**Fitting the topic model**

We used fastTopics [countClust2017, fastTopics2022] to fit a topic model to the (UMI?) counts, with *K* = 16 topics. fastTopics implements the following two-step approach to fit the topic model: (1) fit a non-negative matrix factorization based on a *Poisson model* (“Poisson NMF”) [hien2021]; (2) recover maximum-likelihood estimates (MLEs) of the topic model parameters by a simple reparameterization [fastTopics2022].

In detail, we took the following steps. First, we removed genes with very low expression (total UMI count ≤ 20). Therefore, the *K* = 16 topic model was fit to UMI counts for 364 samples and 28,209 genes. Second, we ran 20 expectation maximization (EM) updates, without extrapolation, to get close to an MLE solution (“prefitting phase”). This prefitting phase was implemented in R by calling fit\_poisson\_nmf from fastTopics with the following settings: numiter = 20, method = "em", init.method = "random", control = list(nc = 8). Third, we performed an additional 180 coordinate descent (CD) updates, with extrapolation, to improve the fit (“refinement phase”). This refinement phase was implemented by calling fit\_poisson\_nmf with the following settings: method = "scd", numiter = 180, control = list(numiter = 4, nc = 8, extrapolate = TRUE), in which the model fit was initialized using the fit obtained from the prefitting phase. Finally, the topic model was recovered from the Poisson NMF model by calling function poisson2multinom. The convergence diagnostics suggested that, after a total of 200 iterations of the Poisson NMF optimization, the parameter estimates were close to an MLE; the change in log-likelihood between successive iterations was less than 1 × 10-5, and the largest residual in the first-order (“Karush-Kuhn-Tucker”) conditions was less than 1.

Reassuringly, the estimated topics captured the predominant expression patterns, most of which identify the 13 tissues (the one exception was iLN and SP, which shared the same expression pattern). Two other topics (topics 1 and 6) captured variation specific to two tissues (LI and PBMC), and one topic (topic 9) captured changes in expression over time that were not specific to any one tissue.

*If necessary, we can add details here about fitting topic models separately to each tissue.*

**Visualizing the topic proportions**

The *n* × *K* matrix of topic proportions, *L*, where *n* denotes the number of RNA-seq samples and *K* is the number of topics, can be viewed as an embedding of the samples in a (*K* − 1)- dimensional space. A simple way to visualize this embedding in 2-d (or 3-d) is to apply a nonlinear dimensionality reduction technique such as *t*-SNE [tsne2008] to *L*. An alternative powerful approach, first suggested by [countClust2017], is to visualize all *K* − 1 dimensions simultaneously using a Structure plot, which has been used with great success in population genetics [rosenberg2002]. The Structure plot is essentially a stacked bar chart, in which bars correspond to samples (rows of *L*) and bar heights (in different colors, one for each topic) are determined by the topic proportions. To arrange the samples in the Structure plot, we first grouped the RNA-seq samples by tissue, then we ordered them within each tissue by time point.

**Differential expression analysis allowing for grades of membership**

To annotate the topics, we used the grade-of-membership differential expression (GoM DE) analysis methods developed in [GoM-DE2022] and implemented in the de\_analysis function in the fastTopics package. In brief, the GoM DE analysis is conceptually similar to a standard DE analysis (e.g., [soneson2018, wang2019]), but extends the idea of comparing expression between groups by allowing the cells to have *partial membership to multiple groups* (here, the groups are the topics in the topic model). We called the de\_analysis function with the following settings: shrink.method = "ash", pseudocount = 0.1 and control = list(ns = 1e5, nc = 8). We performed a second DE analysis, with the same settings, after merging topics 1 and 5 (capturing variation in PBMC expression), 2 and 13 (BM) and topics 6 and 14 (LI). The GoM DE analysis produces, for each gene *j* and topic *k*, estimates of differences in expression, and statistics quantifying support for these differences. In the de\_analysis interface, expression differences are defined by the “least extreme” log-fold change (“l.e. LFC”), which is defined for gene *j* and topic *k* as the log-fold change that is the smallest in magnitude among topic pairs (*k*, *l*). After computing initial estimates, the GoM DE analysis (with shrink.method = “ash”) performs an adaptive shrinkage step [ash2016], separately for each topic, to stabilize the l.e. LFC estimates. We used the posterior mean estimates, posterior standard errors, posterior z-scores (posterior mean/s.e.) and local false sign rates (*lfsr*) produced by the adaptive shrinkage step to report results of the GoM DE analysis. Note that the *lfsr* can be interpreted similarly to the *q*-value [qvalue2003, qvalue2003b], for example, although the *lfsr* tends to be more conservative than quantities such as the *q*-value that control for the false discovery rate [ash2016].

**Gene sets**

Mouse gene sets for the gene set enrichment analyses (GSEA) were compiled from the following gene set databases: NCBI BioSystems [biosystems2009]; Pathway Commons [pc2010, pc2019]; and MSigDB [gsea2005, msigdb2011, msigdb2015], which included Gene Ontology (GO) gene sets [go2000, go2020]. Specifically, we downloaded bsid2info.gz and biosystems\_gene.gz from the NCBI FTP site (https://ftp.ncbi.nih.gov/gene) on March 22, 2020; PathwayCommons12.All.hgnc.gmt.gz from the Pathway Commons website (https://www.pathwaycommons.org) on March 20, 2020; and msigdb\_v7.2.xml.gz from the MSigDB website (https://www.gsea-msigdb.org) on October 15, 2020. For the gene set enrichment analyses, we also downloaded the mouse gene information (“gene info”) file Mus\_musculus.gene\_info.gz from the NCBI FTP site on October 15, 2020. To facilitate integration of these gene sets into our analyses, we have compiled these gene sets into an R package [pathways2021].

We performed two gene set enrichment analyses. In the first GSEA, we included all gene sets other than the following MSigDB collections: C1, C3, C4 and C6, and gene sets labeled as “archived”. In the second GSEA, we focused on the curated pathways, specifically gene sets belonging to the GO and CP subcategories in the MSigDB C2 gene set collection. In both analyses, we removed gene sets with fewer than 10 genes and with more than 400 genes. After removing these gene sets, 21,442 candidate gene sets remained for the first analysis, and 8,939 gene sets remained for the second analysis.

**Gene set enrichment analysis**

We took a simple multiple linear regression approach to the gene set enrichment analysis (GSEA), in which we modeled, for a given topic *k*, the (least extreme) LFC estimates for all genes using a multiple linear regression model in which the regression variables were gene-set memberships, 1 if gene *i* belongs to gene set *j*, otherwise 0. The model fitting for the multiple linear regression model was implemented in SuSiE [susie2020]. The idea behind this simple multiple linear regression approach is that the most relevant gene sets are those that best explain the log-fold changes, and therefore in the multiple regression we sought to identify these gene sets by finding regression coefficients that are nonzero with high probability. Modeling the LFC estimates also helped distinguish among DE genes that show only a slight increase in expression versus those that are highly overexpressed. This simple multiple linear approach ignored uncertainty in the LFC estimates. So, to address this issue, we shrunk the LFC estimates prior to running the GSEA; that is, we defined the regression outcome to be the posterior mean LFC estimate after applying adaptive shrinkage, as described above. This had the effect that genes that we were more uncertain about had an LFC estimate that was zero or near zero.

A benefit to using SuSiE is that it automatically organizes similar or redundant gene sets into *credible sets* (CSs) [susie2020], making it easier to quickly recognize complementary gene sets.

In detail, the GSEA was performed as follows. We performed a separate GSEA for each topic, Specifically, for topic *k*, we ran the susieR function susie with the following options: L = 10, intercept = TRUE, standardize = FALSE, estimate\_residual\_variance = TRUE, refine = FALSE}, compute\_univariate\_zscore = FALSE and min\_abs\_corr = 0. We set *L* = 10 so that SuSiE returned at most 10 credible sets—that is, at most 10 enriched gene sets.

For a given topic *k*, we reported a gene set as being enriched if it was included in at least one CS. We organized the enriched gene sets by credible set (specifically, 95% credible sets). We also recorded the Bayes factor for each CS, which gives a measure of the level of support for that CS. For each gene set included in a CS, we reported the posterior inclusion probability (PIP), and the posterior mean estimate of the regression coefficient. In the results, we refer to the regression coefficient as the enrichment coefficient for a given gene set *j* since it is an estimate of the expected increase in the LFC for genes that belong to gene set *j* relative to genes that do not belong to the gene set.

Often, a CS contained only one gene set, in which case the PIP for that gene set was close to 1. In several other cases, the CS contained multiple similar gene sets; in these cases, the smaller PIPs indicate that it is difficult to choose among the gene sets because they are similar to each other. (Note that the sum of the PIPs in a 95% CS should always be above 0.95 and less than 1.) Occasionally, SuSiE returned a CS with a small Bayes factor containing a large number of gene sets. When this happened, we did not include these gene sets in the results.

**Computing environment for topic modeling and gene set enrichment analyses**

Most computations on real data sets were performed in R 3.5.1 [R2018], linked to the OpenBLAS 0.2.19 optimized numerical libraries, on Linux machines (Scientific Linux 7.4) with Intel Xeon E5-2680v4 (“Broadwell”) processors. For performing the model topic model optimization and DE analysis, which included multithreaded computations, as many as 8 CPUs and 16 GB of memory were used. More details about the computing environment, including the R packages used, were recorded in the workflowr [workflowr2019] pages in the companion code repository [TO DO: create Zenodo repository].

**References**

[ash2016]: Stephens, M.: False discovery rates: a new deal. *Biostatistics* 18(2), 041 (2016)

[biosystems2009] Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W., Bryant, S.H.: The NCBI biosystems database. *Nucleic Acids Research* 38(supplement-1), 492–496 (2009)

[cistopic2019] González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., Aerts, S.: cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods* 16(5), 397–400 (2019)

[countClust2017] Dey, K.K., Hsiao, C.J., Stephens, M.: Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics* 13(3), 1006599 (2017)

[fastTopics2022] Carbonetto, P., Sarkar, A., Wang, Z., Stephens, M.: Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv* 2105.13440 (2021)

[go2000] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1), 25–29 (2000)

[go2020] The Gene Ontology Consortium: the Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* 49(D1), 325–334 (2020)

[GoM-DE2022] Carbonetto, P., Luo, K., Sarkar, A., Hung, A., Pott, S., Stephens. M.: Interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. *Manuscript in preparation.*

[gsea2005] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43), 15545–15550 (2005)

[hien2021] Hien, L.T.K., Gillis, N.: Algorithms for nonnegative matrix factorization with the Kullback–Leibler divergence. *Journal of Scientific Computing* 87(3), 93 (2021)

[msigdb2011] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P.: Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12), 1739–1740 (2011)

[msigdb2015] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P.: The molecular signatures database hallmark gene set collection. *Cell Systems* 1(6), 417–425 (2015)

[pathways2021] Carbonetto, P., Stephens, M.: pathways: gene set enrichment analysis using human and mouse gene sets. R package version 0.1-20 (2021). https://github.com/stephenslab/pathways

[pc2010] Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* 39(supplement 1), 685–690 (2010)

[pc2019] Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., Mistry, H., Mosier, L., Dlin, J., Wen, Q., O’Callaghan, C., Li, W., Elder, G., Smith, P.T., Dallago, C., Cerami, E., Gross, B., Dogrusoz, U., Demir, E., Bader, G.D., Sander, C.: Pathway Commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Research* 48(D1), 489–497 (2019)

[qvalue2003] Storey, J. D.: The positive false discovery rate: a Bayesian interpretation and the q -value. *Annals of Statistics* 31(6), 2013–2035 (2003).

[qvalue2003b] Storey, J. D., Tibshirani, R.: Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100(16), 9440–9445 (2003).

[R] R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org

[rosenberg2002] Rosenberg, N.A.: Genetic structure of human populations. *Science* 298(5602), 2381–2385 (2002)

[soneson2018] Soneson, C., Robinson, M.D.: Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods* 15(4), 255–261 (2018)

[susie2020] Wang, G., Sarkar, A., Carbonetto, P., Stephens, M.: A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society, Series B* 82(5), 1273–1300 (2020)

[tsne2008] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(86), 2579–2605 (2008)

[wang2019] Wang, T., Li, B., Nelson, C.E., Nabavi, S.: Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics* 20, 1–16 (2019)

[workflowr2019] Blischak, J.D., Carbonetto, P., Stephens, M.: Creating and sharing reproducible research code the workflowr way [version 1; peer review: 3 approved]. *F1000Research* 8(1749) (2019)