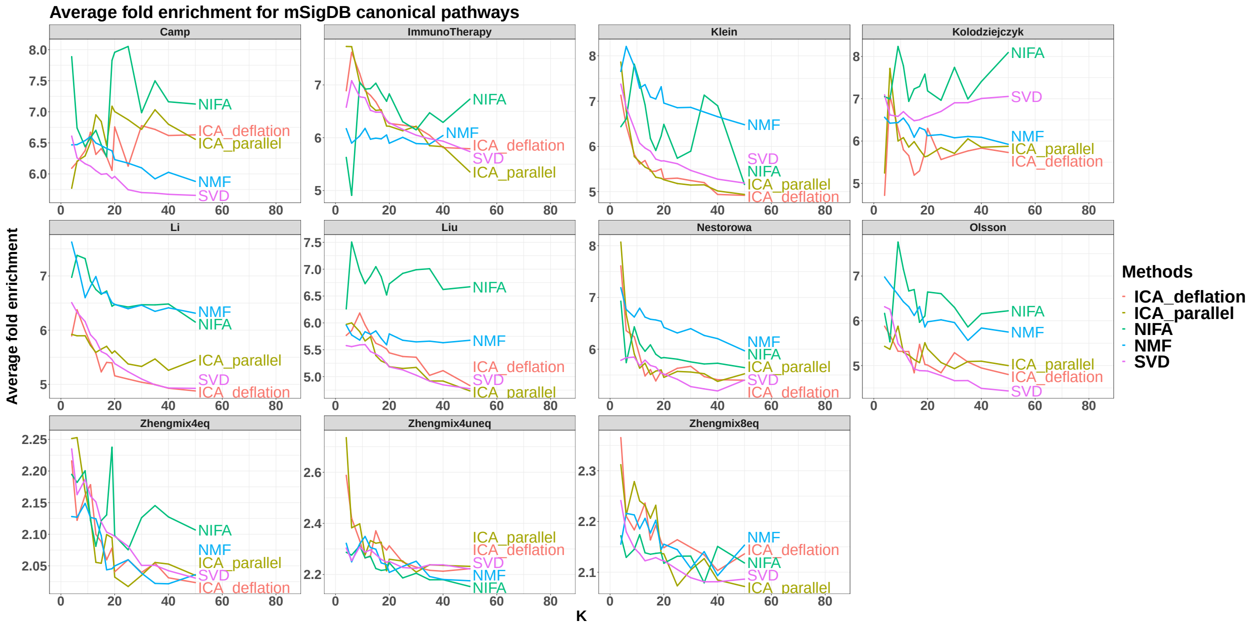


Figure 44: Pathway enrichment of ”canonical pathways” from mSigDB [100]. Enrichment is quantified as average fold enrichment among all factor-pathways pairs where the pathway is significantly over-represented in the top 500 loading genes (hypergeometric test, FDR<0.05). The first two rows are biological datasets. The last row (Zhengmix4eq, Zhengmix4uneq and Zhenmix8eq) are simulated datasets. All SimKumar datasets are excluded from this evaluation as they were not supplied with real gene names.

We then quantify the overall biological enrichment of a single loading vector as the mean fold enrichment for pathways that are significant at FDR<0.05. Pathway enrichment metrics summarized across all factors are plotted in Fig. 44 and Fig. 45 for canonical pathways and xCell respectively. We find that not surprisingly the performance of all methods is much better for real biological datasets than simulated ones (Zhengmix4eq, Zhengmix4uneq and

Zhenmix8eq). We also find that among the biological datasets NIFA is a consistently top performer in both "canonical pathway" and xCell evaluations, though the effect is more dramatic for xCell.

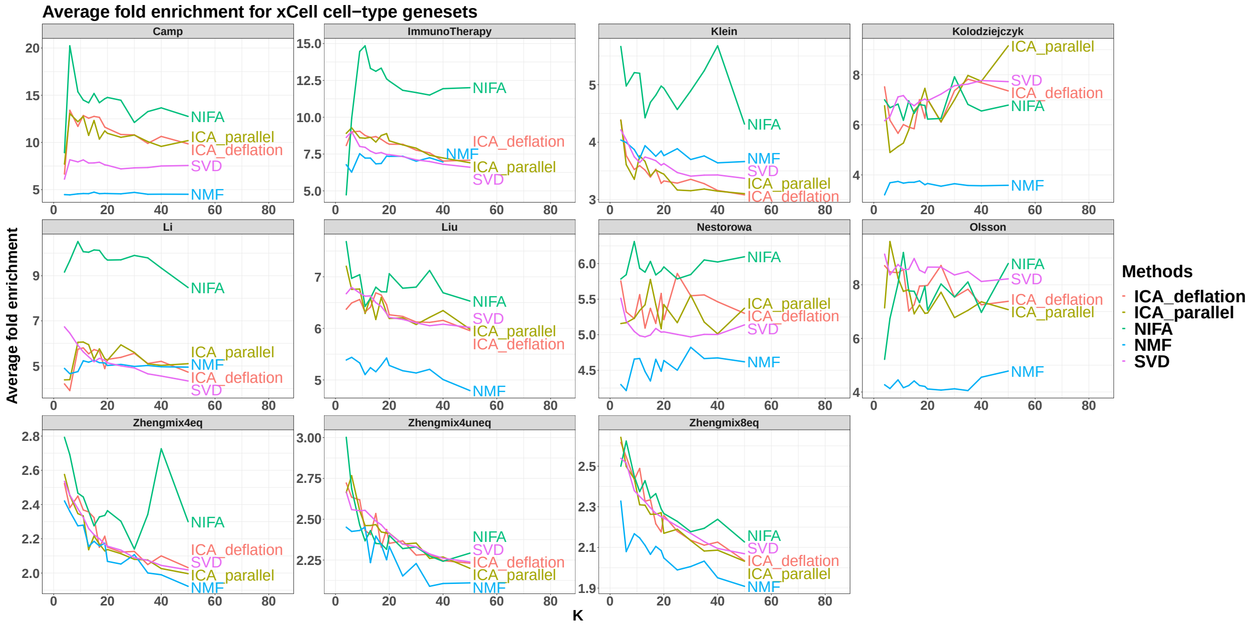


Figure 45: Pathway enrichment for xCell cell-type signatures. This figure is generated identically to Fig. 44 but using the xCell genesets.

4.3.3 In-depth evaluation of the Sade-Feldman *et al.* immunotherapy dataset

Antibodies that block immune checkpoint proteins, including CTLA4, PD-1, and PD-L1 are increasingly used to treat a variety of cancers. While checkpoint inhibitor (CI) therapy can be remarkable effective, not all patients respond [81]. Determining the biological factors that facilitate or impede response to CIs remains an important and unresolved problem

In order to demonstrate how NIFA can be used to gain biological insight we performed several in-depth analyses of the Sade-Feldman *et al.* immunotherapy dataset. This dataset consists of 16,291 individual immune cells from 48 tumor samples of melanoma patients treated with checkpoint inhibitors. The dataset contains both pre-treatment and post-treatment samples and the patients are classified into responders and non-responders.

We applied our NIFA model to the entire single-cell dataset using K=25 which corresponds to the k with maximal correlation with known cell-type annotations (see Fig. 43). The distributions of the inferred factors and the corresponding inferred Gaussian mixture fits are plotted in Fig. 46. Each NIFA factor that has the best correspondence to human annotations is given the same name. NIFA finds both uni- and multi-modal factors and as expected the multi-modal factors are more likely to correspond to cell types.

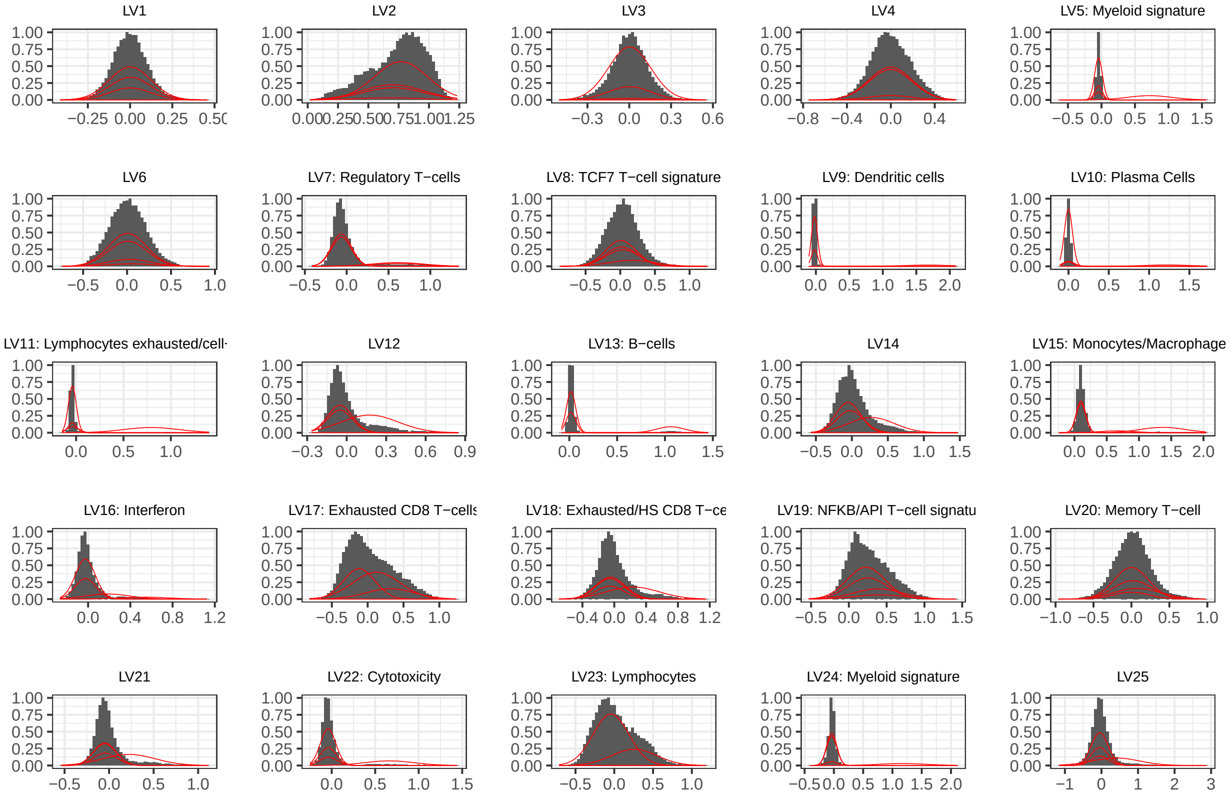


Figure 46: Factor histograms and mixture model fits for the Sade-Feldman *et al.* immunotherapy dataset. Factors that best correspond to cell types identified in the original study are labeled accordingly. NIFA finds both multi-modal and unimodal factors and as expected the multimodal factors are more likely to represent cell types

In order to investigate which variables are associated with immunotherapy response the resulting factors were mean aggregated to a single value for each unique patient sample. We also summarized the human annotated cell-type indicators as their mean values, corresponding to fraction of cells in sample. The resulting summary statistics were tested for association with response using Wilcoxon ranksum and Benjamini-Hochberg FDR adjusted (separately for NIFA factors or human annotations). Pre- and post-treatment samples were analyzed separately and the resulting variables that were significant in either the pre-treatment or post-treatment comparison at FDR<0.2 are plotted in Fig. 47A. Top loading genes corresponding to each significant NIFA factor are show in Table 9.

Each NIFA factor that has the best correspondence to human annotations is given the same name and the results are grouped with grey ovals in Fig. 47. We find that for these matched variables there is an overall good correspondence between the results of NIFA factors and human cell-type annotations. Specifically, both methods discover B-cells as the variable most positively predictive of response and a CD8 T-cell/exhaustion/cell-cycle signature (termed "Lymphocytes exhausted/cell-cycle" in the original study) as the most negatively predictive.

For some subtle pattern the results of NIFA and human annotations can diverge. Human annotations such as "Lymphocytes" and "Cytotoxicity" were not well reproduced by NIFA (correlation of 0.46 and 0.46 respectively) and the corresponding NIFA variables are not significant. On the other hand, NIFA found three different T-cell signatures (8, 19 and 20) which were all associated with the "memory T-cell" human annotation and all were significantly predictive of response. One of these signatures has TCF7 as a top loading gene and thus NIFA was able to independently discover one of the key findings of the original study – that the fraction of TCF7 positive T-cells is highly associated with response.

Aside from generally reproducing the main findings of the original study NIFA was able to uncover additional patterns. For example, we find that presence of Tregs is negatively associated with response in the post-treatment samples. The corresponding human annotation is however not significant despite the fact that the two variables are highly correlated (Pearson correlation = 0.71). Human regulatory T-cells are difficult to identify from a transcriptional profiles. There are no genes that are *unique* to this cell-type. The canonical transcription factor (FOXP3) and surface marker (CD25/IL2RA) can also be transiently expressed by non-Treg CD4 cells [29]; on the other hand, because of noise in scRNA-seq data the absence of these markers doesn't exclude Treg status. Upon closer inspection, we find that NIFA is more conservative in designating Tregs than the human annotation counterpart. Using the NIFA mixture components we can perform a hard cell-type assignment based on the probability of being in the high-expression component being>0.5. Using this cutoff, NIFA finds out 1,418 Tregs, in contrast to 1,740 of human annotated ones. We find that these discrepancy is highly non-random and that the NIFA Tregs are more likely to express both FOXP3 and IL2RA (Fig. 47C) indicating that the NIFA Treg signature is more specific.

ID	name	genes
5	Myeloid signature	FCN1, LYZ, TIMP1, S100A9, S100A8, SERPINA1, VCAN, IL1B, IFI30, PLAUR
7	**Regulatory T-cells**	CTLA4, TNFRSF18, RGS1, TIGIT, CD4, BATF, PIM2, PRDM1, FOXP3, ARID5B
8	TCF7 T-cell signature	DGKA, DDX17, SMG1P1, DENND2D, ARHGEF1, DOCK8, NPIPB5, NLRC5, TCF7, N4BP2L2
9	**Dendritic cells**	GZMB, IGJ, PLAC8, NAPSB, ALOX5AP, GPR183, AC096579.7, IRF7, BCL11A, CLIC3
11	**Lymphocytes exhausted/cell-cycle**	STMN1, RRM2, TUBA1B, TYMS, KIAA0101, TUBB, HIST1H4C, HMGB2, NUSAP1, CDK1
13	**B-cells**	CD74, IGHM, MS4A1, CD79A, IGKC, IRF8, CD79B, CD37, BCL11A, CD52
14		HSPD1, FLNA, BIRC3, REL, HSPE1, COTL1, WARS, PSME2, HSPB1, SLC25A3
15	**Monocytes/Macrophages**	CD74, FTL, CTSB, B2M, FTH1, PSAP, S100A11, IFI30, VIM, ALDOA
16	Interferon	ISG15, IFI44L, MX1, IFI6, XAF1, STAT1, ISG20, IFITM1, TRIM22, IFI44
17	**Exhausted CD8 T-cells**	GZMA, RAC2, CLIC1, NKG7, CORO1A, IL32, ARPC1B, CNN2, LCK, PSMB9
19	NFKB/AP1 T-cell signature	VIM, NFKBIA, FOS, TNFAIP3, ANXA1, SLC2A3, CD52, B2M, JUNB, S100A4
20	**Memory T-cells**	EEF1B2, GAS5, TOMM7, LDHB, SELL, IL7R, EIF3E, COX7C, EIF3L, FAIM3
24	Myeloid signature	MT1X, MT1F, CD14, MT1E, FN1, FBP1, MT2A, S100A4, S100A9, AGTRAP
25		C1QBP, NME1, HSP90AB1, GTF3A, NHP2, PPA1, CCT7, CNBP, CCT2, SNHG1

Table 9: Top loading genes for each factor found to be associated with immunotherapy response. To facilitate biological interpretability ribosomal genes and genes that were not provided with HGNC symbols were not considered. Factors that best match the known human annotations were named accordingly and are in bold. Other factors could be clearly identified as coherent biological pathways based on the loadings. Factors 14 and 25 did not have a clear correspondence to any pathways or cell-type signatures and are left unnamed.

Overall, within this dataset a large number of the human-annotated cell types and NIFA factors are associated with response but some general patterns emerge. Specifically, the presence of myeloid cells is negatively associated with response while presence of lymphocytes, exclusive of those with an exhaustion-like phenotype (for example B-cells, CD4 memory cells), is positively associated with response (see Fig. 47A). The general trend that a high myeloid to lymphocyte ratio is associated with worse outcome is observed across a variety of cancers [155]. Our NIFA-based analysis however finds a myeloid signature (NIFA latent factor 24) that correspond to a subset of annotated "Monocytes/Macrophages" cells is *positively* associated with response, with an effect size that is similar to the lymphocyte populations

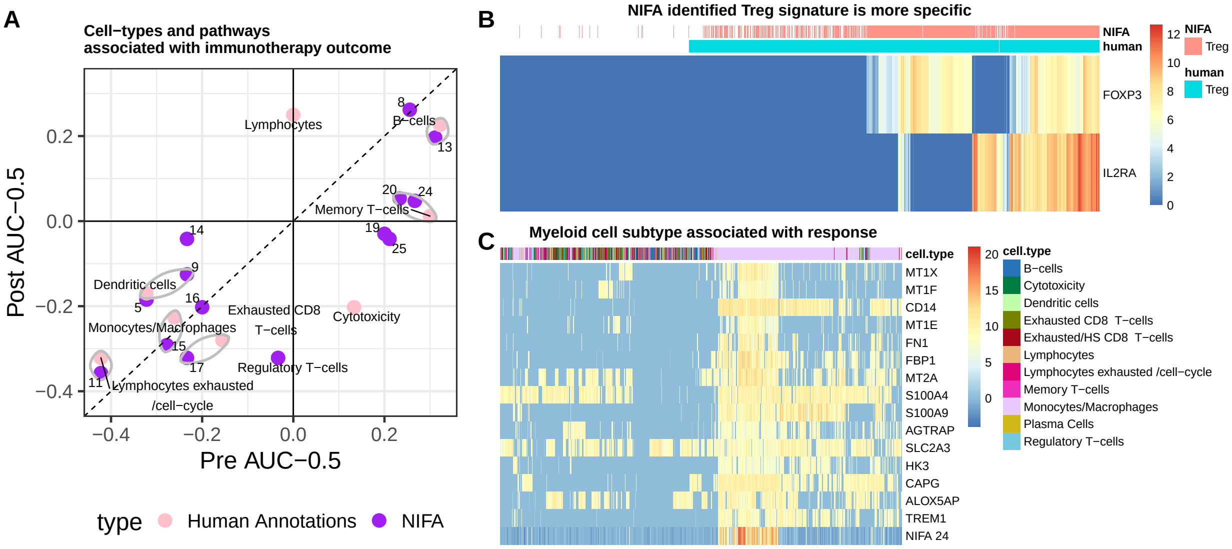


Figure 47: NIFA analysis of signatures associated with immunotherapy response. **A.** NIFA-derived signatures of human-annotated cell-types are mean aggregated per patient sample and the resulting summary statistics are tested for association with immunotherapy response. Variables that are significant at FDR<0.2 are shown with their respective normalized and centered ranksum statistics (ranksum/(number-of-positives × number-of-negatives)-0.5, equivalent to binary classification AUC-0.5). Pre-treatment effects are on the x-axis and post-treatment effects are on the y-axis. NIFA variables most closely matched to a human annotation are grouped with grey ellipses. **B.** Differences in Treg (Regulatory T-cells) identification between NIFA and human annotations. Heatmap of canonical Treg marker genes (FOXP3 and IL2RA) across all cells annotated as Tregs by either method and 1,000 randomly sampled other T cells. Overall, NIFA identifies fewer Treg cells and has a higher correlation with FOXP3 and IL2RA expression. While the NIFA Treg factor is significantly negatively associated with response in post-treatment samples, the corresponding human annotation is not (panel A). **C.** A new myeloid signature positively associated with response. Heatmap of top loading genes along with the factor values for NIFA factor 24 across all cells identified as ”Monocytes/Macrophages” and 1,500 randomly sampled cells. NIFA identifies a subset of the Monocytes/Macrophages calls with unique gene expression. While general myeloid signatures (that is Monocytes/Macrophages and Dendritic cells) were negatively associate with response, the NIFA-24 signature has the opposite pattern (see panel A).

(Fig. 47 A). This myeloid subset is identified by high levels of metallothionein genes (MT1X, MT1F, MT1E and MT2A) and some metabolic genes (see Fig. 47C). Metallothioneins are a family of small proteins that play important roles in metal homeostasis and protection against heavy metal toxicity, DNA damage and oxidative stress [141]. Their induction in tumor-associated macrophages (TAMs) has been noted [50] but to our knowledge this is the first report of an association with clinical outcome.

4.4 Discussion

We propose a factor analysis model designed specifically for single-cell data. The model combines features of PCA, ICA and NMF. Specifically, our model optimizes the PCA-like matrix reconstruction objective with mixture of Gaussians priors on the factors which encourages decomposition along multi-modal directions. We also adopt truncated Gaussian priors on the loadings thus imposing an NMF-like strict non-negativity constraint. Using a variational Bayes framework allows us to automatically fit hyperparameters such as the number of Gaussian mixtures. We evaluate our model using both known cell identity and pathway information and demonstrate that NIFA generates biologically coherent factors that align well with prior knowledge.

One additional feature of our model is that the fully Bayesian framework is readily extensible. For example, it easily supports gene-specific priors for the loadings. This makes it possible to use known biological pathways as additional constraints. We plan on developing this extension in our future work.

5.0 Conclusions & Future Directions

Normalization is a key step in order to reveal biological variations from RNA-seq data. Generative modelling is one category of normalization techniques which not only tackle the problem of representing overall variation structure but also annotate each single variation with different criteria correspondingly.

In chapter 2, we present PLIER which estimate pathway-level regulations by aligning LVs as much as possible with known gene sets. A key observation is that PLIER greatly improves the interstudy concordance which indicates a great reproduciablity in biological findings. A further and more challenging question is whether we could model the pathway-level effects at a finer resolution. PLIER models continuous pathway effects as an aggregation over all cell types and isolates cell-type proportions and pathway effects into separate factors. It is more informative if it is possible to estimate the interactions of these two which represents the celltype-specific pathway effects. Another opportunity is to take the structure of subjects into consideration. For example, if RNA-seq samples come from different tissues, including group-level constraint make it possible to infer tissue-specific pathway effects.

In chapter 3, we illustrate the power of reweighing the contribution of structured factors decomposed by SVD to maximize biological findings without any explicit knowledge about the dataset. If there is no significant signal at all, in which case we can't apply DataRemix directly, we can first apply DataRemix on a related but easier task and the optimized parameters can be transferable to the initial task. This significantly enhances DataRemix's capability to improve biological utility of a broader range of legacy datasets. DataRemix is a flexible parameterization method and it is easy to apply on top of more complicated decomposition frameworks, e.g., HCP by introducing additional parameters other than k, p and μ. One caveat is to be careful about possible overfittings. Another opportunity is to use DataRemix as a measure to quantify how informative a dataset is given different biological objectives. Thus dataset can be assessed quantitatively by its performance with respect to all available tasks.

In chapter 4, we propose a universal framework-NIFA that combines the features of PCA, ICA and NMF for scRNA-seq data. It projects scRNA-seq into uni- and multi-modal factors isolating discrete and continuous variations into separate latent variables. A straightforward extension is to incorporate more structured priors as constraint on the loading part. For example we can align the loadings with known gene sets with spike and slab priors. Another challenging question is how to apply NIFA on large-scale scRNA-seq dataset. With the advancement of scRNA-seq technology, now it is practical to sequence millions of cells in one cohort. One possible strategy is to subsample the large-scale dataset in a systematic manner , apply NIFA and project the resulting factors back into the original high-dimensional space.

5.1 Discussions & Future Directions

5.1.1 The Limitation of Factor Analysis

As a special case of generative modeling, matrix decomposition cannot help generate any new data samples directly from learned parameters. If only loadings and latent factors are learned from a specified optimization setting, you cannot directly generate new samples out of them since you don't explicitly model the noise which provides the variations. Even within a probabilistic framework, e.g. NIFA, it's not valid to generate new samples from the learned posterior distribution. Firstly, the prior distributions are only used to extracted certain patterns and it's hard to verify their validity. Secondly even if the RNA-seq profile can be recovered perfectly from the posterior distribution with respect to likelihood, there is still no guarantee that the posterior distribution is biologically meaningful.

Two state-of-the-art choices to generate new samples in the general field of machine learning are Generative Adversarial Network (GAN) [56] and Variational Autoencoder (VAE) [71]. Meanwhile some generators have been developed specifically for the RNA-seq data. Flux [57] is one of the computational sequencers that simulate the RNA-seq following the shotgun sequencing process. It simulates all intermediate steps and outputs as those from the wet lab experiment based on a limited set of parameters. Meanwhile there are statistical computational sequencers, such as Splatter [169]. Based on the statistical characteristics of the gene expression profiles, the final gene expression values are simulated in a hierarchical way.

The overall objectives of such generative models vary and it highly depends on the goals of study. In some studies, generating new samples from given data works as a transformation which maintains all structural information without leaking any privacy-related information. Other studies attempt to generate new samples which overcome the limitations of wet-lab experiment in order to evaluate computational frameworks in a more controlled setting.

In conclusion factor analysis is helpful to extract specific patterns with respect to certain valid assumptions on the hidden structure of the data. Factor analysis cannot contribute directly to generating new samples, but it's very helpful as a tool for validation.

5.1.2 Incorporate a Richer Set of Biological Knowledge into Factor Analysis

In the PLIER framework, we constrain the loadings to align with most relevant automatically selected subset of pathways in order to identify specific genesets that regulate gene expression. In order to extract specific patterns in the data, we need to formulate corresponding regularization and challenges at the optimization side may arise. Limited biological structures have been incorporated into generative models so far [33]. Some of the promising ideas are limited by the feasibility of efficiently solving the optimization problem. For example, celltype-specific pathway effects are of great interest, but in order to extract certain patterns we need to add an extra hierarchy to the formulation of PLIER which leads to additional difficulties in optimization.

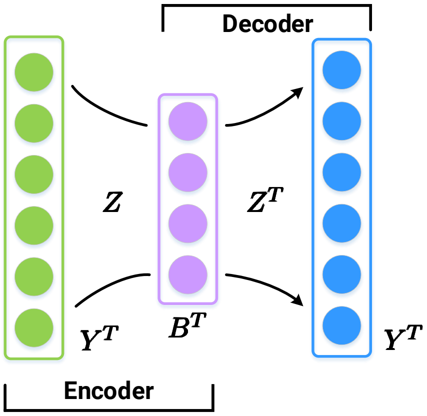


Figure 48: Reformulation of PLIER framework as an autoencoder in the language of neural network.

$$||Y^T - Y^T Z Z^T||_F^2 + \lambda_1 ||Z - CU||_F^2 + \lambda_2 ||Y^T Z||_F^2 + \lambda_3 ||U||_{L^1} \tag{5.1}$$
$$\text{subject to} \quad U \geq 0, \quad Z \geq 0.$$

One of the potential solutions is to formulate the optimization problem in the language of automatic differentiation (AD). As shown in Fig. 48, we reformulate the PLIER framework as a standard autoencoder with regularizations. Loadings Z serves as the weights which connects layers in the network and latent factors B correspond to the compressed layers while the constraints still hold (see Eq. 5.1).

In the PLIER framework, the constraint we impose is $||Z - CU||_F^2$ and in a more general form it can be rewritten as a function $f(Z, C, U)$. As long as f is differentiable, the overall optimization problem can be solved with the help of the automatic differentiation system. Thus we can impose more biologically-meaningful structures into the matrix decomposition framework.

5.1.3 Define Cell Identity and Cell State in the Age of Single-cell Multi-omics

Figuring out the actual cell identity is the most asked question when people run scRNA-seq experiment. People have made substantial progress but the challenges remain [72]. Cell type is a long-standing concept which keeps evolving as we have a better understanding of the underlying molecular mechanism. It has been defined from the perspectives of morphology, cell location, neighbors, transceiptome, proteome and cell functions [113, 167]. Some of the modern views of cell types are listed below.

- **Core regulatory complexes (CoRCs)** From an evolutionary point of view, cell type is defined by evolutionary change of core regulatory complexes (CoRCs) of transcription factors [10].
- **Cell dynamic** Same types of cells should have same functionality if the surrounding environment is similar. If cells are treated as black boxes, with the same inputs, cells which produce similar outputs belong to the same category. Such functional profiles records how cells interact with their environment and provides a dynamic definition of cell types[31]. Instead of a static snapshot, this dynamic profile provides a more precise description of the cell based on its functionality.

- **Cell Lineage** Cell types can be annotated based on its lineage history. A unique differential path distinguishes a certain cell type from the others.

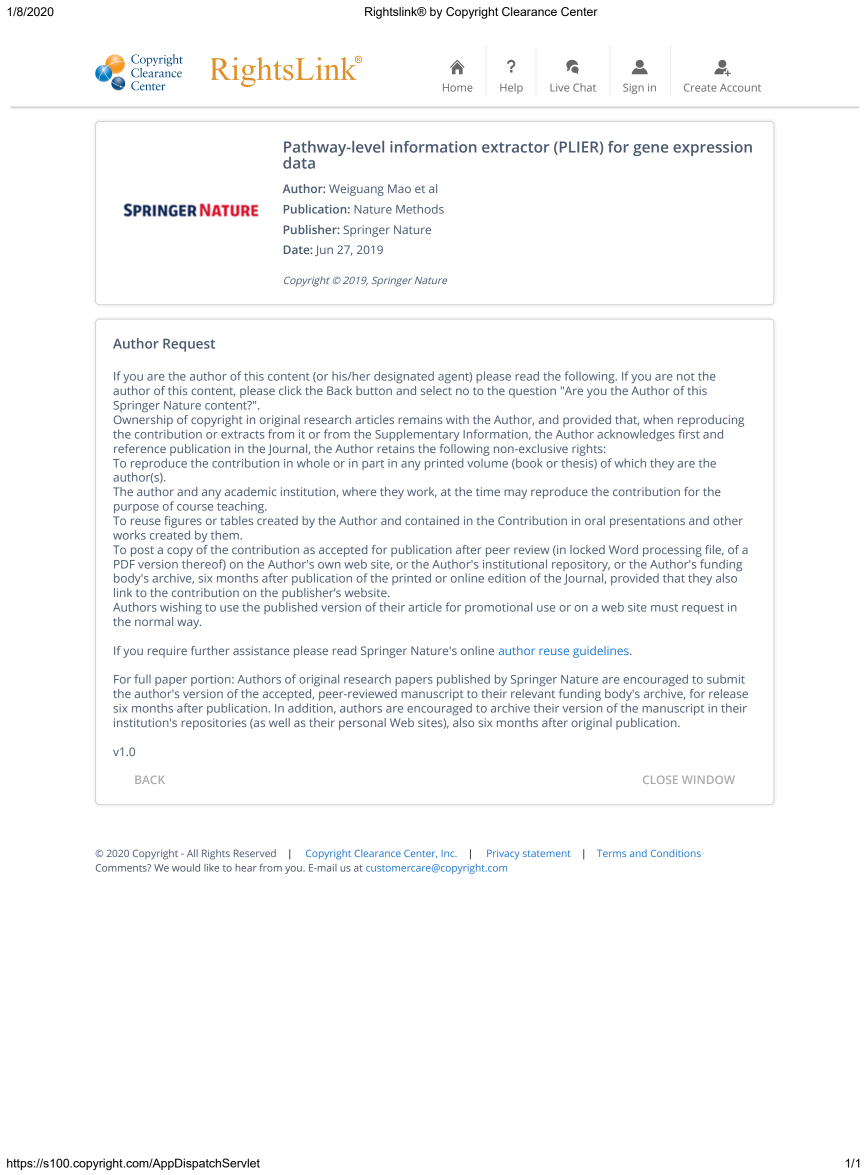
Cell state is another concept convoluted with cell type [156]. Cell type can be thought as a collection of corresponding cell states. Wadding landscape [162] has become the classic illustration to show how gene activity affects transition between cell state or how cell fate is triggered [112]. Classic examples include different cell phases in the cell cycle and different cell conditions of CD8+ T cells [161].

Single-cell sequencing technique makes it possible to generate a molecular profile at single-cell resolution for the first time. Along with scRNA-seq and CITE-seq [151], single-cell ChIp-seq, single-cell DNase-seq and single-cell Hi-C have been developed to decipher the heterogeneity of chromatin accessibilities in a cell population. Besides these developed techniques, single-cell proteomics [142] is promising to measure the activities of proteins as the basic functional element in the cell. The other emerging technique called spatial transcriptomics [147] enables us to measure gene activities in a morphological context.

All these perspectives add up to a holistic view of cell identity and this leads to a broad new field of single-cell multi-omics [12, 64]. Several computational models have been developed to leverage single-cell multi-omics data, such as MOFA [13] and MOFA+ [11]. Also some of general-purpose frameworks have been adopted to single-cell multi-omics, such as iCluster [139, 111] and GFA [89]. These experimental and computational advancements enable us to retrieve a more comprehensive and quantitative description of molecular activities at single-cell resolution which makes it possible to provide a quantitative definition of cell types and cell states.

Appendix A

Permissions to Reuse Copyrighted Content



1/8/2020 Rightslink® by Copyright Clearance Center

Pathway-level information extractor (PLIER) for gene expression data

Author: Weiguang Mao et al

Publication: Nature Methods

Publisher: Springer Nature

Date: Jun 27, 2019

Copyright © 2019, Springer Nature

Author Request

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".
Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online author reuse guidelines.

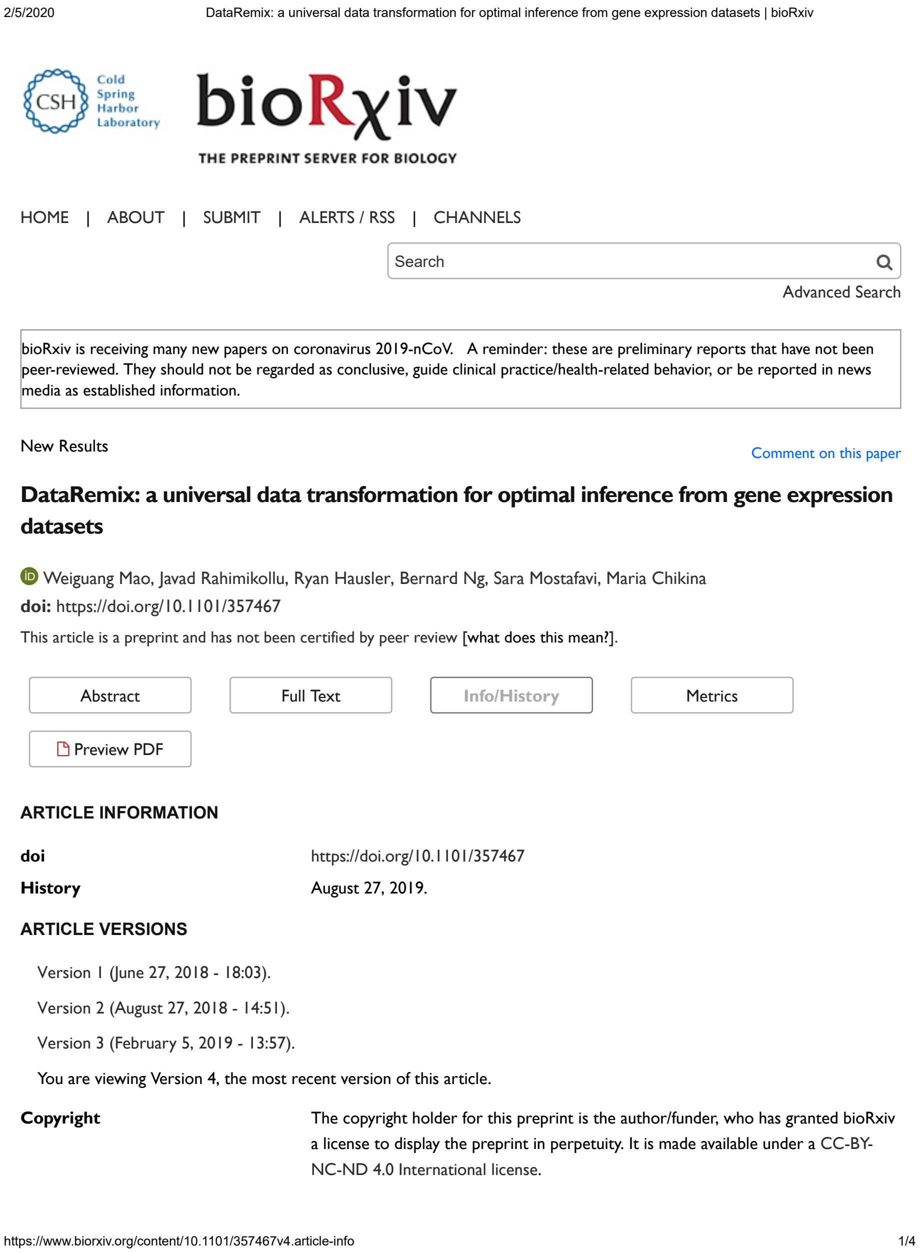
For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0

© 2020 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

https://s100.copyright.com/AppDispatchServlet 1/1

Figure 49: Copyright Permission for PLIER [109].



2/5/2020 DataRemix: a universal data transformation for optimal inference from gene expression datasets | bioRxiv

Cold
Spring
Harbor
Laboratory
CSH
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS
Search
Advanced Search

bioRxiv is receiving many new papers on coronavirus 2019-nCoV. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

DataRemix: a universal data transformation for optimal inference from gene expression datasets

Weiguang Mao, Javad Rahimikollu, Ryan Hausler, Bernard Ng, Sara Mostafavi, Maria Chikina

doi: https://doi.org/10.1101/357467

This article is a preprint and has not been certified by peer review [what does this mean?].

ARTICLE INFORMATION

doi	https://doi.org/10.1101/357467
History	August 27, 2019.

ARTICLE VERSIONS

Version 1 (June 27, 2018 - 18:03).

Version 2 (August 27, 2018 - 14:51).

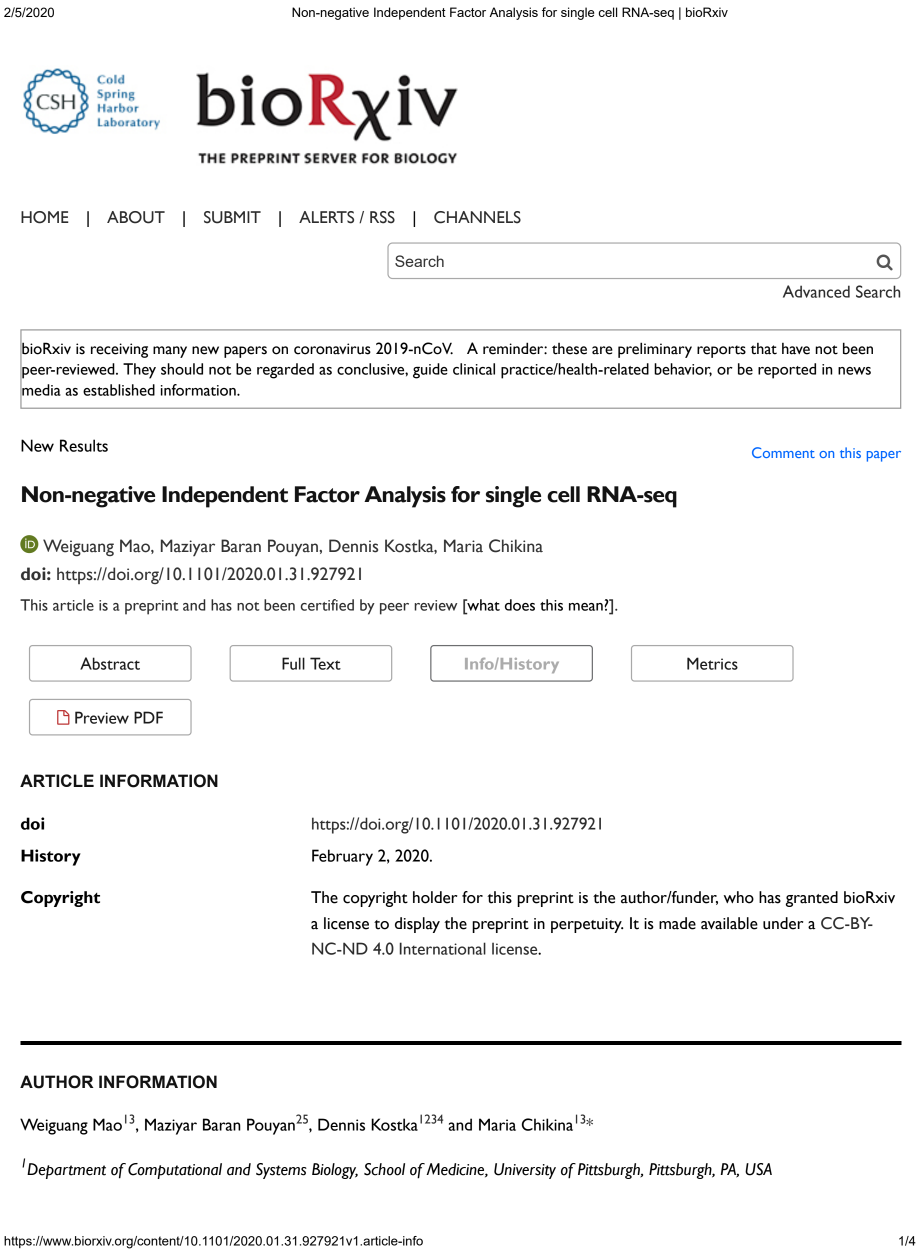
Version 3 (February 5, 2019 - 13:57).

You are viewing Version 4, the most recent version of this article.

Copyright The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

https://www.biorxiv.org/content/10.1101/357467v4.article-info 1/4

Figure 50: Copyright Permission for DataRemix [107].



2/5/2020 Non-negative Independent Factor Analysis for single cell RNA-seq | bioRxiv

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS
Search
Advanced Search

bioRxiv is receiving many new papers on coronavirus 2019-nCoV. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

Comment on this paper

Non-negative Independent Factor Analysis for single cell RNA-seq

Weiguang Mao, Maziyar Baran Pouyan, Dennis Kostka, Maria Chikina

doi: https://doi.org/10.1101/2020.01.31.927921

This article is a preprint and has not been certified by peer review [what does this mean?].

ARTICLE INFORMATION

doi	https://doi.org/10.1101/2020.01.31.927921
History	February 2, 2020.
Copyright	The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

AUTHOR INFORMATION

Weiguang Mao[13], Maziyar Baran Pouyan[25], Dennis Kostka[1234] and Maria Chikina[13]*

[1]*Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA*

https://www.biorxiv.org/content/10.1101/2020.01.31.927921v1.article-info 1/4

Figure 51: Copyright Permission for NIFA [108].

Appendix B

Supplementary Notes

B.1 PLIER

	pathway	LV.index	AUC	p.value	FDR
1	KEGG_SPLICEOSOME	2	0.69	0.000230	0.000754
2	MILI_PSEUDOPODIA_CHEMOTAXIS_UP	2	0.69	0.005779	0.012480
3	MIPS_ANTI_HDAC2_COMPLEX	3	0.83	0.016842	0.031018
4	MIPS_ALL_1_SUPERCOMPLEX	3	0.78	0.011502	0.022561
5	MIPS_RC_COMPLEX_DURING_S_PHASE_OF_CELL_CYCLE	3	0.99	0.005541	0.012083
6	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_BIOSYNTHESIS	3	0.95	0.009092	0.018548
7	MIPS_INTEGRATOR_RNAPII_COMPLEX	3	0.98	0.005849	0.012566
8	MIPS_TFTC_TYPE_HISTONE_ACETYL_TRANSFERASE_COMPLEX	3	0.97	0.007403	0.015533
9	MIPS_P2X7_RECEPTOR_SIGNALLING_COMPLEX	3	0.93	0.011882	0.023219
10	WELCSH_BRCA1_TARGETS_DN	3	0.63	0.005763	0.012480
11	FAELT_B_CLL_WITH_VH3_21_UP	3	0.69	0.021097	0.037702
12	DEBIASI_APOPTOSIS_BY_REOVIRUS_INFECTION_DN	3	0.74	7.56e-11	1.29e-09
13	KLEIN_PRIMARY_EFFUSION_LYMPHOMA_UP	3	0.69	0.016030	0.029704
14	REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	4	0.61	0.000456	0.001412
15	REACTOME_DEADENYLATION_OF_MRNA	5	0.89	0.007313	0.015392
16	KEGG_PROTEASOME	5	0.85	0.000148	0.000490
17	MIPS_PA700_COMPLEX	5	0.92	0.001106	0.002933
18	MIPS_INO80_CHROMATIN_REMODELING_COMPLEX	5	0.90	0.019300	0.034867
19	MIPS_MLL1_WDR5_COMPLEX	5	0.86	0.001653	0.004168
20	REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION	5	0.79	0.015640	0.029210
21	MOOTHA_TCA	5	0.97	0.001387	0.003604
22	CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_DN	5	0.68	4.68e-14	1.29e-12
23	MIPS_HES1_PROMOTER_NOTCH_ENHANCER_COMPLEX	7	0.99	0.005060	0.011182
24	DAZARD_RESPONSE_TO_UV_NHEK_DN	7	0.64	5.18e-05	0.000199
25	MARTINEZ_RESPONSE_TO_TRABECTEDIN	7	0.71	0.009547	0.019097
26	KEGG_RIBOSOME	8	0.65	0.012098	0.023454
27	PUIFFE_INVASION_INHIBITED_BY_ASCITES_DN	8	0.66	0.001007	0.002693
28	REACTOME_EGFR_DOWNREGULATION	9	0.77	0.023705	0.041468
29	LIN_APC_TARGETS	9	0.66	0.013443	0.025465
30	JOHNSTONE_PARVB_TARGETS_2_UP	9	0.60	0.024368	0.042404
31	WOOD_EBV_EBNA1_TARGETS_UP	9	0.69	0.000942	0.002548
32	DASU_IL6_SIGNALING_UP	9	0.70	0.010567	0.020913
33	HOWLIN_CITED1_TARGETS_1_DN	10	0.85	0.000682	0.001956
34	GRANDVAUX_IRF3_TARGETS_DN	10	0.88	0.024748	0.042840
35	LEE_RECENT_THYMIC_EMIGRANT	10	0.61	0.005036	0.011166
36	HAHTOLA_SEZARY_SYNDROM_DN	13	0.80	0.001391	0.003604
37	CHEN_LVAD_SUPPORT_OF_FAILING_HEART_UP	13	0.66	0.006483	0.013820
38	FLOTHO_PEDIATRIC_ALL_THERAPY_RESPONSE_UP	14	0.70	0.010979	0.021629
39	MemoryTcell-RO-unactivated	15	0.79	0.002048	0.005047
40	TCELLA7	15	0.97	1.82e-07	1.33e-06
41	TCELLA8	15	0.96	9.21e-08	7.36e-07
42	SVM T cells CD4 memory resting	15	0.91	4.06e-08	3.67e-07
43	SVM T cells regulatory (Tregs)	15	0.91	4.33e-07	2.93e-06
44	NKA1	16	0.84	6.24e-06	3.02e-05
45	PID_PRLSIGNALINGEVENTSPATHWAY	16	0.85	0.003277	0.007760
46	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP	16	0.72	0.001528	0.003913
47	EBAUER_MYOGENIC_TARGETS_OF_PAX3_FOXO1_FUSION	16	0.73	0.010111	0.020131
48	ABE_VEGFA_TARGETS_30MIN	16	0.75	0.027548	0.047316
49	CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_2	17	0.68	2.50e-09	2.90e-08
50	Neutrophil-Resting	19	0.63	0.000130	0.000437
51	CHUNG_BLISTER_CYTOTOXICITY_DN	19	0.87	4.49e-05	0.000180
52	Neutrophil-Resting	20	0.70	9.96e-10	1.27e-08
53	KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	20	0.75	0.000741	0.002073

54	GSE59184_Il10_6H	20	0.64	8.49e-05	0.000303
55	ALTEMEIER_RESPONSE_TO_LPS_WITH_MECHANICAL_VENTILATION	20	0.77	9.94e-07	6.05e-06
56	PID_IL27PATHWAY	21	0.86	0.003031	0.007256
57	REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES	21	0.81	0.000231	0.000754
58	REACTOME_INTERFERON_GAMMA_SIGNALING	21	0.91	1.32e-07	1.03e-06
59	GSE33057_Ifng	21	0.72	5.99e-09	6.51e-08
60	GSE19182_Ifng	21	0.92	< 2e-16	6.52e-15
61	GSE36287_Ifng	21	0.77	2.22e-14	7.00e-13
62	JISON_SICKLE_CELL_DISEASE_UP	21	0.73	4.12e-08	3.67e-07
63	UROSEVIC_RESPONSE_TO_IMIQUIMOD	21	0.98	0.000118	0.000407
64	WIELAND_UP_BY_HBV_INFECTION	21	0.76	8.68e-06	4.00e-05
65	SANA_RESPONSE_TO_IFNG_UP	21	0.96	1.14e-10	1.80e-09
66	BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	21	0.83	4.26e-07	2.91e-06
67	Neutrophil-Resting	22	0.89	< 2e-16	< 2e-16
68	GRAN2	22	0.78	1.25e-06	7.35e-06
69	SVM Neutrophils	22	0.98	3.13e-10	4.51e-09
70	BROWN_MYELOID_CELL_DEVELOPMENT_UP	22	0.68	0.000123	0.000424
71	SENGUPTA_EBNA1_ANTICORRELATED	23	0.69	5.24e-05	0.000199
72	REACTOME_MRNA_SPLICING_MINOR_PATHWAY	24	0.69	0.022348	0.039617
73	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	24	0.87	8.54e-07	5.39e-06
74	MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL	24	0.84	7.12e-05	0.000258
75	MIPS_55S_RIBOSOME_MITOCHONDRIAL	24	0.92	7.64e-10	9.93e-09
76	LI_DCP2_BOUND_MRNA	24	0.81	1.69e-06	9.74e-06
77	YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_17	24	0.72	6.60e-07	4.25e-06
78	YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_13	24	0.83	1.16e-12	2.47e-11
79	MOREAUX_B_LYMPHOCYTE_MATURATION_BY_TACI_DN	24	0.74	0.000426	0.001338
80	LU_EZH2_TARGETS_UP	24	0.78	3.78e-14	1.09e-12
81	CHICAS_RB1_TARGETS_LOW_SERUM	24	0.72	0.000530	0.001584
82	STARK_PREFRONTAL_CORTEX_22Q11_DELETION_DN	24	0.67	5.77e-10	7.66e-09
83	KEGG_DNA_REPLICATION	26	0.93	1.01e-05	4.60e-05
84	MIPS_CENP_A_NAC_CAD_COMPLEX	26	0.87	0.026876	0.046282
85	REACTOME_EXTENSION_OF_TELOMERES	26	0.97	4.99e-05	0.000194
86	DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP	26	0.81	< 2e-16	< 2e-16
87	WHITFIELD_CELL_CYCLE_G1_S	26	0.67	0.000468	0.001444
88	EGUCHI_CELL_CYCLE_RB1_TARGETS	26	0.96	0.000336	0.001076
89	FRASOR_RESPONSE_TO_SERM_OR_FULVESTRANT_DN	26	0.77	0.001589	0.004038
90	ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_24HR	26	0.76	1.79e-06	1.02e-05
91	HAHTOLA_MYCOSIS_FUNGOIDES_CD4_UP	27	0.83	1.40e-05	6.17e-05
92	ZWANG_CLASS_3_TRANSIENTLY_INDUCED_BY_EGF	27	0.66	9.52e-05	0.000336
93	PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_DN	27	0.87	1.03e-12	2.29e-11
94	BURTON_ADIPOGENESIS_PEAK_AT_2HR	27	0.78	0.000937	0.002547
95	TIAN_TNF_SIGNALING_NOT_VIA_NFKB	27	0.93	0.000736	0.002068
96	BURTON_ADIPOGENESIS_1	27	0.93	5.26e-05	0.000199
97	ZHENG_FOXP3_TARGETS_IN_T_LYMPHOCYTE_DN	28	0.75	0.006550	0.013919
98	DAZARD_RESPONSE_TO_UV_NHEK_DN	28	0.66	1.83e-06	1.03e-05
99	NKcell-control	30	0.90	7.00e-07	4.46e-06
100	NKA2	30	0.99	4.01e-12	8.05e-11
101	NKA3	30	0.84	4.64e-05	0.000183
102	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	30	0.71	0.000125	0.000424
103	SVM NK cells resting	30	0.90	1.35e-07	1.03e-06
104	SVM NK cells activated	30	0.86	5.08e-07	3.36e-06
105	HAHTOLA_SEZARY_SYNDROM_DN	30	0.94	2.97e-06	1.58e-05
106	CHAN_INTERFERON_PRODUCING_DENDRITIC_CELL	30	0.89	0.016039	0.029704
107	NKcell-control	31	0.66	0.026138	0.045129
108	NKA2	31	0.87	2.91e-07	2.03e-06
109	TCELLA1	31	0.86	1.94e-05	8.39e-05
110	Monocyte-Day0	34	0.84	6.53e-15	2.28e-13
111	MONO2	34	0.86	1.45e-10	2.24e-09
112	SVM Monocytes	34	0.94	6.72e-08	5.57e-07
113	ROSS_AML_WITH_CBFB_MYH11_FUSION	34	0.77	0.000808	0.002242
114	YAGI_AML_RELAPSE_PROGNOSIS	34	0.72	0.012359	0.023820
115	KAMIKUBO_MYELOID_CEBPA_NETWORK	34	0.96	6.15e-05	0.000229
116	VALK_AML_CLUSTER_5	34	0.94	2.68e-05	0.000111
117	Bcell-naïve	35	0.90	1.18e-09	1.45e-08
118	BASO1	35	0.99	0.000524	0.001573
119	BCELLA1	35	0.92	3.21e-11	5.75e-10
120	KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	35	0.77	5.69e-05	0.000214
121	SVM B cells naive	35	0.93	1.70e-08	1.66e-07
122	MORI_PLASMA_CELL_DN	35	0.97	4.22e-06	2.15e-05
123	ZHAN_MULTIPLE_MYELOMA_CD1_VS_CD2_DN	35	0.74	0.002417	0.005913
124	SHIN_B_CELL_LYMPHOMA_CLUSTER_9	35	0.83	0.012998	0.024763
125	HADDAD_B_LYMPHOCYTE_PROGENITOR	35	0.65	2.21e-05	9.40e-05
126	KLEIN_PRIMARY_EFFUSION_LYMPHOMA_DN	35	0.85	3.42e-06	1.76e-05
127	BROWNE_INTERFERON_RESPONSIVE_GENES	37	0.89	8.34e-08	6.75e-07

128	BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	37	0.84	5.11e-07	3.36e-06
129	DAUER_STAT3_TARGETS_DN	37	0.75	0.001773	0.004420
130	TIAN_TNF_SIGNALING_VIA_NFKB	40	0.99	2.08e-05	8.96e-05
131	LINDSTEDT_DENDRITIC_CELL_MATURATION_B	40	0.78	0.000509	0.001541
132	GILMORE_CORE_NFKB_PATHWAY	40	0.99	0.004679	0.010480
133	SEKI_INFLAMMATORY_RESPONSE_LPS_UP	40	0.82	7.65e-06	3.60e-05
134	ERY3	43	0.61	0.000124	0.000424
135	REACTOME_SYNTHESIS_AND_INTERCONVERSION_OF_NUCLEOTIDE_DI_AND_TRIPHOSPHATES	43	0.87	0.007622	0.015891
136	VALK_AML_CLUSTER_7	43	0.83	0.002589	0.006286
137	MEGA2	44	0.80	6.35e-05	0.000234
138	RAGHAVACHARI_PLATELET_SPECIFIC_GENES	44	0.95	9.00e-12	1.71e-10
139	WIERENGA_STAT5A_TARGETS_DN	44	0.70	1.79e-07	1.32e-06
140	GNATENKO_PLATELET_SIGNATURE	44	0.95	2.08e-07	1.49e-06
141	TAVOR_CEBPA_TARGETS_DN	44	0.78	0.005430	0.011882
142	PlasmaCell-FromBoneMarrow	47	0.83	4.28e-10	5.92e-09
143	PlasmaCell-FromPBMC	47	0.93	1.54e-13	3.92e-12
144	REACTOME_UNFOLDED_PROTEIN_RESPONSE	47	0.76	0.000132	0.000441
145	REACTOME_ASPARAGINE_N_LINKED_GLYCOSYLATION	47	0.92	2.93e-10	4.32e-09
146	SVM Plasma cells	47	0.79	0.000923	0.002517
147	PELLICCIOTTA_HDAC_IN_ANTIGEN_PRESENTATION_UP	47	0.90	9.34e-08	7.37e-07
148	MORI_PLASMA_CELL_UP	47	0.97	6.21e-08	5.35e-07
149	PASQUALUCCI_LYMPHOMA_BY_GC_STAGE_UP	47	0.65	4.54e-05	0.000181
150	TARTE_PLASMA_CELL_VS_B_LYMPHOCYTE_UP	47	0.94	1.04e-10	1.69e-09
151	SHAFFER_IRF4_TARGETS_IN_ACTIVATED_DENDRITIC_CELL	47	0.83	6.72e-06	3.22e-05
152	LANDIS_ERBB2_BREAST_TUMORS_65_UP	47	0.77	0.022800	0.040310
153	SHAFFER_IRF4_TARGETS_IN_PLASMA_CELL_VS_MATURE_B_LYMPHOCYTE	47	0.82	6.75e-06	3.22e-05
154	SVM Eosinophils	48	0.87	4.29e-06	2.17e-05
155	TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_8D_DN	48	0.59	0.022915	0.040405
156	NAKAJIMA_EOSINOPHIL	48	0.93	2.42e-05	0.000102
157	WANG_NEOPLASTIC_TRANSFORMATION_BY_CCND1_MYC	49	0.90	0.015334	0.028720
158	REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	50	0.79	1.49e-06	8.68e-06
159	MIPS_EIF3_COMPLEX	50	1.00	0.005006	0.011137
160	MIPS_60S_RIBOSOMAL_SUBUNIT_CYTOPLASMIC	50	0.87	2.18e-05	9.31e-05
161	REACTOME_TRANSLATION	50	0.82	1.68e-10	2.52e-09
162	BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES	50	0.75	0.000507	0.001541
163	HOLLEMAN_ASPARAGINASE_RESISTANCE_B_ALL_UP	50	0.91	0.000387	0.001233
164	IRITANI_MAD1_TARGETS_DN	50	0.74	0.003536	0.008313
165	BILANGES_RAPAMYCIN_SENSITIVE_VIA_TSC1_AND_TSC2	50	0.82	1.09e-05	4.90e-05
166	CD8Tcell-N0	52	0.91	2.64e-05	0.000111
167	TCELLA2	52	0.88	2.73e-07	1.93e-06
168	CD4Tcell-Th1-restimulated48hour	54	0.90	4.96e-06	2.45e-05
169	REACTOME_KINESINS	54	0.80	0.009502	0.019097
170	REACTOME_G1_S_SPECIFIC_TRANSCRIPTION	54	1.00	0.000417	0.001318
171	REACTOME_CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION	54	0.96	0.000893	0.002457
172	DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP	54	0.81	$<$ 2e-16	$<$ 2e-16
173	CHANG_CYCLING_GENES	54	0.89	6.31e-16	2.32e-14
174	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA2_UP	54	0.84	2.28e-06	1.24e-05
175	MOLENAAR_TARGETS_OF_CCND1_AND_CDK4_DN	54	0.90	3.94e-07	2.72e-06
176	MORI_PRE_BI_LYMPHOCYTE_UP	54	0.82	9.64e-07	5.97e-06
177	LEE_EARLY_T_LYMPHOCYTE_UP	54	0.90	7.42e-12	1.45e-10
178	KUMAMOTO_RESPONSE_TO_NUTLIN_3A_DN	54	1.00	0.003259	0.007744
179	MORI_IMMATURE_B_LYMPHOCYTE_DN	54	0.87	1.03e-09	1.29e-08
180	ISHIDA_E2F_TARGETS	54	0.97	4.46e-09	5.02e-08
181	ZHAN_MULTIPLE_MYELOMA_PR_UP	54	0.99	3.10e-08	2.97e-07
182	ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_6HR	54	0.85	3.24e-08	3.07e-07
183	GAVIN_FOXP3_TARGETS_CLUSTER_P6	54	0.76	8.17e-06	3.79e-05
184	SHEPARD_BMYB_TARGETS	54	0.82	1.79e-05	7.81e-05
185	GENTLES_LEUKEMIC_STEM_CELL_DN	54	0.86	0.007927	0.016423
186	Neutrophil-Resting	56	0.66	1.89e-06	1.06e-05
187	BOSCO_ALLERGEN_INDUCED_TH2_ASSOCIATED_MODULE	58	0.72	5.98e-06	2.94e-05
188	Bcell-Memory_IgG_IgA	59	0.94	5.64e-10	7.63e-09
189	Bcell-Memory_IgM	59	0.93	1.34e-09	1.62e-08
190	BCELLA2	59	0.93	5.30e-09	5.86e-08
191	SVM B cells memory	59	0.92	4.15e-08	3.67e-07
192	ZHAN_MULTIPLE_MYELOMA_CD1_VS_CD2_DN	59	0.78	0.000765	0.002131
193	ZHAN_MULTIPLE_MYELOMA_CD1_DN	59	0.81	0.000403	0.001279
194	ERY4	61	0.80	$<$ 2e-16	$<$ 2e-16
195	ERY5	61	0.84	$<$ 2e-16	$<$ 2e-16
196	BIOCARTA_AHSP_PATHWAY	61	0.99	0.004183	0.009564
197	STEINER_ERYTHROCYTE_MEMBRANE_GENES	61	0.99	0.000589	0.001727

198	LIN_NPAS4_TARGETS_DN	61	0.70	0.008358	0.017156
199	VALK_AML_CLUSTER_8	61	0.99	1.26e-05	5.63e-05
200	JAATINEN_HEMATOPOIETIC_STEM_CELL_DN	61	0.68	3.21e-06	1.67e-05
201	VALK_AML_CLUSTER_7	61	0.99	1.34e-05	5.92e-05
202	Neutrophil-Resting	62	0.63	6.40e-05	0.000234
203	MARTENS_BOUND_BY_PML_RARA_FUSION	64	0.63	1.98e-06	1.11e-05
204	PARK_TRETINOIN_RESPONSE	64	0.93	0.012289	0.023754
205	PARK_TRETINOIN_RESPONSE_AND_PML_RARA_FUSION	64	0.83	0.001350	0.003524
206	PHONG_TNF_TARGETS_UP	65	0.67	0.018848	0.034154
207	NKcell-control	67	0.77	0.000703	0.001991
208	ERY2	67	0.80	4.70e-05	0.000184
209	MONO1	67	0.70	0.000945	0.002548
210	SVM NK cells resting	67	0.81	5.20e-05	0.000199
211	SVM Mast cells resting	67	0.71	0.016494	0.030461
212	LIAN_NEUTROPHIL_GRANULE_CONSTITUENTS	67	0.84	0.004349	0.009875
213	KAMIKUBO_MYELOID_CEBPA_NETWORK	67	0.78	0.008010	0.016493
214	VALK_AML_WITH_CEBPA	67	0.80	0.001112	0.002937
215	MARTINELLI_IMMATURE_NEUTROPHIL_UP	67	1.00	0.003628	0.008499
216	Neutrophil-Resting	68	0.58	0.012507	0.024036
217	MIPS_12S_U11_SNRNP	69	0.82	0.018854	0.034154
218	KEGG_SPLICEOSOME	69	0.65	0.002186	0.005368
219	LI_DCP2_BOUND_MRNA	69	0.70	0.000953	0.002558
220	DENDA1	70	0.88	1.03e-06	6.24e-06
221	RICKMAN_METASTASIS_DN	71	0.66	3.44e-05	0.000141
222	SENGUPTA_EBNA1_ANTICORRELATED	71	0.64	0.003105	0.007405
223	LU_EZH2_TARGETS_UP	71	0.65	4.60e-05	0.000183
224	KEGG_CHRONIC_MYELOID_LEUKEMIA	74	0.72	0.001646	0.004164
225	GSE19182_Ifng	76	0.77	1.63e-07	1.21e-06
226	MOSERLE_IFNA_RESPONSE	76	1.00	2.43e-06	1.31e-05
227	STAMBOLSKY_TARGETS_OF_MUTATED_TP53_DN	76	0.74	0.003465	0.008175
228	ZHAN_MULTIPLE_MYELOMA_LB_DN	76	0.80	0.001590	0.004038
229	GRANDVAUX_IFN_RESPONSE_NOT_VIA_IRF3	76	1.00	0.004046	0.009314
230	ROETH_TERT_TARGETS_UP	76	0.93	0.011907	0.023219
231	HECKER_IFNB1_TARGETS	76	0.83	5.00e-08	4.36e-07
232	BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS	76	0.95	6.76e-05	0.000246
233	BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	76	0.86	6.72e-08	5.57e-07
234	DAUER_STAT3_TARGETS_DN	76	0.84	2.69e-05	0.000111
235	TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_3D_UP	76	0.75	7.29e-08	5.97e-07
236	NIELSEN_SYNOVIAL_SARCOMA_DN	76	0.87	0.007593	0.015881
237	Neutrophil-Resting	77	0.81	< 2e-16	< 2e-16
238	GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_MAGENTA_UP	77	0.82	0.003661	0.008545
239	THEILGAARD_NEUTROPHIL_AT_SKIN_WOUND_DN	77	0.71	6.56e-08	5.57e-07
240	REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G	78	0.76	0.001936	0.004789
241	MIPS_26S_PROTEASOME	78	0.77	0.021513	0.038342
242	MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL	78	0.70	0.010317	0.020479
243	CD4Tcell-Th1-restimulated12hour	80	0.72	0.023316	0.040920
244	TCELLA4	80	0.90	1.05e-05	4.74e-05
245	BIOCARTA_NO2IL12_PATHWAY	80	0.99	0.000710	0.002004
246	BIOCARTA_IL12_PATHWAY	80	0.78	0.017357	0.031878
247	SVM T cells CD8	80	0.93	4.31e-09	4.93e-08
248	SVM T cells follicular helper	80	0.73	0.001821	0.004522
249	SVM T cells gamma delta	80	0.97	1.80e-09	2.13e-08
250	SVM NK cells resting	80	0.93	4.07e-08	3.67e-07
251	ONO_AML1_TARGETS_UP	80	0.81	0.010994	0.021629
252	PlasmaCell-FromPBMC	81	0.63	0.012751	0.024362
253	HSC1	82	0.91	8.10e-11	1.34e-09
254	HSC3	82	0.79	9.72e-07	5.97e-06
255	MEGA1	82	0.99	0.000600	0.001753
256	JAATINEN_HEMATOPOIETIC_STEM_CELL_UP	82	0.82	< 2e-16	< 2e-16
257	Neutrophil-Resting	83	0.69	1.63e-08	1.61e-07
258	CHUNG_BLISTER_CYTOTOXICITY_DN	83	0.81	0.000514	0.001550
259	EOS2	85	0.77	0.011951	0.023236
260	KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	85	0.67	0.014054	0.026546
261	SVM Mast cells resting	85	0.74	0.008669	0.017740
262	SVM Mast cells activated	85	0.81	0.001769	0.004420
263	NAKAJIMA_MAST_CELL	85	0.76	0.002983	0.007165
264	VANDESLUIS_COMMD1_TARGETS_GROUP_4_UP	87	0.95	0.005857	0.012566
265	CHIBA_RESPONSE_TO_TSA_DN	87	0.77	0.017507	0.032020
266	KORKOLA_CORRELATED_WITH_POU5F1	87	0.95	3.46e-05	0.000141
267	LI_WILMS_TUMOR	87	0.98	0.003963	0.009187
268	KORKOLA_EMBRYONIC_CARCINOMA_VS_SEMINOMA_UP	87	0.98	0.000503	0.001538
269	BHATTACHARYA_EMBRYONIC_STEM_CELL	87	0.87	8.83e-09	9.15e-08
270	BENPORATH_ES_1	87	0.82	< 2e-16	< 2e-16
271	REACTOME_CHOLESTEROL_BIOSYNTHESIS	88	0.81	0.009440	0.019097
272	KEGG_STEROID_BIOSYNTHESIS	88	1.00	0.000624	0.001814

273	SHAFFER_IRF4_TARGETS_IN_ACTIVATED_B_LYMPHOCYTE	88	0.76	0.000108	0.000378
274	HORTON_SREBF_TARGETS	88	0.99	0.000125	0.000424
275	SCHMIDT_POR_TARGETS_IN_LIMB_BUD_UP	88	1.00	9.57e-05	0.000336
276	ERY3	89	0.67	8.73e-09	9.15e-08
277	VALK_AML_CLUSTER_7	89	0.78	0.009563	0.019097
278	MIPS_SNF2H_COHESIN_NURD_COMPLEX	90	0.85	0.013406	0.025465
279	CHEN_HOXA5_TARGETS_9HR_UP	90	0.60	0.006872	0.014511
280	HAMAI_APOPTOSIS_VIA_TRAIL_UP	90	0.77	< 2e-16	< 2e-16
281	ZHANG_TLX_TARGETS_36HR_DN	90	0.71	3.07e-06	1.61e-05
282	DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	90	0.64	5.58e-11	9.74e-10
283	Neutrophil-Resting	92	0.61	0.001435	0.003687
284	MIZUSHIMA_AUTOPHAGOSOME_FORMATION	92	0.82	0.019573	0.035263
285	MIPS_28S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL	93	0.93	6.36e-05	0.000234
286	KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	93	0.82	0.000429	0.001340
287	MIPS_55S_RIBOSOME_MITOCHONDRIAL	93	0.94	3.97e-10	5.60e-09
288	YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_11	93	0.77	8.10e-06	3.78e-05
289	KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	93	0.71	< 2e-16	< 2e-16
290	MANALO_HYPOXIA_DN	93	0.81	< 2e-16	2.11e-15
291	SCHLOSSER_MYC_TARGETS_AND_SERUM_RESPONSE_UP	93	0.94	6.06e-07	3.94e-06
292	ERY3	94	0.89	< 2e-16	< 2e-16
293	ERY4	94	0.93	< 2e-16	< 2e-16
294	ERY5	94	0.90	< 2e-16	< 2e-16
295	IVANOVA_HEMATOPOIESIS_MATURE_CELL	94	0.68	2.05e-07	1.48e-06
296	KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	96	0.70	0.000479	0.001471
297	SVM T cells regulatory (Tregs)	96	0.82	8.17e-05	0.000293
298	SVM T cells gamma delta	96	0.74	0.001402	0.003617
299	REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	97	0.75	6.87e-14	1.82e-12
300	KEGG_OXIDATIVE_PHOSPHORYLATION	102	0.67	0.002565	0.006251
301	MIPS_SPLICEOSOME	102	0.61	0.017531	0.032020
302	MIPS_28S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL	104	0.71	0.024479	0.042486
303	REACTOME_FORMATION_OF_ATP_BY_CHEMIOSMOTIC_COUPLING	104	0.98	0.005241	0.011544
304	REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT	104	0.75	6.12e-06	2.98e-05
305	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS_	104	0.67	0.007717	0.016040
306	MIPS_55S_RIBOSOME_MITOCHONDRIAL	104	0.68	0.004315	0.009830
307	KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	105	0.69	0.022285	0.039612
308	BIOCARTA_PROTEASOME_PATHWAY	105	0.89	0.000701	0.001991
309	REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY_CCT_TRIC	105	0.87	0.009529	0.019097
310	MIPS_SPLICEOSOME	105	0.70	3.81e-05	0.000154
311	KEGG_LYSOSOME	108	0.79	1.35e-07	1.03e-06
312	KEGG_OTHER_GLYCAN_DEGRADATION	108	1.00	0.004558	0.010278
313	REACTOME_N_GLYCAN_TRIMMING_IN_THE_ER_AND_CALNEXIN_CALRETICULIN_CYCLE	108	0.89	0.020439	0.036624
314	REACTOME_POST_TRANSLATIONAL_PROTEIN_MODIFICATION	108	0.78	1.15e-08	1.15e-07
315	KEGG_N_GLYCAN_BIOSYNTHESIS	108	0.77	0.001674	0.004204
316	SVM Neutrophils	108	0.80	6.02e-05	0.000225
317	ZHAN_V1_LATE_DIFFERENTIATION_GENES_UP	108	0.83	0.001330	0.003484
318	MILI_PSEUDOPODIA_HAPTOTAXIS_DN	108	0.75	< 2e-16	< 2e-16
319	APPEL_IMATINIB_RESPONSE	108	0.99	4.67e-06	2.34e-05
320	VALK_AML_CLUSTER_5	108	0.73	0.020406	0.036624
321	BILANGES_RAPAMYCIN_SENSITIVE_VIA_TSC1_AND_TSC2	112	0.76	0.000270	0.000876
322	DEN_INTERACT_WITH_LCA5	112	0.74	0.023330	0.040920
323	GUTIERREZ_CHRONIC_LYMPHOCYTIC_LEUKEMIA_DN	113	0.86	7.16e-06	3.39e-05
324	BILBAN_B_CLL_LPL_UP	113	0.73	0.001128	0.002967
325	KLEIN_PRIMARY_EFFUSION_LYMPHOMA_DN	113	0.86	3.28e-06	1.70e-05
326	DendriticCell-Control	119	0.65	7.20e-05	0.000260
327	DENDA2	119	0.72	0.006058	0.012956
328	KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION	119	0.76	0.002786	0.006717
329	KEGG_ASTHMA	119	0.78	0.015870	0.029555
330	SVM Dendritic cells resting	119	0.97	1.16e-06	6.93e-06
331	REACTOME_AMYLOIDS	120	0.73	0.007973	0.016468
332	REACTOME_MEIOTIC_RECOMBINATION	120	0.80	0.000297	0.000955
333	REACTOME_RNA_POL_I_PROMOTER_OPENING	120	0.98	4.76e-06	2.37e-05
334	REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION	120	0.66	0.004675	0.010480
335	KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	120	0.81	9.17e-07	5.74e-06
336	MIPS_BAF_COMPLEX	121	0.98	0.006587	0.013952
337	MAYBURD_RESPONSE_TO_L663536_DN	121	0.70	0.009148	0.018604
338	DISTECHE_ESCAPED_FROM_X_INACTIVATION	122	1.00	0.004092	0.009388
339	BURTON_ADIPOGENESIS_11	123	0.73	0.003896	0.009062

340	MILI_PSEUDOPODIA_HAPTOTAXIS_UP	123	0.65	4.15e-08	3.67e-07
341	HAMAI_APOPTOSIS_VIA_TRAIL_UP	123	0.74	< 2e-16	< 2e-16
342	PUIFFE_INVASION_INHIBITED_BY_ASCITES_DN	123	0.64	0.003988	0.009213
343	CD4Tcell-N0	126	0.96	2.45e-06	1.31e-05
344	MemoryTcell-RO-unactivated	126	0.93	9.60e-06	4.39e-05
345	TCELLA6	126	0.91	1.49e-07	1.12e-06
346	SVM T cells CD4 naive	126	0.72	0.004427	0.010018
347	SVM T cells CD4 memory resting	126	0.88	4.88e-07	3.27e-06
348	SVM T cells regulatory (Tregs)	126	0.90	1.22e-06	7.19e-06
349	ZHENG_FOXP3_TARGETS_IN_T_LYMPHOCYTE_DN	126	0.83	0.000581	0.001711
350	LEE_EARLY_T_LYMPHOCYTE_DN	126	0.75	0.000823	0.002275
351	LEE_NAIVE_T_LYMPHOCYTE	126	0.98	0.000900	0.002465
352	JAATINEN_HEMATOPOIETIC_STEM_CELL_DN	126	0.64	0.000189	0.000625
353	Neutrophil-Resting	130	0.70	7.88e-09	8.43e-08
354	PID_IL8CXCR2_PATHWAY	130	0.85	0.000631	0.001826
355	BIOCARTA_SALMONELLA_PATHWAY	130	0.88	0.023951	0.041788
356	PID_P38ALPHABETAPATHWAY	130	0.89	0.000634	0.001828
357	BIOCARTA_RAB_PATHWAY	130	0.94	0.009527	0.019097
358	BIOCARTA_CDC42RAC_PATHWAY	130	0.99	0.000700	0.001991
359	CHUNG_BLISTER_CYTOTOXICITY_DN	130	0.82	0.000439	0.001366
360	GAZDA_DIAMOND_BLACKFAN_ANEMIA_PROGENITOR_UP	130	0.76	0.005349	0.011742
361	THEILGAARD_NEUTROPHIL_AT_SKIN_WOUND_DN	130	0.78	1.59e-12	3.29e-11
362	REACTOME_ACTIVATED_AMPK_STIMULATES_FATTY_ACID_OXIDATION_IN_MUSCLE	131	0.84	0.014514	0.027338
363	MEGA2	133	0.79	0.000292	0.000943
364	RAGHAVACHARI_PLATELET_SPECIFIC_GENES	133	0.76	0.000114	0.000394
365	WIERENGA_STAT5A_TARGETS_DN	133	0.63	0.001028	0.002736
366	REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	134	0.60	0.002727	0.006599
367	PYEON_HPV_POSITIVE_TUMORS_UP	136	0.71	0.000548	0.001630
368	HOEBEKE_LYMPHOID_STEM_CELL_UP	136	0.64	0.012576	0.024099
369	REACTOME_PEPTIDE_CHAIN_ELONGATION	137	0.99	2.00e-14	6.63e-13
370	KEGG_RIBOSOME	137	1.00	3.08e-14	9.29e-13
371	MIPS_40S_RIBOSOMAL_SUBUNIT_CYTOPLASMIC	137	1.00	2.26e-06	1.24e-05
372	MIPS_TRBP_CONTAINING_COMPLEX_1	137	1.00	9.17e-05	0.000325
373	MIPS_NOP56P_ASSOCIATED_PRE_RRNA_COMPLEX	137	0.73	4.79e-05	0.000187
374	MIPS_RIBOSOME_CYTOPLASMIC	137	1.00	1.67e-13	4.10e-12
375	MIPS_60S_RIBOSOMAL_SUBUNIT_CYTOPLASMIC	137	1.00	9.61e-09	9.80e-08
376	REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION	137	0.93	3.88e-13	9.18e-12
377	BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES	137	0.98	2.33e-11	4.29e-10
378	HOLLEMAN_ASPARAGINASE_RESISTANCE_B_ALL_UP	137	0.93	0.000131	0.000440
379	PECE_MAMMARY_STEM_CELL_UP	137	0.88	6.57e-13	1.50e-11
380	BILANGES_SERUM_RESPONSE_TRANSLATION	137	0.93	3.21e-05	0.000132
381	CHNG_MULTIPLE_MYELOMA_HYPERPLOID_UP	137	0.89	2.23e-06	1.23e-05
382	BURTON_ADIPOGENESIS_12	138	0.74	0.015020	0.028210
383	ZHANG_TLX_TARGETS_36HR_DN	138	0.65	0.000563	0.001667
384	REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY_CCT_TRIC	140	0.83	0.017619	0.032092
385	VISALA_RESPONSE_TO_HEAT_SHOCK_AND_AGING_DN	140	1.00	0.004787	0.010686
386	FU_INTERACT_WITH_ALKBH8	140	0.99	0.005614	0.012203

Table 10: A complete list of LV-geneset associations.

B.2 DataRemix

B.2.1 Fourier transform of the exponential kernel: equation 3.5

The original paper which proposed to use Fourier features to construct an approximate feature map [129] lists analytic formulations for three popular kernels, which are Gaussian kernel, Laplacian kernel and Cauchy kernel. Here we provide the detailed derivation for the Fourier transform of the exponential kernel in the three-dimensional space.

$$p(\omega) = \frac{1}{(2\pi)^3} \int \exp(-i\vec{\omega}^T \vec{\Delta}) \exp(-\frac{\|\vec{\Delta}\|_2}{2}) d\vec{\Delta}$$

First, we take the substitution $w = \|\vec{\omega}\|_2$ and $r = \|\vec{\Delta}\|_2$. We assume that $\vec{\omega}$ is parallel to the polar direction.

$$\begin{aligned}
p(\omega) &= \frac{1}{(2\pi)^3} \int \exp(-i\vec{\omega}^T \vec{\Delta}) \exp(-\frac{\|\vec{\Delta}\|_2}{2}) d\vec{\Delta} \\
&= \frac{1}{(2\pi)^3} \int_0^\infty \int_0^{2\pi} \int_0^\pi \exp(-iwr\cos\theta) \exp(-\frac{r}{2}) \cdot r^2 \sin\theta dr d\theta d\phi \\
&= \frac{1}{(2\pi)^3} \int_0^\infty r^2 \exp(-\frac{r}{2}) dr \int_0^{2\pi} d\phi \int_0^\pi \exp(-iwr\cos\theta) \cdot \sin\theta d\theta \\
&= \frac{1}{(2\pi)^2} \int_0^\infty r^2 \exp(-\frac{r}{2}) dr \int_0^\pi \exp(-iwr\cos\theta) \cdot \sin\theta d\theta
\end{aligned}$$

Here we make the substitution $z = \cos\theta$. Thus $\sin\theta d\theta = -dz$.

$$p(\omega) = \frac{1}{(2\pi)^2} \int_0^\infty r^2 \exp(-\frac{r}{2}) dr \int_1^{-1} -\exp(-iwrz) dz$$

Make another substitution $t = -iwrz$, where $dz = -\frac{1}{iwr} dt$

$$\begin{aligned}
p(\omega) &= \frac{1}{(2\pi)^2} \int_0^\infty r^2 \exp(-\frac{r}{2}) dr \int_1^{-1} -\exp(-iwrz) dz \\
&= \frac{1}{(2\pi)^2} \int_0^\infty r^2 \exp(-\frac{r}{2}) dr \int_{-iwr}^{iwr} \exp(t) \cdot \frac{1}{iwr} dt \\
&= \frac{1}{(2\pi)^2} \int_0^\infty r \exp(-\frac{r}{2}) dr \int_{-iwr}^{iwr} \exp(t) \cdot \frac{1}{iw} dt
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^2 \cdot iw} \int_0^\infty r \exp(-\frac{r}{2}) dr \int_{-iwr}^{iwr} \exp(t) dt \\
&= \frac{2}{(2\pi)^2 \cdot iw} \int_0^\infty r \exp(-\frac{r}{2}) dr \cdot i \sin(wr) \\
&= \frac{2}{(2\pi)^2 w} \int_0^\infty r \exp(-\frac{r}{2}) \sin(wr) dr \\
&= \frac{2}{(2\pi)^2 w} \cdot \frac{-4w \exp(-\frac{r}{2})(4w^2 r + r + 4)\cos(wr)}{(4w^2+1)^2} \Bigg|_0^\infty \\
&= \frac{2}{(2\pi)^2 w} \cdot \frac{16w}{(4w^2+1)^2} \\
&= \frac{8}{\pi^2 (4w^2+1)^2} \\
&= \frac{8}{\pi^2 (4\|\vec{\omega}\|_2^2+1)^2}
\end{aligned}
$$

B.2.2 Posterior sampling: equation 3.7

The formula for sampling from the posterior can be intuitively understood in terms of the parallels with linear regression. Since in the feature space, any sample $f(\lambda)$ from Gaussian Process can be approximated by $\Phi(\lambda)^T\theta$, we can think of this as a simple linear regression: substitute $f(\lambda) = y - \epsilon, \epsilon \sim N(0, \sigma^2)$ and wish to solve $y = \Phi(\lambda)^T\theta + \epsilon, \epsilon \sim N(0, \sigma^2)$ for θ. If θ comes with a Gaussian prior $N(0, I)$, then the posterior distribution of θ is $N(A^{-1}\Phi(\vec{\lambda})\vec{y}, \sigma^2 A^{-1})$, where $A = \Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + \sigma^2 I$. The mean value $(\Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + \sigma^2 I)^{-1}\Phi(\vec{\lambda})\vec{y}$ is the same as the ridge regression estimator or MAP (Maximum a posteriori estimation) estimator of θ. For a full formal derivation see bellow:

$$
P(\theta|\vec{\lambda}, \vec{y}) \propto P(\vec{y}|\vec{\lambda}, \theta)P(\theta)
$$

where $P(\vec{y}|\vec{\lambda},\theta) \sim N(\Phi(\vec{\lambda})^T\theta, \sigma^2 I), P(\vec{\theta}) \sim N(0, I)$

$$
\begin{aligned}
P(\theta|\vec{\lambda},\vec{y}) &\propto \frac{1}{\sqrt{(2\pi)^t}\sigma^t} \exp\left(-\frac{1}{2}(\vec{y}-\Phi(\lambda)^T\theta)^T \cdot \sigma^{-2} I \cdot (\vec{y}-\phi(\lambda)^T\theta)\right) \cdot \frac{1}{\sqrt{(2\pi)^t}} \exp\left(-\frac{1}{2}\theta^T\theta\right) \\
&\propto \exp\left(-\frac{1}{2}(\vec{y}-\Phi(\lambda)^T\theta)^T \cdot \sigma^{-2} I \cdot (\vec{y}-\Phi(\lambda)^T\theta)\right) \cdot \exp\left(-\frac{1}{2}\theta^T\theta\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\vec{y}^T\vec{y} - 2\vec{y}^T\Phi(\lambda)^T\theta + \theta^T\Phi(\lambda)\Phi(\lambda)^T\theta\right] - \frac{1}{2}\theta^T\theta\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left[\vec{y}^T\vec{y} - 2\vec{y}^T\Phi(\lambda)^T\theta + \theta^T\Phi(\lambda)\Phi(\lambda)^T\theta\right] - \frac{1}{2}\theta^T\theta\right) \cdot \text{constant} \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\vec{y}^T\vec{y} - 2\vec{y}^T\Phi(\lambda)^T\theta + \theta^T\Phi(\lambda)\Phi(\lambda)^T\theta\right] - \frac{1}{2}\theta^T\theta\right) \\
&\cdot \exp\left(-\frac{1}{2\sigma^2}\vec{y}^T\left[\frac{1}{\sigma^2}\Phi(\lambda)^T(\frac{1}{\sigma^2}\Phi(\lambda)\Phi(\lambda)^T - I)^{-1}\Phi(\lambda) - I\right]\vec{y}\right) \\
&= \exp\left(-\frac{1}{2}(\theta-\hat{\theta})^T(\frac{1}{\sigma^2}\Phi(\lambda)\Phi(\lambda)^T + I)(\theta-\hat{\theta})\right)
\end{aligned}
$$

where $\hat{\theta} = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\Phi(\lambda)\Phi(\lambda)^T + I)^{-1}\Phi(\lambda)\vec{y}$. Thus,

$$
\begin{aligned}
P(\theta|\vec{\lambda},\vec{y}) &\propto \exp\left(-\frac{1}{2}(\theta-\hat{\theta})^T B(\theta-\hat{\theta})\right) \\
&\propto N(\frac{1}{\sigma^2}B^{-1}\Phi(\vec{\lambda})\vec{y}, B^{-1})
\end{aligned}
$$

where $B = \frac{1}{\sigma^2}\Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + I$. Here, B is a simple transformation of the previously defined quantity $A = \Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + \sigma^2 I$ with $B = \frac{1}{\sigma^2}A$. Equivalently, we get,

$$P(\theta|\vec{\lambda},\vec{y}) \propto N(A^{-1}\Phi(\vec{\lambda})\vec{y}, \sigma^2 A^{-1})$$

where $A = \Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + \sigma^2 I$.

B.3 NIFA

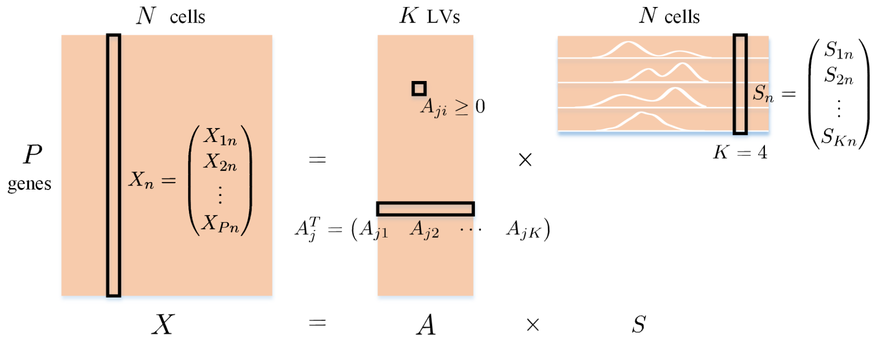

N cells
K LVs
N cells
P
genes
$X_n = \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{Pn} \end{pmatrix}$
$=$
$A_{ji} \geq 0$
$A_j^T = (A_{j1} \quad A_{j2} \quad \cdots \quad A_{jK})$
$\times$
$S_n = \begin{pmatrix} S_{1n} \\ S_{2n} \\ \vdots \\ S_{Kn} \end{pmatrix}$
$K = 4$
X
$=$
A
$\times$
S

Figure 52: A schematic representation of the NIFA model.

Each latent variable (each row of S) is associated with M component distribution. And we assume a Gaussian noise model.

$$X = AS + noise \tag{B.1}$$

The noise model is set to be simple Gaussian $N(0, \Sigma^{-1})$ where $\Sigma^{-1} = \begin{pmatrix} \beta & & & \\ & \beta & & \\ & & \ddots & \\ & & & \beta \end{pmatrix}$.

B.3.1 Joint likelihood

$$P = \prod_{n=1}^{N} P(X_n | A, S_n, \beta) P(S_n | \epsilon, \mu, \sigma) \cdot \prod_{j=1}^{P} \prod_{i=1}^{K} P(A_{ji} | \eta_{ji}, \lambda_i) \prod_{i,m,n} P(\epsilon_{imn}) \cdot \prod_{i,m} P(\mu_{im}) P(\sigma_{im}) \cdot P(\beta)$$

$$\begin{aligned} & P(X_n | A, S_n, \beta) \\ =& |\det(2\pi\Sigma)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (X_n - AS_n)^T \Sigma^{-1} (X_n - AS_n) \right) \\ =& (2\pi)^{-\frac{P}{2}} \beta^{\frac{P}{2}} \exp\left(-\frac{\beta}{2} \sum_{j=1}^{P} (X_{jn} - A_j^T S_n)^2 \right) \end{aligned} \tag{B.2}$$

X represents the RNA-seq matrix with dimension P-by-N, A denotes loading matrix with dimension P-by-K, S stands for latent variables with dimension K-by-N and β is defined by the noise model.

$$P(S_n|\epsilon, \mu, \sigma) = \prod_{i=1}^{K} \prod_{m=1}^{M} N(S_{in}|\mu_{im}, \sigma_{im})^{\epsilon_{imn}} \tag{B.3}$$

Each latent variable is associated with M component distributions (Gaussian distribution). Each component distribution comes with μ_{im} as the mean and σ_{im} as the inverse of the variance. ϵ_{imn} is a set of binary latent variables, and $\sum_{m=1}^{M} \epsilon_{imn} = 1$. If S_{in} is generated from component j, then $\epsilon_{imn} = 1$ if $m = j$ and $\epsilon_{imn} = 1$ if $m \neq j$.

$$\begin{aligned} &P(A_{ji}|\eta_{ji}, \lambda_i) \\ =&(2\pi)^{-\frac{1}{2}}(\lambda_i)^{\frac{1}{2}} \exp(-\frac{1}{2}\lambda_i(A_{ji} - \eta_{ji})^2) \cdot \frac{1}{1 - \Phi(-\eta_{ji}\lambda_i^{\frac{1}{2}})} \end{aligned} \tag{B.4}$$

The loading matrix A is modelled with a truncated normal prior with $a = 0$ and $b = \infty$. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

$$P(\epsilon_{imn}) = \pi_{im}^{\epsilon_{imn}} \tag{B.5}$$

The latent variables ϵ_{imn} are governed by the mixing proportion π_{im}.

$$P(\mu_{im}) = N(\mu_{im}|\rho_{im}, \phi_{im}) \tag{B.6}$$

μ_{im} comes with a Gaussian distribution as prior with ρ_{im} as the mean and ϕ_{im} as the inverse of the variance.

$$P(\sigma_{im}) = Gamma(\sigma_{im}|a_{\sigma_{im}}, b_{\sigma_{im}}) \tag{B.7}$$

σ_{im} comes with a Gamma distribution as prior with $a_{\sigma_{im}}$ and $b_{\sigma_{im}}$ as hyper parameters.

$$P(\beta) = Gamma(\beta|a_\beta, b_\beta) \tag{B.8}$$

β comes with a Gamma distribution as prior with a_β and b_β as hyper parameters.

B.3.2 Variational update functions

$$Q = \prod_{n=1}^{N} q(S_n) \cdot \prod_{j,i} q(A_{ji}) \cdot \prod_{i,m,n} q(\epsilon_{imn}) \cdot \prod_{i,m} q(\mu_{im}) q(\sigma_{im}) \cdot q(\beta)$$

The derivations are detailed in the appendix. Thus,

$$q(S_n) = N(S_n|\mu^*_m, \Sigma^*_m) \tag{B.9}$$

where

$$\begin{cases} \mu^*_m = \Sigma^*_m \left[\sum_{m=1}^{M} B_m^{-1} \mu_m + < \beta > \sum_{j=1}^{P} X_{jn} < A_j > \right] \\ \Sigma^*_m = \left[\sum_{m=1}^{M} B_m^{-1} + < \beta > \sum_{j=1}^{P} < A_j A_j^T > \right]^{-1} \end{cases}$$

$$\mu_m = \begin{pmatrix} < \mu_{1m} > \\ < \mu_{2m} > \\ \vdots \\ < \mu_{Km} > \end{pmatrix}, B_m^{-1} = \begin{pmatrix} < \epsilon_{1mn} >< \sigma_{1m} > & & \\ & \ddots & \\ & & < \epsilon_{Kmn} >< \sigma_{Km} > \end{pmatrix}$$

$$q(A_{ji}) = f(A_{ji}|\eta^*_{ji}, \lambda^*_i, 0, \infty) \tag{B.10}$$

which is the truncated normal distribution with $a = 0$ and $b = \infty$. where

$$\begin{cases} \lambda^*_i = \lambda_i + < \beta > \sum_{n=1}^{N} < S^2_{in} > \\ \eta^*_{ji} = \frac{\lambda_i \eta_{ji} - <\beta> \left[\sum_{n=1}^{N} <S_{in}> (\sum_{m \neq i} <A_{jm}><S_{mn}> - X_{jn}) \right]}{\lambda_i + <\beta> \sum_{n=1}^{N} <S^2_{in}>} \end{cases}$$

$$q(\epsilon_{imn}) = \widetilde{\lambda_{imn}}^{\epsilon_{imn}} \tag{B.11}$$

where

$$\widetilde{\lambda_{imn}} = \frac{e^{\lambda_{imn}}}{\sum_{m=1}^{M} e^{\lambda_{imn}}}$$
$$\lambda_{imn} = -\frac{1}{2} \log(2\pi) + \frac{1}{2} < \log \sigma_{im} > + \log \pi_{im} - \frac{1}{2} < \sigma_{im} > \left[< S^2_{in} > -2 < S_{in} >< \mu_{im} > + < \mu^2_{im} > \right]$$

$$q(\mu_{im}) = N(\mu_{im}|\rho^*_{im}, \phi^*_{im}) \tag{B.12}$$

where

$$\begin{cases} \rho^*_{im} = (\phi^*_{im} < \sigma_{im} >)^{-1} \left[< \sigma_{im} > \sum_{n=1}^{N} < \epsilon_{imn} >< S_{in} > + \phi_{im}\rho_{im} < \sigma_{im} > \right] \\ \phi^*_{im} = \phi_{im} + \sum_{n=1}^{N} < \epsilon_{imn} > \end{cases}$$

$$q(\sigma_{im}) = Gamma(\sigma_{im} | a^*_{\sigma_{im}}, b^*_{\sigma_{im}}) \tag{B.13}$$

where

$$\begin{cases} a^*_{\sigma_{im}} = a_{\sigma_{im}} + \frac{1}{2}\sum_{n=1}^{N} < \epsilon_{imn} > + \frac{1}{2} \\ b^*_{\sigma_{im}} = b_{\sigma_{im}} + \dfrac{1}{2}\displaystyle\sum_{n=1}^{N} < \epsilon_{imn} > (< S^2_{in} > -2 < S_{in} >< \mu_{im} > + < \mu^2_{im} >) \\ \qquad + \dfrac{1}{2} < \phi_{im} > (< \mu^2_{im} > -2 < \mu_{im} > \rho_{im} + \rho^2_{im}) \end{cases}$$

$$q(\beta) = Gamma(\beta | a^*_\beta, b^*_\beta) \tag{B.14}$$

where

$$\begin{cases} a^*_\beta = a_\beta + \frac{NP}{2} \\ b^*_\beta = b_\beta + \frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{P}[X^2_{in} - 2X_{in} < A^T_j >< S_n > + Tr(< A_j A^T_j >< S_n S^T_n >)] \end{cases}$$

$$\pi_{im} = \frac{1}{N}\sum_{n=1}^{N} \widetilde{\lambda_{imn}} \tag{B.15}$$

B.3.2.1 Update functions of μ_j and Σ_j

$$\delta_i = \sqrt{(\lambda^*_i)^{-1}}$$

$$Z_{ji} = 1 - \Phi(-\frac{\eta^*_{ji}}{\delta_i})$$

$$\mu_j \quad \Leftarrow \quad \mu_{ji} = \eta^*_{ji} + \frac{\phi(-\frac{\eta^*_{ji}}{\delta_i})}{Z_{ji}}\delta_i \tag{B.16}$$

$$\Sigma_j \quad \Leftarrow \quad \Sigma^{ii}_j = \delta^2 \left[1 + \frac{-\frac{\eta^*_{ji}}{\delta_i}\phi(-\frac{\eta^*_{ji}}{\delta_i})}{Z} - \left(\frac{\phi(-\frac{\eta^*_{ji}}{\delta_i})}{Z} \right)^2 \right] \tag{B.17}$$

where Σ_j is a diagonal matrix and Σ^{ii}_j stands for the ith diagonal values of Σ_j.

B.3.2.2 Update functions of expectations

$$< S_n >= \mu_m^* \tag{B.18}$$

$$< S_n S_n^T >= \mu_m^* {\mu_m^*}^T + \Sigma_m^* \tag{B.19}$$

$$< \epsilon_{imn} >= \widetilde{\lambda_{imn}} \tag{B.20}$$

$$< \sigma_{im} >= \frac{a_{\sigma_{im}}^*}{b_{\sigma_{im}}^*} \tag{B.21}$$

$$< A_j >= \mu_j \tag{B.22}$$

$$< A_j A_j^T >= \mu_j \mu_j^T + \Sigma_j \tag{B.23}$$

$$< \log \sigma_{im} >= \psi(a_{\sigma_{im}}^*) - \ln(b_{\sigma_{im}}^*) \tag{B.24}$$

$$< \mu_{im} >= \rho_{im}^* \tag{B.25}$$

$$< \mu_{im}^2 >= (\phi_{im}^* \sigma_{im}^*)^{-1} + (\rho_{im}^*)^2 \tag{B.26}$$

$$< \beta >= \frac{a_\beta^*}{b_\beta^*} \tag{B.27}$$

B.3.2.3 Derivation for eq B.10 The loading matrix A is modelled with a truncated normal prior with $a = 0$ and $b = \infty$. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

$$
\begin{aligned}
&\log P(A_{ji}|\eta_{ji}, \lambda_i) \\
&= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\lambda_i - \frac{1}{2}\lambda_i(A_{ji} - \eta_{ji})^2 - \log(1 - \Phi(-\eta_{ji}\lambda_i^{\frac{1}{2}})) \\
\xRightarrow{\text{remove constant}} & -\frac{1}{2}\lambda_i(A_{ji}^2 - 2A_{ji}\eta_{ji})
\end{aligned}
$$

$$
\begin{aligned}
&\prod_{n=1}^{N}(2\pi)^{-\frac{P}{2}}\beta^{\frac{P}{2}}\exp\left(-\frac{\beta}{2}\sum_{j=1}^{P}(X_{jn} - A_j^T S_n)^2\right)\cdot P(A_{ji}|\eta_{ji}, \lambda_i) \\
&= \prod_{n=1}^{N}(2\pi)^{-\frac{P}{2}}\beta^{\frac{P}{2}}\exp\left(-\frac{\beta}{2}\sum_{j=1}^{P}(A_{ji}S_{in} + \sum_{m\neq i}A_{jm}S_{mn} - X_{jn})^2\right)\cdot P(A_{ji}|\eta_{ji}, \lambda_i) \\
\xRightarrow{\log} & \sum_{n=1}^{N}\left[-\frac{P}{2}\log(2\pi) + \frac{P}{2}\log\beta - \frac{\beta}{2}\sum_{j=1}^{P}(A_{ji}S_{in} + \sum_{m\neq i}A_{jm}S_{mn} - X_{jn})^2\right] \\
& + \log P(A_{ji}|\eta_{ji}, \lambda_i) \\
\xRightarrow{\text{remove constant}} & \sum_{n=1}^{N}\left[-\frac{\beta}{2}(A_{ji}S_{in} + \sum_{m\neq i}A_{jm}S_{mn} - X_{jn})^2\right] + \log P(A_{ji}|\eta_{ji}, \lambda_i) \\
\xRightarrow{\text{remove constant}} & \sum_{n=1}^{N}\left[-\frac{\beta}{2}\left(A_{ji}^2 S_{in}^2 + 2A_{ji}S_{in}(\sum_{m\neq i}A_{jm}S_{mn} - X_{jn})\right)\right] + \log P(A_{ji}|\eta_{ji}, \lambda_i) \\
&= -\frac{\beta}{2}\left[A_{ji}^2\sum_{n=1}^{N}S_{in}^2 + 2A_{ji}\sum_{n=1}^{N}S_{in}(\sum_{m\neq i}A_{jm}S_{mn} - X_{jn})\right] - \frac{1}{2}\lambda_i(A_{ji}^2 - 2A_{ji}\eta_{ji}) \\
&= A_{ji}^2\left(-\frac{\beta}{2}\sum_{n=1}^{N}S_{in}^2 - \frac{1}{2}\lambda_i\right) + A_{ji}\left[-\beta\sum_{n=1}^{N}S_{in}(\sum_{m\neq i}A_{jm}S_{mn} - X_{jn}) + \eta_{ji}\lambda_i\right] \\
\xRightarrow{<\cdot>_Q} & A_{ji}^2(-\frac{<\beta>}{2}\sum_{n=1}^{N}<S_{in}>^2 - \frac{1}{2}\lambda_i) + A_{ji}\left[-<\beta>\sum_{n=1}^{N}<S_{in}>\right. \\
& \left.(\sum_{m\neq i}<A_{jm}><S_{mn}> - X_{jn}) + \eta_{ji}\lambda_i\right]
\end{aligned}
$$

Bibliography

[1] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7), 2009.

[2] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

[3] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[4] Genevera I Allen, Logan Grosenick, and Jonathan Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159, 2014.

[5] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[6] Orly Alter, Patrick O Brown, and David Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences*, 100(6):3351–3356, 2003.

[7] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[8] Carl A Anderson, Gabrielle Boucher, Charlie W Lees, Andre Franke, Mauro D'Amato, Kent D Taylor, James C Lee, Philippe Goyette, Marcin Imielinski, Anna Latiano, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics*, 43(3):246, 2011.

[9] Dvir Aran, Zicheng Hu, and Atul J Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18(1):220, 2017.

[10] Detlev Arendt, Jacob M Musser, Clare VH Baker, Aviv Bergman, Connie Cepko, Douglas H Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D Laubichler, et al. The origin and evolution of cell types. *Nature Reviews Genetics*, 17(12):744, 2016.

[11] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a probabilistic framework for comprehensive integration of structured single-cell data. *BioRxiv*, page 837104, 2019.

[12] Ricard Argelaguet, Stephen J Clark, Hisham Mohammed, L Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W Hanna, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, pages 1–5, 2019.

[13] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6), 2018.

[14] William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429, 2016.

[15] Kinjal Basu and Souvik Ghosh. Analysis of thompson sampling for gaussian process optimization in the bandit setting. *arXiv preprint arXiv:1705.06808*, 2017.

[16] Alexis Battle et al. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome Res*, 24(1):14–24, Jan 2014.

[17] Etienne Becht, Nicolas A Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, Wolf H Fridman, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology*, 17(1):218, 2016.

[18] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[19] Babraham Bioinformatics. Fastqc: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute*, 2011.

[20] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[21] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

[22] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015.

[23] Florian Buettner, Naruemon Pratanwanich, Davis J McCarthy, John C Marioni, and Oliver Stegle. f-sclvm: scalable and versatile factor analysis for single-cell rna-seq. *Genome biology*, 18(1):212, 2017.

[24] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[25] Kerstin Bunte, Eemeli Leppäaho, Inka Saarinen, and Samuel Kaski. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 32(16):2457–2463, 2016.

[26] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

[27] J Gray Camp, Keisuke Sekine, Tobias Gerber, Henry Loeffler-Wirth, Hans Binder, Malgorzata Gac, Sabina Kanton, Jorge Kageyama, Georg Damm, Daniel Seehofer, et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659):533–538, 2017.

[28] Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2), 2011.

[29] Xin Chen and Joost J Oppenheim. Resolving the identity myth: key markers of functional cd4+ foxp3+ regulatory t cells. *International immunopharmacology*, 11(10):1489–1496, 2011.

[30] Maria Chikina, Elena Zaslavsky, and Stuart C Sealfon. Cellcode: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*, 31(10):1584–1591, 2015.

[31] Hans Clevers, Susanne Rafelski, Michael Elowitz, Allon Klein, Jay Shendure, Cole Trapnell, Ed Lein, Emma Lundberg, Matthias Uhlen, Alfonso Martinez-Arias, et al. What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Systems*, 4(3):255–259, 2017.

[32] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[33] Jake Crawford and Casey S Greene. Incorporating biological structure into machine learning models in biomedicine. *Current Opinion in Biotechnology*, 63:126–134, 2020.

[34] Alan Dabney and John D. Storey. *qvalue: Q-value estimation for false discovery rate control*, 2014. R package version 1.38.0.

[35] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284, 2016.

[36] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

[37] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[38] Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using rna-seq. *Appl. Bioinformatics*, pages 1–67, 2015.

[39] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

[40] Jesse M Engreitz, Bernie J Daigle Jr, Jonathan J Marshall, and Russ B Altman. Independent component analysis: mining microarray data for fundamental human gene expression modules. *Journal of biomedical informatics*, 43(6):932–944, 2010.

[41] Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19(5):776–792, 2018.

[42] Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, 13(3):241, 2016.

[43] Anthony J Filiano, Yang Xu, Nicholas J Tustison, Rachel L Marsh, Wendy Baker, Igor Smirnov, Christopher C Overall, Sachin P Gadani, Stephen D Turner, Zhiping Weng, et al. Unexpected role of interferon-γ in regulating neuronal connectivity and social behaviour. *Nature*, 535(7612):425–429, 2016.

[44] Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118, 2010.

[45] Jerome H Friedman. Exploratory projection pursuit. *Journal of the American statistical association*, 82(397):249–266, 1987.

[46] Menachem Fromer, Panos Roussos, Solveig K Sieberts, Jessica S Johnson, David H Kavanagh, Thanneer M Perumal, Douglas M Ruderfer, Edwin C Oh, Aaron Topol, Hardik R Shah, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*, 19(11):1442, 2016.

[47] Johann A Gagnon-Bartsch, Laurent Jacob, and Terence P Speed. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112, 2013.

[48] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[49] Renaud Gaujoux. An introduction to nmf package. *URL: https://cran. r-project. org/package= NMF. R package version 0.20*, 6, 2018.

[50] Yingbin Ge, Rikka Azuma, Bethsebah Gekonge, Alfonso Lopez-Coral, Min Xiao, Gao Zhang, Xiaowei Xu, Luis J Montaner, Zhi Wei, Meenhard Herlyn, et al. Induction of metallothionein expression during monocyte to melanoma-associated macrophage differentiation. *Frontiers in biology*, 7(4):359–367, 2012.

[51] David Gerard and Matthew Stephens. Unifying and generalizing methods for removing unwanted variation based on negative controls. *arXiv preprint arXiv:1705.08393*, 2017.

[52] Christian Gieger, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, Ulrich Elling, Alison H Goodall, Yann Labrune, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, 2011.

[53] Ting Gong, Nicole Hartmann, Isaac S Kohane, Volker Brinkmann, Frank Staedtler, Martin Letzkus, Sandrine Bongiovanni, and Joseph D Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PloS one*, 6(11), 2011.

[54] Ting Gong and Joseph D Szustakowski. Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085, 2013.

[55] Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, 19(1):220, 2018.

[56] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[57] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.

[58] Kasper D Hansen, Rafael A Irizarry, and Zhijin Wu. Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, 2012.

[59] Traver Hart, H Kiyomi Komori, Sarah LaMere, Katie Podshivalova, and Daniel R Salomon. Finding the active genes in deep rna-seq gene expression studies. *BMC genomics*, 14(1):778, 2013.

[60] Tracy SP Heng, Michio W Painter, Kutlu Elpek, Veronika Lukacs-Kornek, Nora Mauermann, Shannon J Turley, Daphne Koller, Francis S Kim, Amy J Wagers, Natasha Asinovski, et al. The immunological genome project: networks of gene expression in immune cells. *Nature immunology*, 9(10):1091–1094, 2008.

[61] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

[62] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–1115, 2013.

[63] Victoria Hore, Ana Viñuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094, 2016.

[64] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26(3):304–319, 2016.

[65] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.

[66] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[67] Laurent Jacob, Johann A Gagnon-Bartsch, and Terence P Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, 2016.

[68] Rodolphe Jenatton et al. Structured sparse principal component analysis. *arXiv preprint arXiv:0909.1440*, 2009.

[69] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[70] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, Dec 2008.

[71] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[72] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[73] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483, 2017.

[74] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[75] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason CH Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell stem cell*, 17(4):471–485, 2015.

[76] Wei Kong, Charles R Vanderburg, Hiromi Gunshin, Jack T Rogers, and Xudong Huang. A review of independent component analysis application to microarray gene expression data. *Biotechniques*, 45(5):501–520, 2008.

[77] Monika S Kowalczyk, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka K Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research*, 25(12):1860–1872, 2015.

[78] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *Journal of proteome research*, 11(8):4120–4131, 2012.

[79] Kimberly R Kukurba and Stephen B Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015.

[80] Roshan M Kumar, Patrick Cahan, Alex K Shalek, Rahul Satija, A Jay DaleyKeyser, Hu Li, Jin Zhang, Keith Pardee, David Gennert, John J Trombetta, et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516(7529):56–61, 2014.

[81] James Larkin, Vanna Chiarion-Sileni, Rene Gonzalez, Jean Jacques Grob, C Lance Cowey, Christopher D Lao, Dirk Schadendorf, Reinhard Dummer, Michael Smylie, Piotr Rutkowski, et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *New England journal of medicine*, 373(1):23–34, 2015.

[82] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

[83] Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y Weiss-Solís, Robin Duque, Hugues Bersini, and Ann Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, 14(4):469–490, 2013.

[84] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[85] Jeffrey T Leek. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research*, 42(21):e161–e161, 2014.

[86] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[87] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[88] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9), 2007.

[89] Eemeli Leppäaho, Muhammad Ammad-ud din, and Samuel Kaski. Gfa: exploratory analysis of multiple data sources with group factor analysis. *The Journal of Machine Learning Research*, 18(1):1294–1298, 2017.

[90] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[91] Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, et al. De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular systems biology*, 15(2), 2019.

[92] Bo Li, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

[93] Bo Li, Eric Severson, Jean-Christophe Pignon, Haoquan Zhao, Taiwen Li, Jesse Novak, Peng Jiang, Hui Shen, Jon C Aster, Scott Rodig, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome biology*, 17(1):174, 2016.

[94] Huipeng Li, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, 49(5):708, 2017.

[95] Peipei Li, Yongjun Piao, Ho Sun Shon, and Keun Ho Ryu. Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, 16(1):347, 2015.

[96] Sheng Li, Paweł P Łabaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po-Yen Wu, May Wang, Charles Wang, et al. Detecting and correcting systematic variation in large-scale rna sequencing data. *Nature biotechnology*, 32(9):888, 2014.

[97] James C Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, 100(26):15522–15527, 2003.

[98] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[99] Yang Liao, Gordon K Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, 41(10):e108–e108, 2013.

[100] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[101] Jennifer Listgarten, Carl Kadie, Eric E. Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A*, 107(38):16465–16470, Sep 2010.

[102] Chang Liu, Maria Chikina, Rahul Deshpande, Ashley V Menk, Ting Wang, Tracy Tabib, Erin A Brunazzi, Kate M Vignali, Ming Sun, Donna B Stolz, et al. Treg cells promote the srebp1-dependent metabolic fitness of tumor-promoting macrophages via repression of cd8+ t cell-derived interferon-γ. *Immunity*, 51(2):381–397, 2019.

[103] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[104] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[105] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6), 2019.

[106] Mitchell J Machiela and Stephen J Chanock. Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21):3555–3557, 2015.

[107] Weiguang Mao, Ryan Hausler, and Maria Chikina. Dataremix: a universal data transformation for optimal inference from gene expression datasets. *bioRxiv*, page 357467, 2018.

[108] Weiguang Mao, Maziyar Baran Pouyan, Dennis Kostka, and Maria Chikina. Non-negative independent factor analysis for single cell rna-seq. *bioRxiv*, 2020.

[109] Weiguang Mao, Elena Zaslavsky, Boris M Hartmann, Stuart C Sealfon, and Maria Chikina. Pathway-level information extractor (plier) for gene expression data. *Nature Methods*, 16(7):607, 2019.

[110] JL Marchini, C Heaton, Maintainer Brian Ripley, and MASS Suggests. Package 'fastica', 2019.

[111] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[112] Naomi Moris, Cristina Pina, and Alfonso Martinez Arias. Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, 17(11):693, 2016.

[113] Samantha A Morris. The evolving concept of cell identity in the single cell era. *Development*, 146(12):dev169748, 2019.

[114] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621, 2008.

[115] S. Mostafavi et al. Type i interferon signaling genes in recurrent major depression: increased expression detected by whole-blood rna sequencing. *Mol Psychiatry*, 19(12):1267–1274, Dec 2014.

[116] Sara Mostafavi, Alexis Battle, Xiaowei Zhu, Alexander E Urban, Douglas Levinson, Stephen B Montgomery, and Daphne Koller. Normalizing rna-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One*, 8(7), 2013.

[117] Sonia Nestorowa, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology*, 128(8):e20–e31, 2016.

[118] Cancer Genome Atlas Research Network et al. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, May 2013.

[119] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[120] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

[121] Bernard Ng, William Casazza, Ellis Patrick, Shinya Tasaki, Gherman Novakovsky, Daniel Felsky, Yiyi Ma, David A Bennett, Chris Gaiteri, Philip L De Jager, et al. Using transcriptomic hidden variables to infer context-specific genotype effects in the brain. *The American Journal of Human Genetics*, 105(3):562–572, 2019.

[122] Noa Novershtern et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144(2):296–309, Jan 2011.

[123] Guillaume Obozinski, Laurent Jacob, et al. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[124] Andre Olsson, Meenakshi Venkatasubramanian, Viren K Chaudhri, Bruce J Aronow, Nathan Salomonis, Harinder Singh, and H Leighton Grimes. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, 537(7622):698–702, 2016.

[125] Lior Pachter. Models for transcript quantification from rna-seq. *arXiv preprint arXiv:1104.3889*, 2011.

[126] Lucia Peixoto, Davide Risso, Shane G Poplawski, Mathieu E Wimmer, Terence P Speed, Marcelo A Wood, and Ted Abel. How data analysis affects power, reproducibility and biological insight of rna-seq studies in complex datasets. *Nucleic acids research*, 43(16):7664–7674, 2015.

[127] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241, 2015.

[128] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[129] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[130] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome biology*, 14(9):3158, 2013.

[131] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[132] Hubert Rehrauer, Lennart Opitz, Ge Tan, Lina Sieverling, and Ralph Schlapbach. Blind spots of quantitative rna-seq: the limits for assessing abundance, differential expression, and isoform switching. *BMC bioinformatics*, 14(1):370, 2013.

[133] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896, 2014.

[134] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

[135] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480, 2011.

[136] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

[137] D. T. Ross et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–235, Mar 2000.

[138] Moshe Sade-Feldman, Keren Yizhak, Stacey L Bjorgaard, John P Ray, Carl G de Boer, Russell W Jenkins, David J Lieb, Jonathan H Chen, Dennie T Frederick, Michal Barzily-Rokni, et al. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175(4):998–1013, 2018.

[139] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

[140] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[141] Manfei Si and Jinghe Lang. The roles of metallothioneins in carcinogenesis. *Journal of hematology & oncology*, 11(1):107, 2018.

[142] Nikolai Slavov. Unpicking the proteome in single cells. *Science*, 367(6477):512–513, 2020.

[143] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[144] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[145] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[146] Sudeep Srivastava and Liang Chen. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic acids research*, 38(17):e170–e170, 2010.

[147] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[148] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010.

[149] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

[150] Genevieve L Stein-O'Brien, Raman Arora, Aedin C Culhane, Alexander V Favorov, Lana X Garmire, Casey S Greene, Loyal A Goff, Yifeng Li, Aloune Ngom, Michael F Ochs, et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics*, 34(10):790–805, 2018.

[151] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865, 2017.

[152] Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, 35(14):i436–i445, 2019.

[153] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[154] Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, and Casey S Greene. Multiplier: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell systems*, 8(5):380–394, 2019.

[155] Vésteinn Thorsson, David L Gibbs, Scott D Brown, Denise Wolf, Dante S Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F Gao, Christopher L Plaisier, James A Eddy, et al. The immune landscape of cancer. *Immunity*, 48(4):812–830, 2018.

[156] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498, 2015.

[157] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511, 2010.

[158] Severin Uebbing. Evaluation of rna-seq normalization methods using challenging datasets. *bioRxiv*, page 401679, 2018.

[159] Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145, 2015.

[160] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.

[161] Anne M Van der Leun, Daniela S Thommen, and Ton N Schumacher. Cd8+ t cell states in human cancer: insights from single-cell analysis. *Nature Reviews Cancer*, pages 1–15, 2020.

[162] CH Waddington. The strategy of the genes: a discussion of some aspects of theoretical biology. 1957.

[163] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, et al. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics*, 45(10):1238, 2013.

[164] Adam Williams, Charalampos G Spilianakis, and Richard A Flavell. Interchromosomal association and gene regulation in trans. *Trends in genetics*, 26(4):188–197, 2010.

[165] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[166] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–437, 2014.

[167] Bo Xia and Itai Yanai. A periodic table of cell types. *Development*, 146(12):dev169854, 2019.

[168] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4(1):1–11, 2013.

[169] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[170] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. *PLoS computational biology*, 14(6):e1006245, 2018.

[171] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2017.

[172] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.

[173] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.