Additional Results And Figures

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.

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

$$Y = ZB + E. \tag{1}$$

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.

Our basic evaluation strategy is based on computing the maximal correlations between simulated and recovered latent variables, and for the purpose of comparisons with other methods we use the absolute value so as to allow factors with reversed sign. Fig. 1 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 [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 as input. SPC is another popular method that can enforce sparsity and positivity, it has one hyperparameter that we set by cross-validation for each component 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

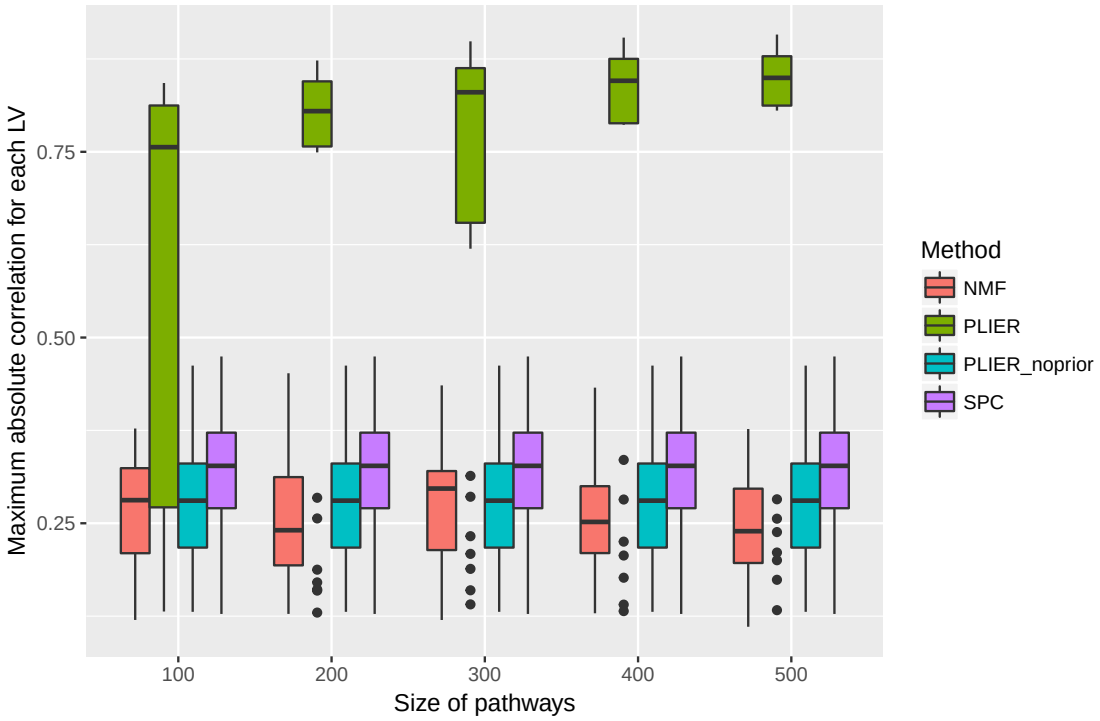

Figure 1: Boxplot of the correlation between simulated LVs and those recovered by various decomposition methods. We compare PLIER against two other methods, NMF [Brunet et al., 2004] and SPC [Witten et al., 2009], as well as PLIER run without using any prior information. In this simulation we provide 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. 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.

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. 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. 3).

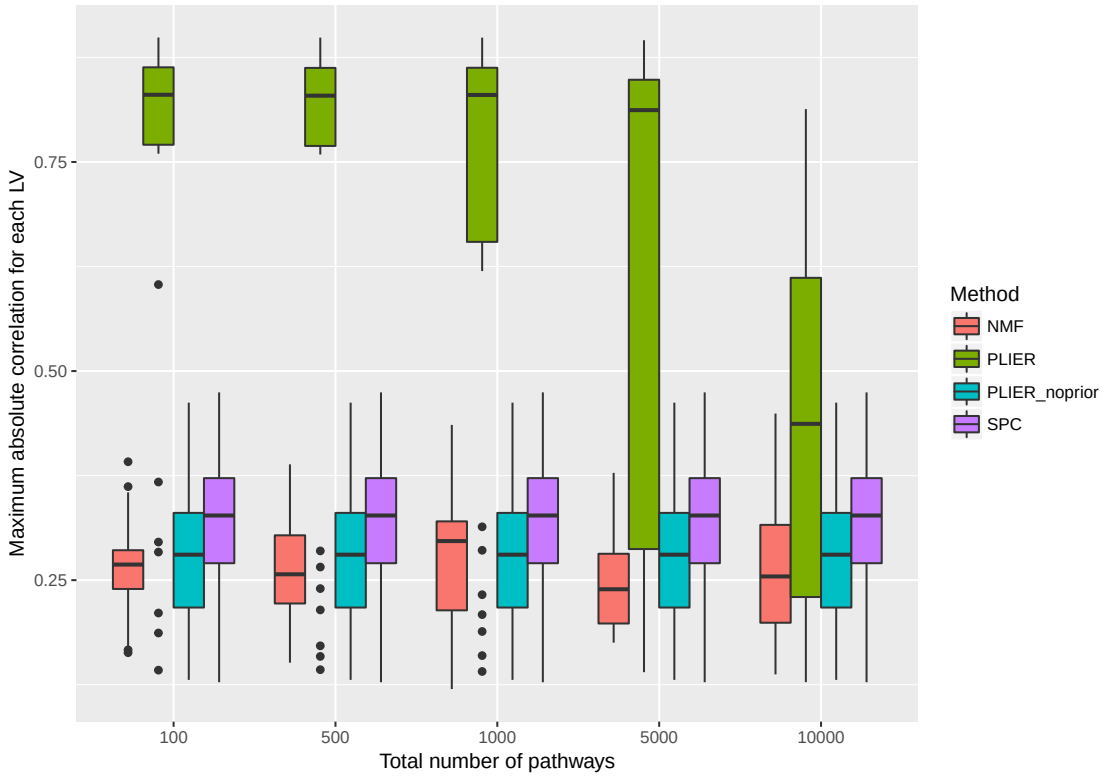

Figure 2: Data is simulated as in Fig. 1 except that the number of genes per pathway is kept at 300 and the number of uninformative pathways is varied. As the prior information gets noisy, PLIER's performance approaches those of others. Statistics were computed using Pearson correlation across 300 samples. 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.

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 effect 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. 4A.

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. 4A, left panel, bottom rows) and produces non-informative LVs (Fig. 4A, right panel, bottom rows).

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. 4B).

Overall we find that almost all LVs with credible prior information association (FDR<0.05,

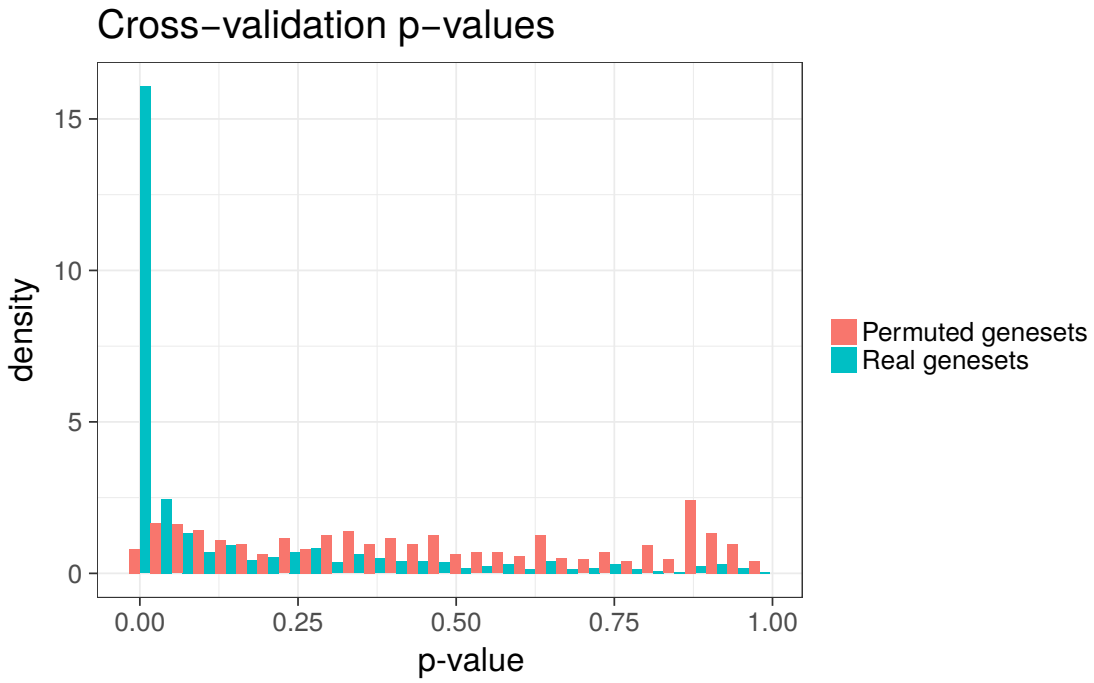

Figure 3: Histogram comparisons of pathway association p-values produced with real genesets and a single run with gene-label permuted genesets. Statistics were calculated the held-out set of genes and genes which are not annotated to the pathway. P-values are calculated with a two-sided Wilcoxon rank-sum test. Uncorrected p-values are plotted in the histogram.

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. 5. 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.

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 [Battle et al., 2014] 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 Supplementary Fig. 1. Moreover, the matching LVs use prior information genesets that are either the same or closely related (see row/column names in Supplementary Fig. 1).

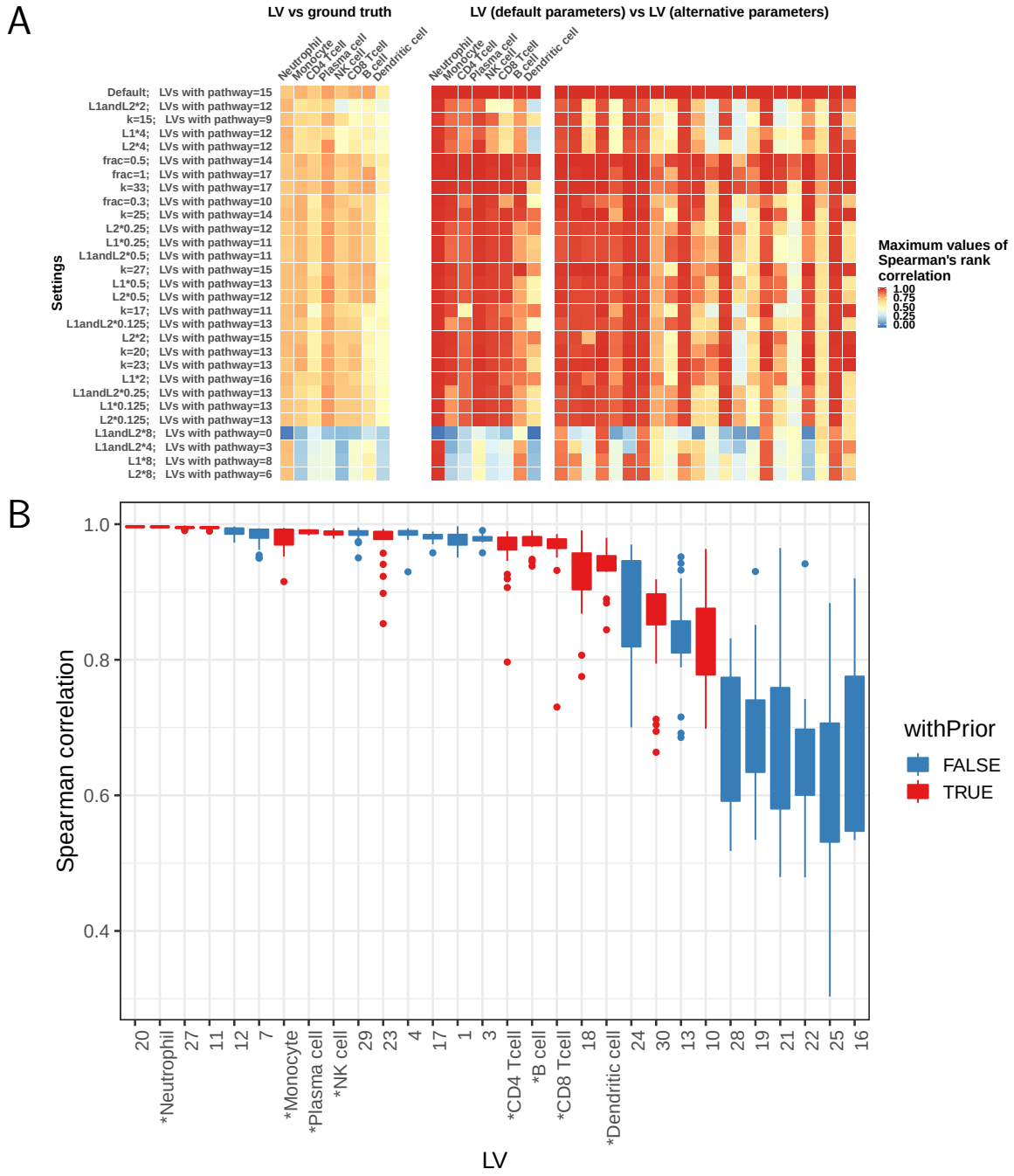

Figure 4: (A) Robustness of LVs with respect to different parameter choices. Row labels indicate the parameter settings and the number of significant pathway associations. The L1 and L2 parameters are reported relative to the default. (Left panel) Maximum rank correlation of LVs with the ground truth 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.

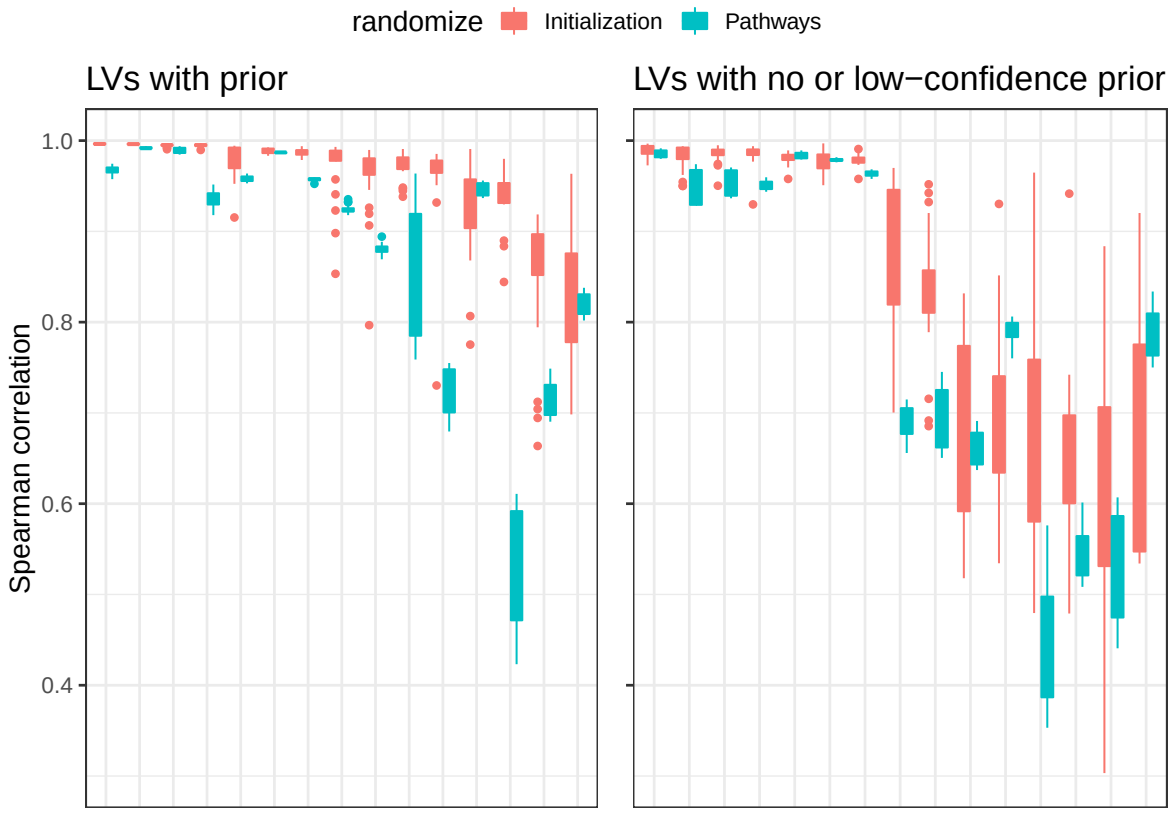

Figure 5: Robustness of LVs with respect to pathway randomization compared to robustness of LVs with respect to random initialization. Statistics were computed using Spearman rank correlation across 35 subjects.

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* (Supplementary Fig. 1, Inset).

Distributions of PLIER loadings

We plot loading statistics from our analysis of the DGN dataset in Fig. 6. 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 "Pathway recovery significance")

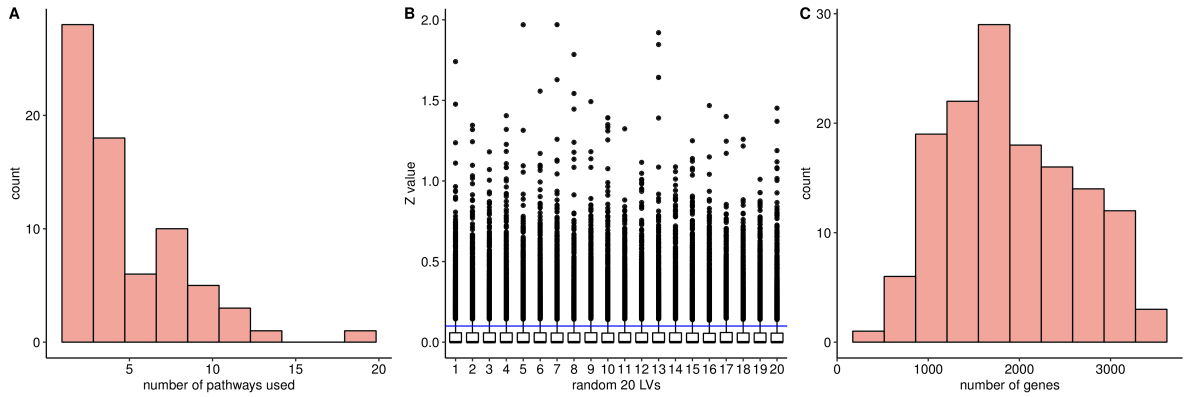

Figure 6: (A) The distribution of number of LV-associated pathways per LV. (B) The boxplot of loading value corresponding to 20 random LVs. Boxplot displays the 25th, 50th and 75th percentiles, with whiskers extending to 1.5x the interquartile range or the range of the data whichever is smallest. (C) The distribution of number of genes with loading values above 0.1.

LV interpretation and naming

The top genes contributing to each genotype-associated LV are depicted in Supplementary Fig. 6. In many cases the identity of the genes and the corresponding PLIER pathway utilization (see U matrix visualized in the main text, Fig. 2) 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 of immune cell types, ImmGen [Heng et al., 2008]. The results are visualized in Supplementary Fig. 6. 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 main text, Table 1) on proportions of various myeloid cell types as determined in a large GWAS study of blood cell-type composition [Astle et al., 2016]. 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 [Novershtern et al., 2011]. 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.

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. For this analysis, rather than using cross validation the SPC sparsity parameter was explicitly optimized to maximize the the eQTL discovery objective.

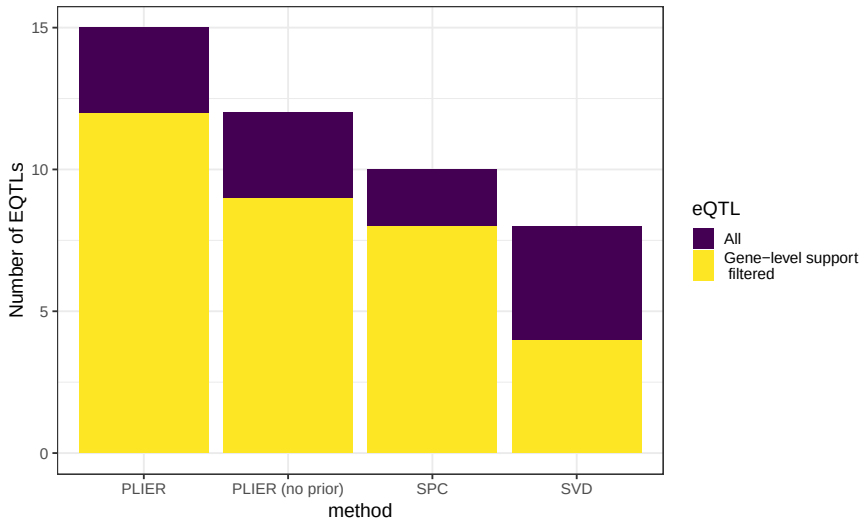

Figure 7: Comparison of eQTL discovery results from different decomposition methods.

Analysis of single-cell RNA sequencing dataset

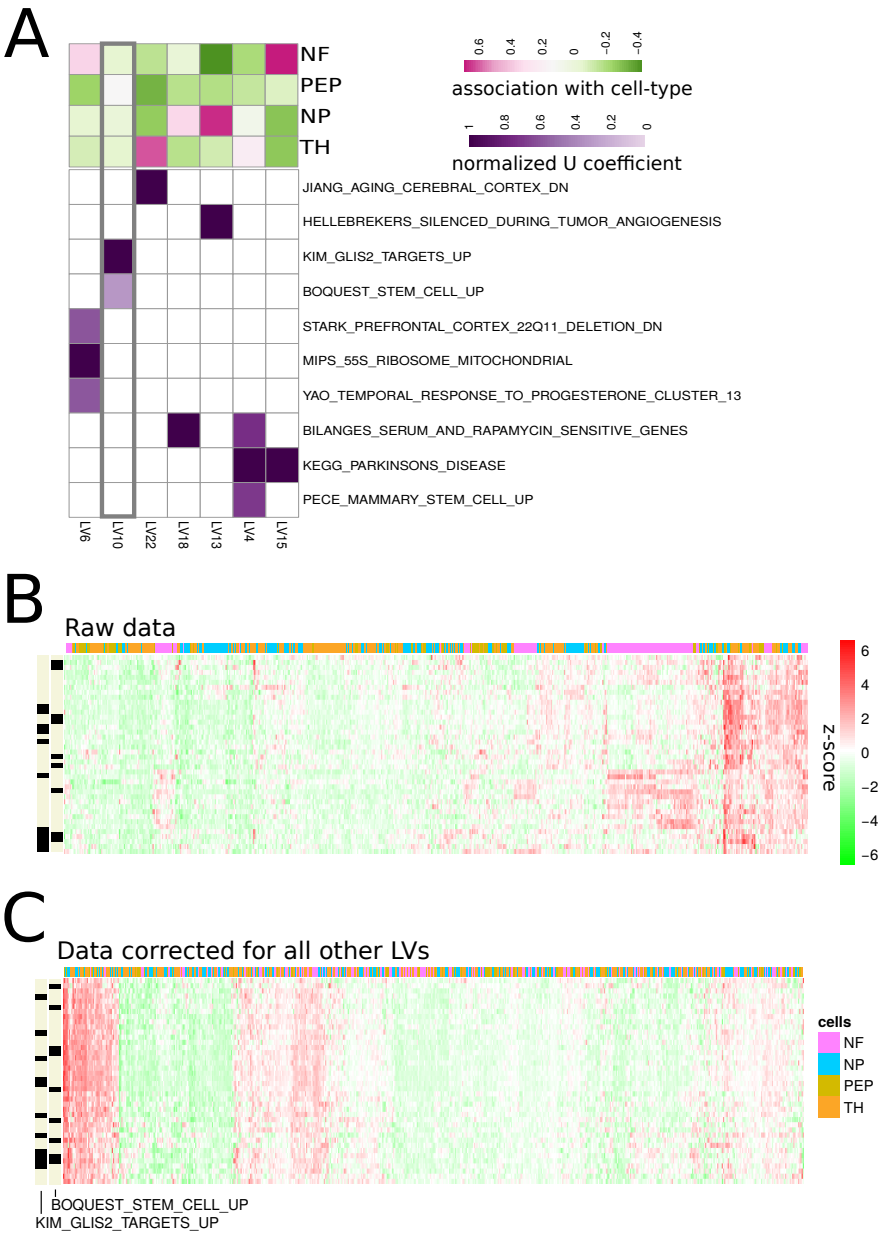

Figure 8: (A) The subset of the U matrix with the highest-confidence (AUC>0.75, FDR<0.01) pathway associations. Spearman rank correlations with cell types (139, 169, 81 and 233 for NF, NP, PEP, and TH respectively) defined in the original paper are displayed above. While many of the LVs are correlated 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.

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 [Usoskin et al., 2015]. 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. 8A). 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. 8B) but stands out clearly when correcting for other sources of variation (Fig. 8C). Thus, PLIER is able to both reveal additional heterogeneity in this complex dataset and associate it with prior information in a single computational step.

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 decompositions 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. 9A and B). Moreover, the matched LVs often use either the same or highly related prior information (Fig. 9B). 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. 9C).

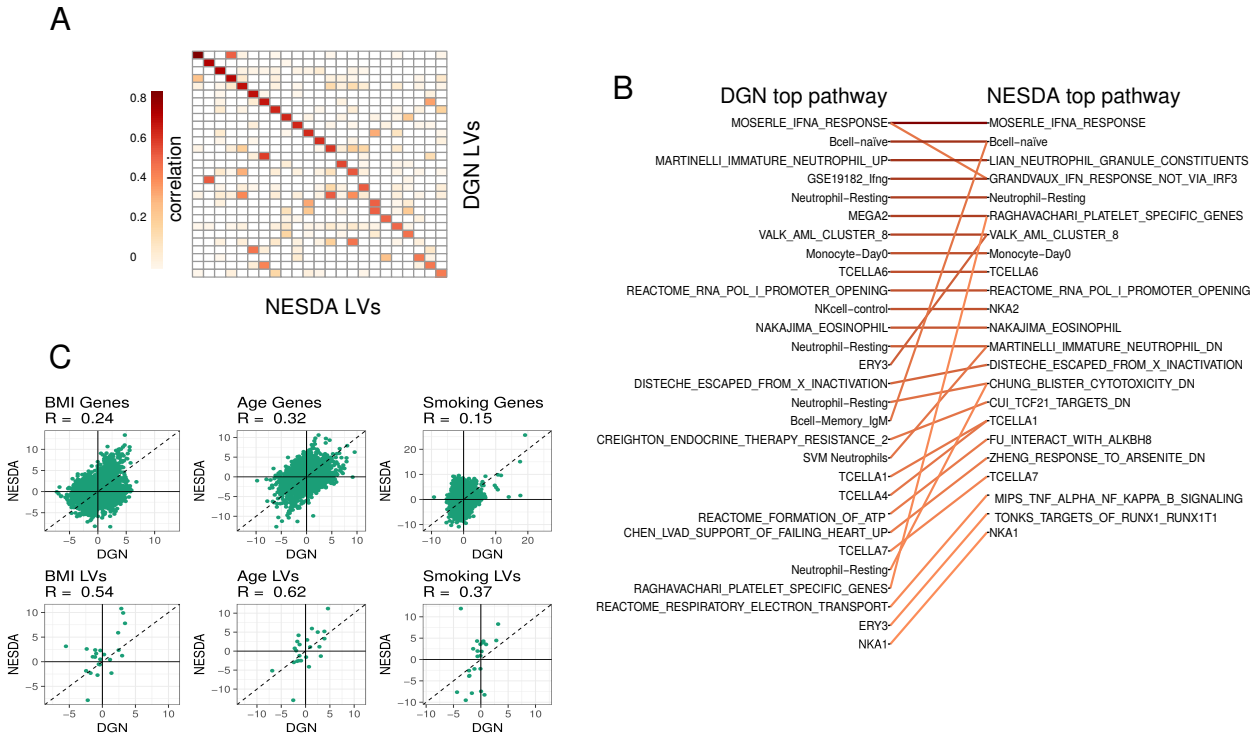

Figure 9: **LV-based meta-analysis increases cross-dataset concordance** Two whole blood datasets DGN (RNAseq) and NESDA (array) were independently decomposed using PLIER. We assessed the correspondence between the resulting LVs by comparing their loadings on the common set of genes. (A) All pairwise 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.

Discussion

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 [Abbas et al., 2009, Novershtern et al., 2011], there are two important distinctions. PLIER is considerably more tolerant of errors in marker genes since the 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 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.

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 [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 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 [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 on large numbers of groups and its native form does not explicitly deal with the issue of geneset/pathway incompleteness.

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.

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`.

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.

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.

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.

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

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.

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.

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

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.