nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-019-0456-1

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

Weiguang Mao[1,2], Elena Zaslavsky[3], Boris M. Hartmann[3], Stuart C. Sealfon[3] and Maria Chikina[1,2*]

A major challenge in gene expression analysis is to accurately infer relevant biological insights, such as variation in cell-type proportion or pathway activity, from global gene expression studies. We present pathway-level information extractor (PLIER) (https://github.com/wgmao/PLIER and http://gobie.csb.pitt.edu/PLIER), a broadly applicable solution for this problem that outperforms available cell proportion inference algorithms and can automatically identify specific pathways that regulate gene expression. Our method improves inter-study replicability and reveals biological insights when applied to *trans*-eQTL (expression quantitative trait loci) identification.

One salient feature of high-dimensional molecular data structure is the presence of groups of correlated measurements. In gene expression datasets, correlated expression among genes commonly represents coordinated transcriptional regulation or, in studies of heterogeneous tissues, variation in the proportion of the different cell types. Identification of the mechanisms that underlie coordinated gene expression changes is crucial for interpretation. Importantly, correlated expression patterns may also be the result of various technical factors, often referred to as 'batch effects'[1]. The challenge is to identify and interpret biologically meaningful signatures while reducing any negative effects of technical noise. To this end we have developed pathway-level information extractor (PLIER), which performs an unsupervised data structure deconvolution and mapping to prior knowledge, and identifies regulation in cell-type proportions or pathway activity while reducing technical noise.

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 (gene sets) and produces a dataset deconvolution that optimizes alignment of LVs to a relevant subset of the available gene sets. The method automatically finds these relevant gene sets among the hundreds to thousands considered (Fig. 1a). Technical noise reduction is also achieved during the deconvolution, as technical factors are segregated preferentially into LVs that do not associate with prior information (Supplementary Figs. 1 and 2).

We first validate the method using cell-type proportion inference because (1) it is an important step of gene expression interpretation, (2) other methods are available for comparative benchmarking, and (3) predictions can be tested against a direct measurement gold standard. For this purpose, we generated a validation dataset comprising 35 human whole-blood samples assayed by both RNA sequencing (RNA-seq) and direct mass cytometry measurement (using CyTOF (Fluidigm)) of cell-type proportion. We applied PLIER to the validation dataset using 605 pathways that included 60 cell-type markers and 555 canonical pathways from the Molecular Signatures Database (MSigDB) of the Broad Institute[2]. We produced a decomposition with 14 LVs annotated with high confidence (area under curve (AUC) of >0.7, false discovery rate (FDR) of <0.05; see Methods for cross-validation procedure) to one or more gene sets, 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 value of 0.71 (range 0.58–0.78) (Fig. 1).

We compared PLIER against the established methods for mixture decomposition inference. These methods rely either on low-rank matrix decomposition or on 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 (SPC) analysis (see Methods for details). For a reference-based approach, we tested Cibersort[3] and NNLS (non-negative least squares)[4]. Both of these approaches combine a regression algorithm with a dedicated cell-type-specific reference matrix that is optimized explicitly for human blood deconvolution.

PLIER performed markedly better than other constrained matrix decomposition methods and surprisingly outperformed the reference-based supervised approaches for four out of the eight cell types. This performance of the essentially unsupervised and generally applicable PLIER method is in part due to the capacity of PLIER to sort through many candidate gene sets and find those that are most informative for the specific dataset. PLIER can be supplied with multiple and even discordant marker sets for the same cell type and will automatically pick the one that models the data.

Although PLIER shows excellent performance when benchmarked for cell-type deconvolution, it is not designed specifically for this task. Instead, it is a general method for estimating pathway activity and can therefore be applied to a wide variety of gene expression interpretation problems.

As an example, we evaluated the usefulness of PLIER for 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[5]. Many *trans*-eQTLs exert their effect by altering the activity of a regulatory protein, which in turn affects the expression of many downstream genes[6]. *Trans*-eQTLs, which provide important insights into gene regulatory networks, are difficult to detect and are less commonly identified than *cis*-eQTLs owing to the multiple hypothesis burden of testing hundreds of thousands of variants and tens of thousands of genes.

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [2]Carnegie Mellon-University of Pittsburgh PhD Program in Computational Biology, Pittsburgh, PA, USA. [3]Department of Neurology and Center for Advanced Research on Diagnostic Assays, Icahn School of Medicine at Mount Sinai, New York, NY, USA. *e-mail: mchikina@pitt.edu

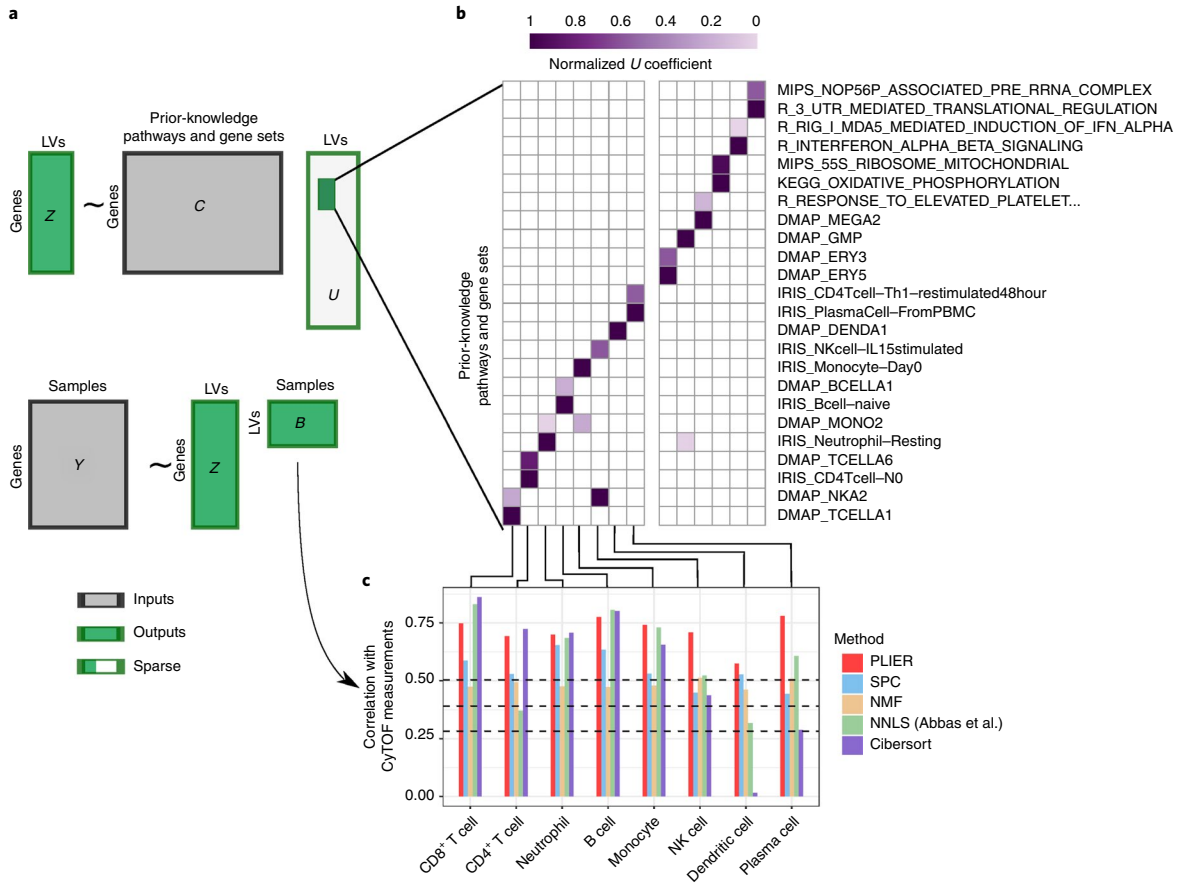

Fig. 1 | PLIER overview. PLIER is a matrix factorization approach that decomposes gene expression data into a product of a small number of LVs and their corresponding gene associations or loadings, while constraining the loadings to align with the most relevant automatically selected subset of prior knowledge. **a**, Given two inputs, the gene expression matrix *Y* and the prior knowledge (represented as binary gene set membership in matrix *C*), the method returns the LVs (*B*), their loadings (*Z*), and an additional sparse matrix (*U*) that specifies which (if any) prior-information gene sets and pathways are used for each LV. 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 heat map, facilitating the identification of the correspondence between specific LVs and prior biological knowledge. As the absolute scale of the *U* matrix is arbitrary, each column is normalized to a maximum of 1. **c**, We validate the LVs 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 (one-tailed test). NK cell, natural killer cell.

We analyzed the recently published Depression Gene Networks (DGN) dataset[5], 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 gene sets comprising biochemical and transcriptional pathways ('canonical pathways' and 'chemical and genetic perturbations' from MSigDB[3]), cell-type markers from multiple sources[3,4,7] and cytokine signatures[8]. 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 gene sets 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 the NF-κB pathway. Overall, we find that 29 LVs were unambiguously related to cell type, canonical pathways or cytokine signaling (see Supplementary Fig. 3 for *U* matrix visualization and Supplementary File 1 for a complete list of LV–gene set associations).

To perform eQTL analysis, we treated the PLIER LVs as quantitative traits (see Methods for details), and identified 12 LVs that showed significant associations with genotypes (Table 1; 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. 2a). 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 that considers only 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[9], we compared the true-positive rate, π_1 (see Methods), for gene-centric and pathway-centric eQTLs and found that the pathway-centric eQTLs were more reproducible at every *P*-value threshold. For example, at a cut-off that corresponds to a gene-level FDR of 0.2, the gene-centric π_1 is ~0.2, while for pathway-centric eQTLs it is ~0.6 (see Supplementary Fig. 4 for replication across a range of cut-offs).

In addition to improving the accuracy of *trans*-eQTL discovery, the PLIER decomposition identifies the pathway (or pathways) associated with the LV eQTL, which can provide precise biological

Table 1 | Summary table of all pathway-level effects found in the DGN dataset

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

Statistics were computed using Spearman rank correlation across 922 subjects with a two-sided test. FDRs are computed using the Benjamini–Hochberg procedure on the total number of tests (number of LVs multiplied by number of SNPs). SNP-LV associations that passed FDR < 0.05 were filtered further 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 gene set association captured in the *U* matrix. Pathways with no positive entries in *U* (indicated with an asterisk) are named based on further analysis of top genes (see Supplementary Figs. 5 and 6 and Supplementary Note) and/or the presence of a putative *cis* eQTL transcriptional mediator. The complete pathway utilization for these LVs can be seen in Fig. 2. The expression patterns for the top 15 genes driving each LV are plotted in Supplementary Fig. 6.

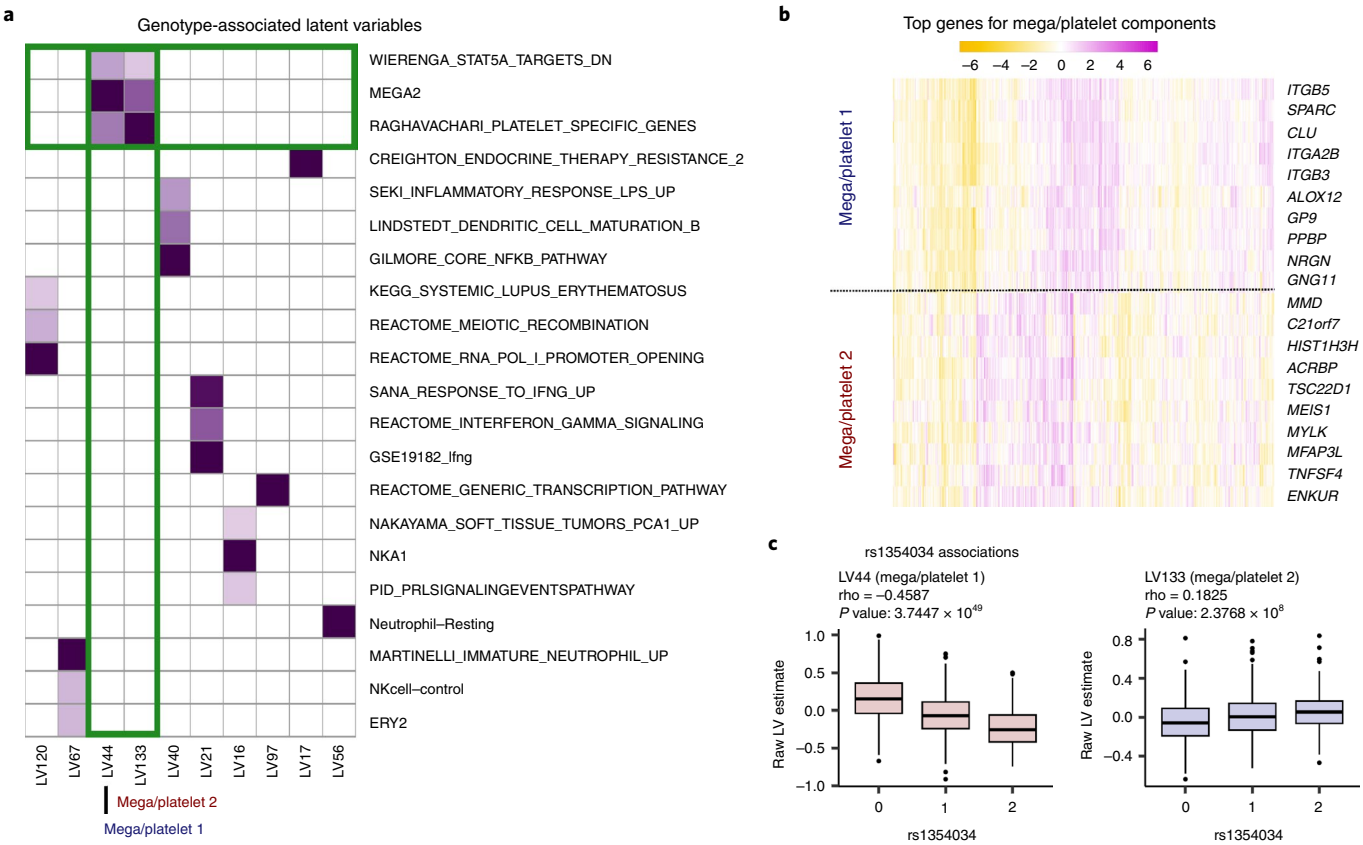

Fig. 2 | a, Pleiotropy of the *ARHGEF3* locus. A heat map of a subset of the *U* matrix corresponding to LVs with a genotype effect (LV eQTLs). Only pathways with a cross-validation FDR of <0.05 are shown. We find that two LVs (LV44 and LV133) share pathway annotations (albeit with different coefficients) that suggest a relationship with megakaryocyte and platelet (mega/platelet) biology. **b**, Heat map of the top genes in the loading for LV44 and LV133. All genes are annotated to the pathways shown in **a**, except *ENKUR*. **c**, Box plots of the association of LV44 and LV133 with SNP rs1354034 ($n = 344$, 429, and 149 for 0, 1, and 2, respectively). Althoughthe 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. The box plots display the 25th, 50th, and 75th percentiles, with whiskers extending to 1.5 times the interquartile range or the range of the data, whichever is smallest. *P* values are from the uncorrected two-tailed Spearman rank correlation test.

Table 2 | Summary table of the associations between the two megakaryocyte/platelet LVs and SNPs known to affect only one platelet phenotype

Phenotype	Reported SNP	Close gene	LV44 *P* value	LV 133 *P* value	Proxy SNP
MPV	rs10876550	*COPZ1*	**1.1847×10^{5}**	0.69933	rs10876550
PLT	rs2911132	*ERAP2*	0.13817361	**2.4417×10^{5}**	rs2549803

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[10]. Although no SNPs outside of the *ARHGEF3* 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 megakaryocyte/platelet LVs with other loci known to affect platelet biology are distinct. Specifically, the early LV133 is more closely related to the process controlling PLT (platelet number), whereas the late LV44 is related to the process controlling MPV (mean platelet volume).

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 the megakaryocyte/platelet lineage, based on their pathway association (Fig. 2a,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[5]. Using PLIER, we find that 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 mean platelet volume (MPV)[10]. Our analysis further shows that the two LVs linked to this SNP are supported by different genes that show distinct expression patterns (Fig. 2b). These results suggest that the two LV eQTLS may distinguish two different processes of megakaryocyte/platelet biology, which is supported by a recent single-cell hematopoetic lineage report. This study shows that genes associated with the two LVs are expressed at different developmental time points[11]. Specifically, mouse orthologs of *MEIS1* and *TSC22D1* (from LV133) are expressed in 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 correlated positively with each other in the DGN dataset. Notably, the effects of the rs1354034 alleles on LV133 and on LV44 are in opposite directions (with minor allele associated with increased expression on LV133 and with decreased expression for LV44) (Fig. 2c). Furthermore, we find that using partial correlation analysis, whereby the LVs are corrected for each other, dramatically improves the eQTL statistics (Supplementary Fig. 7). These results strongly argue that the effects of LV44 and LV133 are independent.

We speculate that the independent regulation of the two LV eQTLs by the same locus results from an effect on different regulatory networks 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 variation[10]. We reason that the effects of rs1354034 on multiple LVs is reflective of its pleiotropic function. Indeed, correlation of the two LVs with SNPs that are known to be specifically linked to MPV or PLT alone shows divergent patterns. In addition to the association with rs1354034, LV133 is most strongly associated with a SNP linked to platelet number, whereas LV44 is most strongly associated with a SNP linked to platelet volume (Table 2). This analysis supports a model in which *ARHGEF3* 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 mechanisms underlying complex biological processes. Other ways in which PLIER can be applied to single-cell RNA-seq or cross-study concordance analysis are presented in Supplementary Note 1.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0456-1.

Received: 15 December 2017; Accepted: 16 May 2019;
Published online: 27 June 2019

References

1. Leek, J. T. et al. *Nat. Rev. Genet.* **11**, 733–739 (2010).
2. Subramanian, A. et al. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
3. Newman, A. M. et al. *Nat. Methods* **12**, 453–457 (2015).
4. Abbas, A. R. et al. *PLoS One* **4**, e6098 (2009).
5. Battle, A. et al. *Genome Res.* **24**, 14–24 (2014).
6. Westra, H.-J. et al. *Nat. Genet.* **45**, 1238 (2013).
7. Novershtern, N. et al. *Cell* **144**, 296–309 (2011).
8. Filiano, A. J. et al. *Nature* **535**, 425–429 (2016).
9. Wright, F. A. et al. *Nat. Genet.* **46**, 430–437 (2014).
10. Gieger, C. et al. *Nature* **480**, 201–208 (2011).
11. Olsson, A. et al. *Nature* **537**, 698–702 (2016).

Acknowledgements

This work was supported by the US National Institutes of Health (NIH) grants U54HG008540 and 5R03MH109008 to M.C., 1R01HG009299 to M.C. and W.M., and 5U19AI117873 and 5U24DK112331 to Z.E., S.S.C., and H.B.M. The authors acknowledge G. Nudelman for help with RNA-seq processing. This study uses data from dbGaP (phs000486.v1). Funding support for the Genetic Association Information Network (GAIN) Major Depression study: Stage 1 Genome-wide Association In Population Based Samples Study (parent studies: NESDA) and the Netherlands Twin Register (NTR)) was provided by the Netherlands Scientific Organization (904-61-090, 904-61-193, 480-04-004, 400-05-717, Netherlands Organisation for Scientific Research (NWO) Genomics, SPI 56-464-1419), the Centre for Neurogenomics and Cognitive Research (CNCR-VU), the European Union (EU/WLRT-2001-01254), ZonMW (geestkracht program, 10-000-1002), NIMH (RO1 MH059160) and matching funds from participating institutes in NESDA and NTR, and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The datasets used for the analyses described in this manuscript were obtained from dbGaP (http://www.ncbi.nlm.nih.gov/gap) through dbGaP accession number phs000486.v1.p1. Samples and associated phenotype data for the GAIN Major Depression study: Stage 1 Genome-wide Association In Population Based Samples Study (PI, P. F. Sullivan, University of North Carolina) were provided by D. I. Boomsma and E. de Geus, VU University Amsterdam (PIs NTR), B. W. Penninx, VU University Medical Center Amsterdam, F. Zitman, Leiden University Medical Center, and W. Nolen, University Medical Center Groningen (PIs and site-PIs, NESDA).

Author contributions

M.C. conceived and led this work. W.M. and M.C. developed the analytical framework, analyzed data, and produced figures. M.C., W.M., Z.E., and S.S.C. drafted the manuscript. W.M. implemented the web interface. B.H.M. collected the RNA-seq and CyTOF data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0456-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C.

Peer review information: Lei Tang and Tal Nawy were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Gene expression measurements are highly correlated, and this correlation structure often reflects the activity of upstream biological processes. This data structure is exploited implicitly any time a clustering is performed, as is often done with cancer datasets to define molecularly distinct subtypes[12,13]. 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 uses 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 applied successfully to gain biological insight from complex datasets[14,15].

However, SVD decompositions have several limitations. By construction PCA and SVD produce 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 non-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, NMF has been applied to cancer gene expression decomposition, yielding more intuitive results[16]. Likewise, methods to introduce sparsity into the matrix decomposition have been proposed[17,18].

However, these methods 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 use of the vast information contained in biological gene sets during the decomposition. We therefore developed a platform that introduces additional constraints to explicitly and iteratively optimize the decomposition using the biological knowledge represented by a compendium of prior information.

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 $||Y-ZB||_F^2$ subject to $\text{rank}(Z)=k$, $\text{rank}(B)=k$.

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 SVD of Y. In an SVD-based decomposition, rows of B are referred to as PCs. As PCs are necessarily orthogonal, which our method does not require, we will use the more general term LVs.

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 gene sets that represent biological pathways, sets of tissue- or cell-type-specific markers, and coordinated transcriptional responses observed in genomewide experiments.

Given n genes and m gene sets, 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 jth gene set. 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

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

subject to $U>0$, $Z>0$.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 gene sets 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 gene sets represent each LV.

The parameter λ_1 keeps a balance between the proportion of prior knowledge that 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 gene set 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.

Although the 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. Although there are no convergence guarantees, in practice this algorithm converges in under a few hundred iterations.

Optimization constants. The optimization has four 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 that can be determined via permutation as in the approach proposed by Leek et al.[19] or the simple 'elbow' approach (num.pc in our package implements both). However, it is logical that the number of constrained LVs 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 ks persist when k is increased. It is also possible to optimize 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 $Z \approx UD^{1/2}$ 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 (Supplementary Note 1). 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 LVs that we wish to be associated with prior information; 0.7 by default. The λ_3 constant is then adjusted periodically by binary search to meet this goal. Although 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.

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. To quantify this, we designed a cross-validation procedure that proceeds as follows. For each pathway included in the entire prior-information compendium, a random one-fifth of the positive genes are set to 0 and this new prior-information matrix is used to run PLIER. Afterward, 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 gene set (which preserves dependencies among pathways) are uniformly distributed (Supplementary Note 1).

Although 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 gene set are used.

Validation data. *Sample processing.* We used anonymized discard samples that have the determination of non-human research. Blood was drawn into Tempus tubes (AB scientific) for RNA and into EDTA tubes for Cyto analysis. RNA was extracted using the MagMAX 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.

CyTOF sample processing. CyTOF antibodies were purchased pre-conjugated from Fluidigm (formerly DVS Sciences), or were purchased purified and then conjugated in-house using MaxPar X8 Polymer Kits (Fluidigm) according to the manufacturer's instructions. Whole-blood samples were processed within 4 h of collection and stained by addition of a titrated panel of antibodies (CD45, CCR6, CD19, CD45RA, CD141, CD4, IgD, CD16, CD127, CD123, CD66b, CD1c, CD27, CXCR3, CCR4, CCR7, CD14, CD56, CD8, CD161, CD24, CD3, CD25, CXCR5, CD38, HLADR) directly to 400 µl of whole blood. After 20 min incubation at room

temperature (25 °C), the samples were treated with 4 ml of BD FACSLyse and incubated for a further 10 min. The samples were then washed and incubated in 0.125 nM Ir intercalator (Fluidigm) diluted in PBS containing 2% formaldehyde, and stored at 4 °C until acquisition.

Immediately prior to acquisition, samples were washed once with PBS, once with de-ionized water, and then resuspended at a concentration of 1 million cells ml^{-1} 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 s^{-1}.

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[20], 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.

RNA-seq methods. The samples were sequenced SE100 to an average depth of 48.8 million reads. Quality assessment was carried out using FastQC v0.11.8[21]. Alignment to GenCode hg38 was done using STAR[22]. Transcript counts are assigned using the FeatureCounts tool (subread package)[23]. The final counts were filtered for genes that had 0 counts in all samples. The data were transformed to RPKM (reads per kilobase of transcript per million mapped reads). We also created a quantile normalized count dataset by filtering all genes that had <3 counts in any samples and by quantile normalizing the log transformed counts. This stringent filtering was performed to avoid data artifacts caused by quantile normalization of low count genes. As RNA samples were processed in two separate batches, both final datasets were corrected for batch difference.

Details of method comparisons. For validation, we compared the Spearman rank correlation of CyTOF-based proportion measurements with estimates obtained from different methods. *P* value thresholds indicated on the plot in Fig. 1c are for the one-tailed test. We compared performance on the validation dataset against four alternative approaches. Two of the approaches are matrix decomposition methods that are commonly applied to gene expression data: SPC analysis and NMF. The other two approaches are reference-based methods that are designed specifically to estimate human blood cell-type proportions: NNLS regression (originally applied to cell-type deconvolution in ref. [4]) and Cibersort[3]. 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 (fragments per kilobase of transcript per million mapped reads).

For NMF we used the default algorithm and matrix norm as implemented in the NMF R package (version 0.20.6)[16]. As NMF requires a positive matrix, we used quantile-normalized log counts, which achieved the best performance among different transformation and normalization methods tested. 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 (version 1.0.9)[18]. 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[18]. Since SPC and NMF do not assign a biological cause to the inferred LVs, 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 we used the CellMix R package implementation (version 1.6.2)[24], and for Cibersort we used the EpiDISH R package implementation (version 1.2.0)[25]. For both methods the input data were raw FPKM values, which is the preferred data transformation for Cibersort and also performed best in our evaluations. As NNLS and Ciberosrt are both reference-based methods and can be used with any reference or basis matrix, we tried both approaches with two different references, one from Abbas et al.[4] and LM22 (leukocyte signature matrix) 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 in Abbas et al.[4], we allowed various combinations of the estimates; for example, we created an 'all-B cell' 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.

Public data. *DGN dataset.* The DGN dataset is not available for public release but can be requested from the US National Institute of Mental Health (NIMH) following instructions in the original publication[26]. The NIMH database contains several normalized versions of these data, and for our study we used '*trans*' normalized data, as described in Battle et al.[5]. These data are already normalized for genotype PCs and all known technical factors, and no further normalizations were performed.

NESDA dataset. The NESDA (Netherlands Study of Depression and Anxiety) dataset[9] was obtained from dbGAP (phs000486.v1). Following suggestions from the NESDA study authors, the NESDA dataset was normalized for known technical factors and the first three genotype PCs using linear regression.

Prior-information gene sets. The generic blood cell-type marker dataset was derived from the IRIS (immune response in silico)[4] and DMAP (differentiation map)[7] datasets. Many canonical marker genes (such as *CD19*, *CD3E*, and *CD8A*) have a multimodal distribution with a high expressor group and one or more low or medium expressor groups. 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 or 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, but our method is flexible and can handle redundancy easily. 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[3] that 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' gene sets from mSigDB, and a set of transcriptional signatures relevant to immune signaling described by Filiano et al.[8].

***Trans*-eQTLs.** For the purposes of our analysis we define valid *trans* associations as gene–SNP pairs for which the gene and all of its homologs (as defined by the Ensembl database[27]) are on a different chromosome from that of the SNP.

Gene-centric eQTLs. We first computed *P* values for all valid *trans* associations using rank correlation, and then computed Benjamini–Hochberg FDR for the total number of valid *trans* association tests.

Pathway-centric eQTLs. As pathway LVs are composed of multiple genes from different chromosomes, all LV–SNP associations are potentially valid *trans* associations. We first perform rank correlation tests on all LV–SNP pairs and compute the Benjamini–Hochberg FDR for the entire set of pathway-level tests (number of LVs multiplied by the number of SNPs). Associations with FDR > 0.05 are not considered further. We subsequently compute gene-level support for pathway-level eQTLs by performing all valid *trans* association tests on the subset of SNPs that had FDR < 0.05 for at least one LV. We compute gene-level FDRs correcting for the total number of tests performed overall (all possible LV tests plus the gene tests on selected SNPs). We note that the *P* value threshold for FDR = 0.2 in the PLIER-centric analysis is higher than in the gene-centric analysis (4.1×10^{-7} versus 7.7×10^{-8}). It is possible that these PLIER-centric FDRs are overly permissive since many SNPs are filtered out on the basis of a lack of pathway-level associations. However, we emphasize that we do not rely on these values for any conclusions in our analysis. They are used only to define the upper limit for associations that are checked for replication. Finally, we filter pathway-level effects with low gene-level support. We defined low gene-level support as 0 *trans* gene–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 *trans* gene in gene-centric analysis, at a permissive FDR. This step is designed to remove any spurious pathway associations that could arise from *cis* genes or *cis* homologs contributing to the pathway-level estimate. We found that this last filtering step removed two associations in the HLA locus where all the associated genes were in *cis*.

Replication. To assess replication in the NESDA dataset, SNPs were matched based on linkage disequilibrium using the LDlink tool with CEU (Northern Europeans from Utah) population[28]. Specifically, if the exact SNP was not present in the NESDA dataset, we selected the SNP with the highest linkage disequilibrium, and if multiple SNPs had the same linkage disequilibrium, we took the one closest in genomic coordinates. We considered a match only if the best linkage disequilibrium was above 0.5. We assessed the relationship between the NESDA replication π_1 and the *P* value obtained in DGN in two different ways. One method uses a consistent cut-off of $\lambda = 0.05$, so that the pi_1 estimate is simply computed as 1 minus the fraction of *P* values above 0.05 divided by 0.95. We also evaluate pi_1 using the method implemented by 'qvalue' of the Bioconductor package[29]. This method selects the optimal λ for each π_1 estimate. We find that the typical value is approximately 0.8, although a different value may be selected at each threshold resulting in more noise in the π_1 curve.

Platelet phenotypes. The sentinel SNPs and their relevant phenotypes (MPV, PLT, or both) are supplied as a supplementary table in Gieger et al.[10]. Proxy SNPs were defined as above.

Software availability. The method, auxiliary functions, and example datasets (including gene expression and cell proportion data used to generate Fig. 1) are

compiled in the PLIER R package available at https://github.com/wgmao/PLIER. PLIER can also be used via an online interface located at http://gobie.csb.pitt.edu/PLIER/.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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 the Gene Expression Ominibus (GSE130824). The DGN dataset can be obtained from NIMH following the instructions provided in ref. [26]. The NESDA dataset can be obtained from dbGAP (identifier: phs000486.v1).

References

12. The Cancer Genome Atlas Research Network. *Nature* **497**, 67–73 (2013).
13. Ross, D. T. et al. *Nat. Genet.* **24**, 227–235 (2000).
14. Alter, O., Brown, P. O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **97**, 10101–10106 (2000).
15. Hore, V. et al. *Nat. Genet.* **48**, 1094 (2016).
16. Brunet, J.-P. et al. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).
17. Zou, H. & Hastie, T. J. R. *Stat. Soc. Ser. B* **67**, 301–320 (2005).
18. Witten, D. M. et al. *Biostatistics* **10**, 515–534 (2009).
19. Leek, J. T. et al. *PLoS Genet.* **3**, e161 (2007).
20. Levine, J. H. et al. *Cell* **162**, 184–197 (2015).
21. Wingett, S. W. & Andrews, S. *F1000Res.* **7**, 1338 (2018).
22. Dobin, A. et al. *Bioinformatics* **29**, 15–21 (2013).
23. Liao, Y., Smyth, G. K. & Shi, W. *Bioinformatics* **30**, 923–930 (2013).
24. Gaujoux, R. & Seoighe, C. *Bioinformatics* **29**, 2211–2212 (2013).
25. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. *Nat. Methods* **15**, 1059–1066 (2018).
26. Mostafavi, S. et al. *Mol. Psychiatry* **19**, 1267–1274 (2014).
27. Zerbino, D. R. et al. *Nucleic Acids Res.* **46**, D754–D761 (2017).
28. Machiela, M. J. & Chanock, S. J. *Bioinformatics* **31**, 3555–3557 (2015).
29. Storey, J. D. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).

nature research

Corresponding author(s): Maria Chikina

Last updated by author(s): Apr 22, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	RNAseq processing: STAR 2.5.2 Subread 1.5.3.
Data analysis	Cytof clustering: Phenograph 1.5. Computation was done using R version 3.4.0. PLIER is available at https://github.com/wgmao/PLIER. PLIER depend on glmnet (version used 2.0-10) and q-value (version used 3.16.7) and rsvd (version used 0.6)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Processed gene expression and cell proportion measurements generated for this study are available through the PLIER package. The Depression Susceptibility Genes and Networks (DGN) dataset can be obtained from NIHM following instructions provided in the original publication \citep{Mostafavi2014}. The NESDA dataset can be obtained from obtained from dbGAP (identifier: phs000486.v1, web: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000486.v1.p1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not perform sample size calculations for the gene expression data generated in this study. The number of samples was determined by availability. We obtained 35 paired gene expression and Cytof profiles and this number was sufficient to generate statistically significant predictions of cell-type proportions using a variety of methods so no further sample collection was needed. All other datasets used are publicly available.
Data exclusions	No data was excluded. One of the gene expression samples does not have corresponding Cytof measurements because of limited material. Material availability was a pre-established exclusion criterion. This sample is included in the PLIER analysis but not in the Cytof validation.
Replication	Our analysis of the in-house generated gene expression data and the Usoskin et al. single cell data are included in the PLIER package documentation. Other datasets used in this study (DGN and NESDA) are not available for public release.
Randomization	Since we do not have any contrasts in our analysis no randomization was necessary.
Blinding	Since we do not have any contrasts in our analysis no blinding was necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging