bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pathway-Level Information ExtractoR (PLIER): a generative model for gene expression data

Weiguang Mao, Maria Chikina (email: mchikina@pitt.edu)

February 7, 2017

Abstract

Genome scale molecular datasets are often highly structured, with many correlated measurements. This general phenomenon can be related to the underlying data generating process. In assays of mixed cell populations, such as blood, variation in cell-type proportion induces a complex correlation structure at the gene-level. Likewise, groups of genes can be co-regulated/co-expressed through shared transcription factors and signaling pathways. Many applications of gene expression analysis rely on their ability to reflect these unobserved biological processes in order to draw mechanistic conclusions. On the other hand, correlated patterns of expression may also reflect nuisance factors, such as batch effects, which interfere with correct biological interpretation. The choice of analysis method is heavily dependent on which of these factors (nuisance or interesting-biological) is believed to account for more variation and the optimal variance analysis strategy remains an open question.

In this study we describe a method to infer a biologically grounded data generating model that provides estimates of underlying biological processes, including explicitly identified pathway-level and cell-type proportion effects. Specifically, we formulate a new matrix decomposition framework, PLIER (Pathway-level Information ExtractoR), that explicitly incorporates prior biological knowledge. Using simulations, we demonstrate the superiority of our method in recovering the true data generating model. Using real data, we show that our approach is able to recover interpretable biological variables, reproduce previous findings in a simplified framework, distinguish biological and technical variation, and provide additional biological insight. The PLIER method and auxiliary functions and data are compiled in the PLIER R package available at https://github.com/wgmao/PLIER.

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

One salient feature of high dimensional molecular data is the presence of groups of correlated measurements, i.e. data structure. In gene expression datasets correlation among genes may be the result of coordinated transcriptional regulation. In such cases, understanding the underlying pathway-level effects can improve statistical power and interpretation. On the other hand, a variety of technical factors, often referred to as "batch effects", are also reflected in the data structure (see Leek et al. [2010] for review). Thus, common variation in gene expression can be a nuisance that interferes with differential expression analysis but may also represent important biological processes. Methods to address these sources of variation have to balance decreasing noise against removing biological signals.

To take an illustrative example we consider the problem of transcriptional profiling in human blood, which is a complex and highly variable mixture of at minimum 20 transcriptionally unique cell-types [Novershtern et al., 2011]. In a typical blood dataset the first few principal components (PCs) produced by Principal Component Analysis (PCA) capture most cell-type composition and technical variation. However, cell-type composition heterogeneity in blood can itself be of interest, complicating the decisions regarding which variation should be removed. For example, in a recent study [Battle et al., 2014] of blood eQTLs, two different normalizations, with different numbers of latent components, were applied to optimize cis and trans eQTL discovery. This was motivated by the observation that biological variation in the dataset, such as cell-type proportion variation, yields significant *trans*-eQTL effects. However, these same effects become a nuisance variable when assessing *cis*-eQTLs.

Ideally, it should be possible to cleanly separate variation that is due to technical factors from that which has a biological origin, while also attributing the latter to specific biological processes. For example, in a recent paper we proposed a strategy that extracts latent variables that have a one-to-one correspondence to variation in specific cell-types. This allows for composition variation to first be analyzed directly and then statistically removed to identify transcriptionally mediated effects, thus retaining all biologically relevant signals. However, this kind of analysis is difficult to automate and our approach relied heavily on some intuition about which cell-types are detectable in specific datasets, which depends critically on the size of the dataset, sample preparation, and biological perturbation.

The problem of finding the correct data generating model has received considerable attention. The widely recognized drawback of PCA is that it produces necessarily orthogonal components that are linear combinations of all original variables. Thus the PCs do not in general align with the true latent variables that drive common variation. One approach is to partition the variation into biological and technical in a supervised fashion, that is based on additional information about the samples, such as the contrast of interest, or known technical factors (see Leek and Storey [2007], Stegle et al. [2010], Listgarten et al. [2010], Kang et al. [2008] and Mostafavi et al. [2013], which formulates a general framework for the variation partition approach). However, while these approaches can dramatically improve statistical power by removing technical variation, they do not provide a generative model for remaining biological variation and heavily depend on the availability and quality of the auxiliary sample data.

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

An alternative approach is to seek an entirely unsupervised model by imposing additional constraints on the matrix decomposition problem in order to identify biologically meaningful variance components [Mahoney et al., 2009, Srebro, 2004, Gillis, 2014]. For example, non-negative matrix factorization (NMF) is a popular choice for gene expression analysis [Brunet et al., 2004, Wang et al., 2013, Gillis, 2014]. Another approach is to introduce sparsity constraints so that only some genes have non-zero loadings (often referred to as sparse PCA) [Zou et al., 2006, Witten et al., 2009]. By lifting the highly restrictive orthogonality requirement and imposing biologically motivated structure (such as sparsity or positivity in the gene loadings), these methods will often naturally groups biologically related genes together producing latent components that align well with known biological factors (such as disease subtype). However, as there is no explicit requirement for the genes associated with each component be biologically related, this outcome is not in general guaranteed.

Importantly, while methods such as NMF and sparse PCA are indeed automated unsupervised approaches (that attempt to find a data generating model without any expert knowledge about the dataset) they are also completely general and can be applied to any high dimensional dataset. On the other hand, an approach specifically designed for gene expression does not have to be completely agnostic; it can incorporate information about gene identities as long as this information is generic and not dataset specific.

In this work, we formulate a alternative matrix decomposition method that explicitly relates the data structure to prior information in the form of pathways and biologically related genesets representing biological pathways (e.g. Kanehisa et al. [2016]), sets of tissue- or cell-type specific markers (e.g. [Abbas et al., 2009]), and coordinated transcriptional responses observed in genome wide experiments (e.g. [Subramanian et al., 2005]). Besides the matrix decomposition, our method returns information that indicates if and how each variance component is associated with the prior information, thus providing an additional dimension of biological interpretability without any further analysis.

Using simulations and real data we demonstrate that our approach can effectively use prior information to achieve superior performance at recovering the data generating model and providing biological insight. Our method is computationally efficient, running in just a few minutes on large clinical datasets, is robust to technical noise, and is readily applicable to other high-dimensional datasets.

2 Methods

Problem Setting

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, B minimizing

$$||Y - ZB||_F^2 \tag{1}$$

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

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

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

singular value decomposition (SVD) of Y. In an SVD based decomposition rows of B are referred to as principle components (PCs). Since PCs are necessarily orthogonal, which we do not require, we will use the term latent variables (LVs).

In order to improve the interpretability of the low dimensional representation, 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 notations above, we form the updated decomposition problem based on the original formulation. We wish to find U, Z, B minimizing

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

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

The first term of the optimization is the same and minimizes the overall reconstruction error. The second term specifies that Z should be "close to" sparse combinations of genesets represented by C and the third terms introduce 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.

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 can set several default parameters that perform well in a range of situations. For example, we find that k should be set to the number of statistically significant PCs. We provide a function `num.pc` in our package that determines the correct k based on permutation following the approach proposed in [Leek et al., 2007]. A good choice for λ_1 and

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. 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 optimum value of constant that controls the sparsity of U is highly dataset dependent; however we found that we can set it heuristically so that the fraction of latent components that have a non-zero U coefficients is approximately controlled. To do this we solve the the ridge regression problem

$$\text{argmin}_U ||Z^{(0)} - CU||_F^2 + 5||U||_F$$

at the first iteration and set λ_3 to the median of the maximum values taken columnwise. This approach is motivated by the observation that the lasso problems can be solved by iteratively thresholding ridge regression coefficients. In our experiments with real data this approach indeed resulted in approximately half of the latent variables being associated with prior information. We use these heuristics for all the analysis in the manuscript and set them as the default behavior for our method.

Pseudo 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 pseudo cross validation procedure that proceeds as follows: we create a new loadings matrix Z' that represents a 4/5 of the genes and recompute new latent variables B'. We then use B' to compute the loadings for the held-out 1/5 of the genes. After completing this for the entire set of genes we can compute the AUC (and the p-value) for each non-zero element of U. While it is also possible to perform true cross validation by running PLIER on a subsets of the raw data this has a significant computational cost and we find that the pseudo cross validation procedure is in good agreement with the true cross validation results. Pseudo cross validation results are provided by default in the output.

Comparison with similar methods

There are several methods that can take prior information about genesets into account in order to a biologically meaningful low dimensional representation. Examples, include Bayesian Factor Analysis [Bunte et al., 2016] that extracts pathway-level latent variables and our previously proposed method CellCODE [Chikina et al., 2015] that estimates cell proportion variation from cell-type marker genesets. However, these methods require that the genesets be specified *a priori* and that gene 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 gene-

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sets 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 these have drawbacks. For example, a related method termed structured sparse PCA [Jenatton et al., 2009] 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 [Obozinski et al., 2011]. However, the implementation is computationally expensive and in its current form only applies to regression without being readily adaptable to matrix decomposition.

Data

The Depression Genen 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 [Mostafavi et al., 2014]. The NIMH database contains several normalized versions of this data and for our study we used "trans" normalized data as described in [Battle et al., 2014].

The generic blood cell-type marker dataset was derived from the IRIS (Immune Response In Silico) [Abbas et al., 2009] and DMAP (Differentiation Map) datasets [Novershtern et al., 2011] 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. We also included cell-type markers from a recent publication Newman et al. [2015] 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 from mSigDB, and a set of transcriptional signatures relevant to immune signaling described in [Filiano et al., 2016]. The entire prior information dataset with 1513 pathways is included in the PLIER package.

3 Results

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

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the matrix decomposition to align with prior knowledge.

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

$$Y = ZB + E. \tag{3}$$

With both Z and $B > 0$. B is drawn from Beta distribution and each column sums to one by design. 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.

Our basic evaluation strategy is based on computing the maximal correlations between simulated and recovered latent variables, for the purpose of comparisons with other methods we use the absolute value so as to allow factors with reversed sign. Figure 1 depicts the results of multiple simulation runs process with four decomposition methods: PLIER, PLIER with no prior information (which can be accomplished by setting λ_3 to a high value), NMF[Brunet et al., 2004] and SPC [Witten et al., 2009]. NMF is a popular decomposition method that is free of hyperparameters (though different matrix norms can be used) however it requires positive data

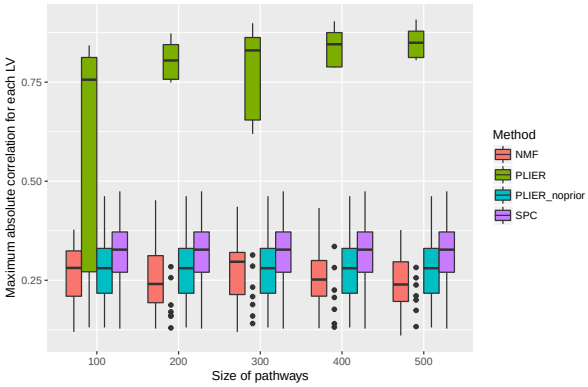

Figure 1: Data is simulated according to the NMF model (see text for details). Boxplots of the correlation between simulated LVs and those recovered by various decomposition methods. We compare PLIER against two other methods, NMF [Brunet et al., 2004] and SPC [Witten et al., 2009], as well as PLIER 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.

as input. SPC is another popular method that can enforce sparsity and positivity, it has one hyperparameter that we set by cross-validation as described in the original paper [Witten et al., 2009]. 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

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

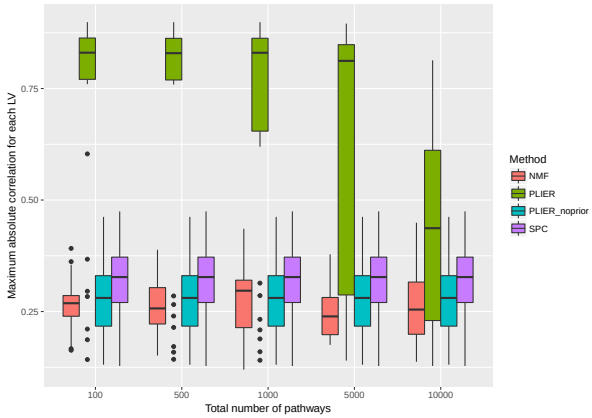

Figure 2: Data is simulated the same as in 1 except that the number of genes per pathway is kept at 300 and the number of uninformative pathways is varied. As the prior information gets noisy PLIER performance approaches that of other constrained decomposition methods

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 (Figure 2).

Real Data Examples

Analysis of large whole blood dataset

To demonstrate the utility of our approach on real data we apply it to the Depression Gene Network (DGN) dataset which consists of 922 whole blood gene expression profiles quantified by RNAseq. One of the goals of the DGN study was to identify gene expression quantitative trait loci (eQTLs). Two groups of eQTLs are typically distinguished: locally acting *cis*-eQTLs that affect a nearby gene, and *trans*-eQTLs that affect gene located far from the polymorphism, possibly on a different chromosome. By definition *cis*-eQTLs affect only a few genes, typically just one, by altering a nearby regulatory sequence while *trans*-eQTLs may be mediated at the pathway-level. The basic model for

Table 1: The *trans*-eQTL "success rate" (defined as the number of eQTLs found significant at an adjusted p-value of 0.05 multiplied by the number of SNPs (651,075)) among different sets of quantitative traits. We consider single gene expression, PLIER latent variables, and a set of latent variables restricted to those that used some prior information.

QT	num eQTLs	num QTs	success rate
PLIER ("with prior" LVs only)	40	62	0.65
PLIER (all LVs)	47	108	0.43
gene-level	1005	15231	0.066

such pathway-level effects is that a locally acting *cis*-eQTL alters the expression for a gene with a regulatory role (such as a transcription factor) which in turn alters the expression of downstream targets, giving rise to several *trans*-eQTLS. Empirically, many SNPs that affect gene expression in *trans* affect multiple genes and in many cases a single *cis* effect, which presumably affects the upstream regulatory factor, can also be found. The converse is not true, however: most *cis*-eQTLs do not give rise to *trans*-eQTLs and indeed *trans*-eQTLs are orders of magnitude less frequent.

Blood datasets are particularly challenging as blood composition varies dramatically across individuals and the composition effect can be considered a nuisance or an interesting-biological variables depending on the question asked. Some cell-type variation has a genetic basis which may manifest as *trans*-eQTLs of cell-type specific genes but the same variation becomes a nuisance variable when evaluating *cis*-eQTLs. The initial analysis of this datasets [Battle et al., 2014] in fact applied two different normalizations, with different numbers of latent components, in order to maximize *cis* and *trans* -eQTLs discovery.

While a dimensionality reducing decomposition cannot capture *cis*-eQTLs it can in principle be used to improve the discovery of pathway-level *trans*-eQTLs if it succeeds

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

at effectively isolating specific pathway-level effects. In the context of a PLIER decomposition we hypothesize that it is possible to identify *trans*-eQTLs by testing the relationship between genotype and PLIER latent variables. In essence, we simply define a new quantitative trait, instead of using the expression of single genes directly we use the estimated pathway-level effect.

We apply PLIER decomposition to the DGN dataset using a set of prior information genesets that includes cell-type markers, "canonical" pathways from mSigDB, and a set of transcriptional signatures relevant to immune signaling described in [Filiano et al., 2016] (see Methods). We produce a 108 dimensional decomposition (see Methods for choosing the decomposition dimension) of which 62 used some prior information. Overall 163 pathways out of 1513 were associated with some LV.

Using these decomposition results we observe that indeed many LVs have a significant association with genotype. For those LVs that have a genotype association and also a high confidence pathway association we visualize the corresponding entries in the decomposition matrix U (which specify how the LVs and prior information genesets are related) in Figure 3. In fact, we find that if LVs are used directly as quantitative traits we can greatly increase the frequency of eQTLs that pass the significance threshold (see Table 1). We note that our approach to multiple hypothesis testing is to apply Bonefrroni correction to each quantitative trait, the same approach used in the original study, and therefore the "success rates" of gene-level and pathway-level eQTLs are directly comparable since the correction does not depended on the number of traits.

Besides improving the rate of eQTL discovery the PLIER decomposition can be used to infer the biological nature of the latent variable which in many cases leads to improved interpretation. For example we can deduce the that the one SNP that yields the largest number of significant *trans*-eQTLs in a gene-level analysis is in-fact associated with a single latent variable, that we infer to represent megakaryocyte/platelet lineage cell-type variation (megakaryocytes are the precursors to platelets). The SNP in question, rs1354034, has been previously associated with platelet volume [Gieger et al., 2011].

We also detect several new associations that were not found in the gene-level analysis. One example of particular interest is a significant association between rs3184504 and an Interferon gamma transcriptional signature. The genes with the highest loadings for this latent variable include canonical interferon-gamma regulated genes (GBP1, STAT1, and TAP1 are among the top 10). This associated SNP is a missense variant in SH2B3 which is a known regulator of the Interferon-gamma pathway [Mori et al., 2014]. While recent genome-wide association studies have revealed a link between polymorphism in this gene and autoimmune diseases, including type 1 diabetes and celiac disease [Hunt et al., 2008, Smyth et al., 2008], to our knowledge a direct link to interferon gamma in human blood is a novel finding.

Technical variation invariance

A key motivation for PLIER is to tease apart technical and biological variation and the hypothesis is that those 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

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

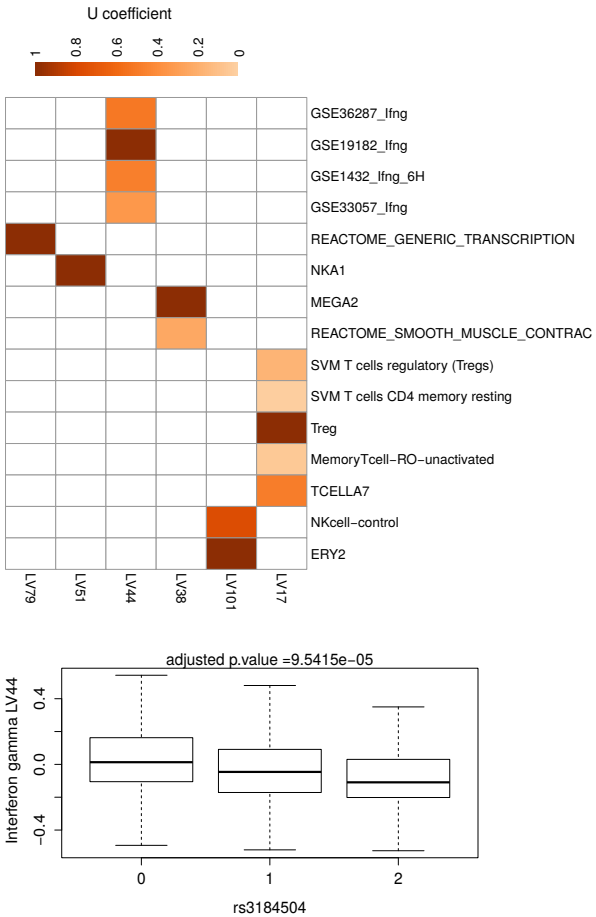

Figure 3: **Top** A subset of the prior information coefficient matrix U computed from the DGN dataset. We focus on the latent variables that both had a significant eQTL association and a prior information geneset association with an AUC of 0.75 (see Methods for cross validation approach). **Bottom** A new significant eQTL that affects the interferon gamma pathway extracted as latent variable LV44.

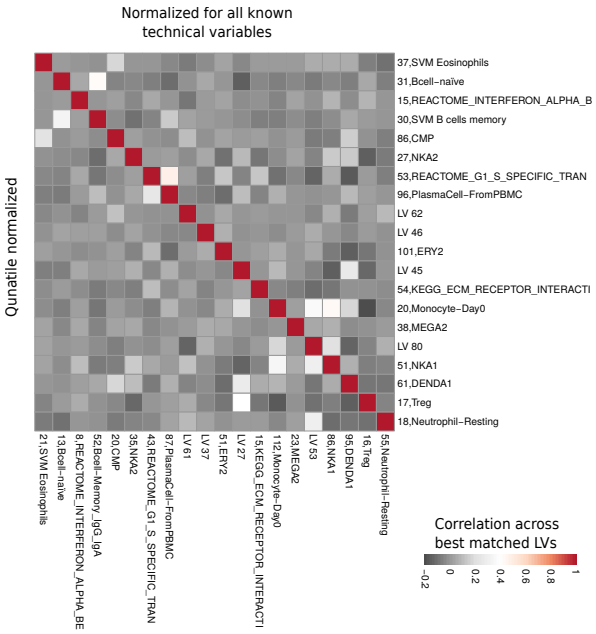

Figure 4: Heatmap of the correlations among LVs from decompositions performed on two different versions of the same data, one normalized for all technical variables and another normalized with quantile normalization. The heatmap shows all pairwise correlations for the top 20 best matched LVs named with their corresponding prior information (if any). Note that the prior information used is almost identical across the two decompositions.

hypothesis by applying PLIER to differently normalized data. In the preceding section we had used a dataset that was normalized for 20 technical variables which reflected information about data collection and RNAseq quality control. We can also apply PLIER to the "naive-normalized" version of data represented by logged counts normalized by quantile normalization. Obtaining two different decompositions we find that many LVs are indeed in one-to-one correspondence and that these are biased towards LVs that made use of prior information (Figure 5). When the matching LVs use prior information we find that the corresponding genesets are likewise either the same or closely related (Figure 4).

4 Discussion

We present a new method Pathway-level Information ExtractoR, PLIER, which incorporates prior information into matrix decomposition of molecular data, yielding a biologically grounded data model. The method can improve the interpretability of gene expression datasets by providing a correspondence between data structure and biologically coherent gene groups. Evaluating the method on real data we show that it is able to find biological latent variables, reproduce previ-

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

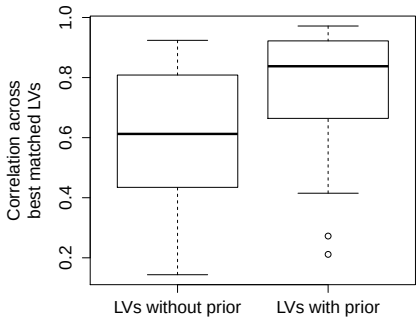

Figure 5: Correlation distributions across all best matched pairs of LVs. LVs that use prior information are more robust to normalization procedure as they are more consistent across differently normalized datasets.

ous findings by recasting *trans*-eQTLs in a latent variable framework and provide credible novel predictions.

References

Alexander R Abbas et al. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*, 4(7): e6098, 2009. doi: 10.1371/journal.pone.0006098. URL `http://dx.doi.org/10.1371/journal.pone.0006098`.

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. doi: 10.1101/gr.155192.113. URL `http://dx.doi.org/10.1101/gr.155192.113`.

Jean-Philippe Brunet et al. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164–4169, Mar 2004. doi: 10.1073/pnas.0308531101. URL `http://dx.doi.org/10.1073/pnas.0308531101`.

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.

Maria Chikina, Elena Zaslavsky, and Stuart C Sealfon. Cellcode: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*, page btv015, 2015.

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.

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.

Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12:257, 2014.

Karen A Hunt, Alexandra Zhernakova, Graham Turner, Graham AR Heap, Lude Franke, Marcel Bruinenberg, Jihane Romanos, Lotte C Dinesen, Anthony W Ryan, Davinder Panesar, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nature genetics*, 40(4):395–402, 2008.

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

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1): D457–D462, Jan 2016. doi: 10.1093/nar/gkv1070. URL `http://dx.doi.org/10.1093/nar/gkv1070`.

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. doi: 10.1534/genetics.108.094201. URL `http://dx.doi.org/10.1534/genetics.108.094201`.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–1735, Sep 2007. doi: 10.1371/journal.pgen.0030161. URL `http://dx.doi.org/10.1371/journal.pgen.0030161`.

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. *Nat Rev Genet*, 11 (10):733–739, Oct 2010. doi: 10.1038/nrg2825. URL `http://dx.doi.org/10.1038/nrg2825`.

Jeffrey T Leek et al. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 2007.

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. doi: 10.1073/pnas.1002425107. URL `http://dx.doi.org/10.1073/pnas.1002425107`.

Michael W Mahoney et al. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

Taizo Mori, Yukiko Iwasaki, Yoichi Seki, Masanori Iseki, Hiroko Katayama, Kazuhiko Yamamoto, Kiyoshi Takatsu, and Satoshi Takaki. Lnk/sh2b3 controls the production and function of dendritic cells and regulates the induction of ifn-γ–producing t cells. *The Journal of Immunology*, 193(4):1728–1736, 2014.

bioRxiv preprint doi: https://doi.org/10.1101/116061; this version posted March 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. doi: 10.1038/mp.2013.161. URL `http://dx.doi.org/10.1038/mp.2013.161`.

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): e68141, 2013. doi: 10.1371/journal.pone.0068141. URL `http://dx.doi.org/10.1371/journal.pone.0068141`.

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.

Noa Novershtern et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144 (2):296–309, Jan 2011. doi: 10.1016/j.cell.2011.01.004. URL `http://dx.doi.org/10.1016/j.cell.2011.01.004`.

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

Deborah J Smyth, Vincent Plagnol, Neil M Walker, Jason D Cooper, Kate Downes, Jennie HM Yang, Joanna MM Howson, Helen Stevens, Ross McManus, Cisca Wijmenga, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *New England Journal of Medicine*, 359(26):2767–2777, 2008.

Nathan Srebro. *Learning with matrix factorizations*. PhD thesis, Citeseer, 2004.

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. doi: 10.1371/journal.pcbi.1000770. URL `http://dx.doi.org/10.1371/journal.pcbi.1000770`.

Aravind Subramanian, Pablo Tamayo, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43): 15545–15550, Oct 2005. doi: 10.1073/pnas.0506580102. URL `http://dx.doi.org/10.1073/pnas.0506580102`.

Yu-Xiong Wang et al. Nonnegative matrix factorization: A comprehensive review. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1336–1353, 2013.

Daniela M Witten et al. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

Hui Zou et al. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.