nature methods

ANALYSIS
https://doi.org/10.1038/s41592-019-0372-4

Evaluating measures of association for single-cell transcriptomics

Michael A. Skinnider[1*], Jordan W. Squair[2] and Leonard J. Foster[1,3*]

Single-cell transcriptomics provides an opportunity to characterize cell-type-specific transcriptional networks, intercellular signaling pathways and cellular diversity with unprecedented resolution by profiling thousands of cells in a single experiment. However, owing to the unique statistical properties of scRNA-seq data, the optimal measures of association for identifying gene-gene and cell-cell relationships from single-cell transcriptomics remain unclear. Here, we conducted a large-scale evaluation of 17 measures of association for their ability to reconstruct cellular networks, cluster cells of the same type and link cell-type-specific transcriptional programs to disease. Measures of proportionality were consistently among the best-performing methods across datasets and tasks. Our analysis provides data-driven guidance for gene and cell network analysis in single-cell transcriptomics.

Single-cell RNA-sequencing (scRNA-seq) permits transcriptome-wide quantification of gene expression in individual cells. Technological advances in single-cell sequencing protocols have led to exponential growth in the number of cells that can be profiled in a single experiment[1]. In response to these advances, a myriad of computational methods have been developed to facilitate preprocessing, differential expression, cell-type identification and pseudotemporal ordering of scRNA-seq datasets, among other tasks[2].

The maturation of scRNA-seq technology provides an opportunity to identify cell-type-specific functional modules[3–6], regulatory networks[7] and their genetic determinants[8] at a resolution that was until recently impossible, as exemplified by the recent discovery of the first quantitative trait loci for coexpression[8]. Because the number of cells profiled by a single scRNA-seq experiment can exceed the number of samples profiled by bulk RNA-seq even in large consortium projects, single-cell transcriptomics provides a fertile ground to characterize new cellular circuits[9]. However, single-cell transcriptomic datasets are distinguished in multiple respects from bulk RNA-seq datasets, in particular by an abundance of dropouts and by overdispersion[10]. The optimal measure of association to identify gene regulatory relationships in these datasets is consequently unclear. A sensitive and specific measure of association that could be used to compare transcriptome-wide expression profiles between pairs of cells would likewise be of great use for comparisons of cell types across species, batches or datasets[11–13], and ultimately for the reconstruction of networks of cell types[14–16].

Here, we conducted a systematic evaluation of 17 measures of association over a range of tasks in single-cell transcriptomics, including gene network analysis, cell clustering and disease gene identification. We first evaluated the functional coherence of gene coexpression networks constructed from a large single-cell transcriptomics compendium with each measure of association, and explored the correspondence between these networks and other biological networks, including protein–protein interaction, cellular signaling and metabolic networks. Following the intuition that measures of association that prioritize biologically meaningful relationships should also result in increased reproducibility across datasets by better discriminating signal from noise, we quantified the reproducibility of coexpression network analysis of the major endocrine cell types of the pancreas across five separate datasets. We further assessed the ability of each measure of association to permit unbiased assignment of cell types from single-cell transcriptomes. Finally, we explored the potential of single-cell transcriptomics to implicate cell-type-specific patterns of gene coexpression in human disease.

Results

To conduct a comprehensive evaluation of measures of association for single-cell transcriptomics, we assembled a large compilation of 213 scRNA-seq datasets, culled from 43 studies published between 2014 and 2018 (Supplementary Data 1). These datasets varied over several orders of magnitude in the numbers of cells and genes they profiled (Fig. 1a,b), had varying proportions of dropouts (Fig. 1c,d) and were generated using a number of different sequencing protocols (Fig. 1e), which allowed us to jointly characterize the effects of technical variables and measures of association on gene network inference.

We constructed gene coexpression networks for each dataset in our scRNA-seq compendium using 17 measures of association, analyzing a total of 3,621 networks. These measures included Pearson, Spearman and Kendall correlations; biweight midcorrelation; weighted rank correlation[17]; three distance metrics (Euclidean, Manhattan and Canberra); and the cosine similarity. In line with the recent suggestion that measures of gene co-occurrence across cell types capture biologically meaningful relationships, we also evaluated the gene co-dependency index[18], as well as the Jaccard coefficient, Dice coefficient and Hamming distance between vectors describing gene joint presence or absence across cells, at any expression level. Because estimates of gene expression derived from sequencing experiments reflect relative rather than absolute abundance[19], we also evaluated two measures of proportionality, ϕ_s and ρ_p (ref. [20]). Finally, we also evaluated mutual information[21], an information theoretic measure, and implemented a recently described

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [2]International Collaboration on Repair Discoveries (ICORD), University of British Columbia, Vancouver, Bristish Columbia, Canada. [3]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, Canada. *e-mail: michael.skinnider@msl.ubc.ca; foster@msl.ubc.ca

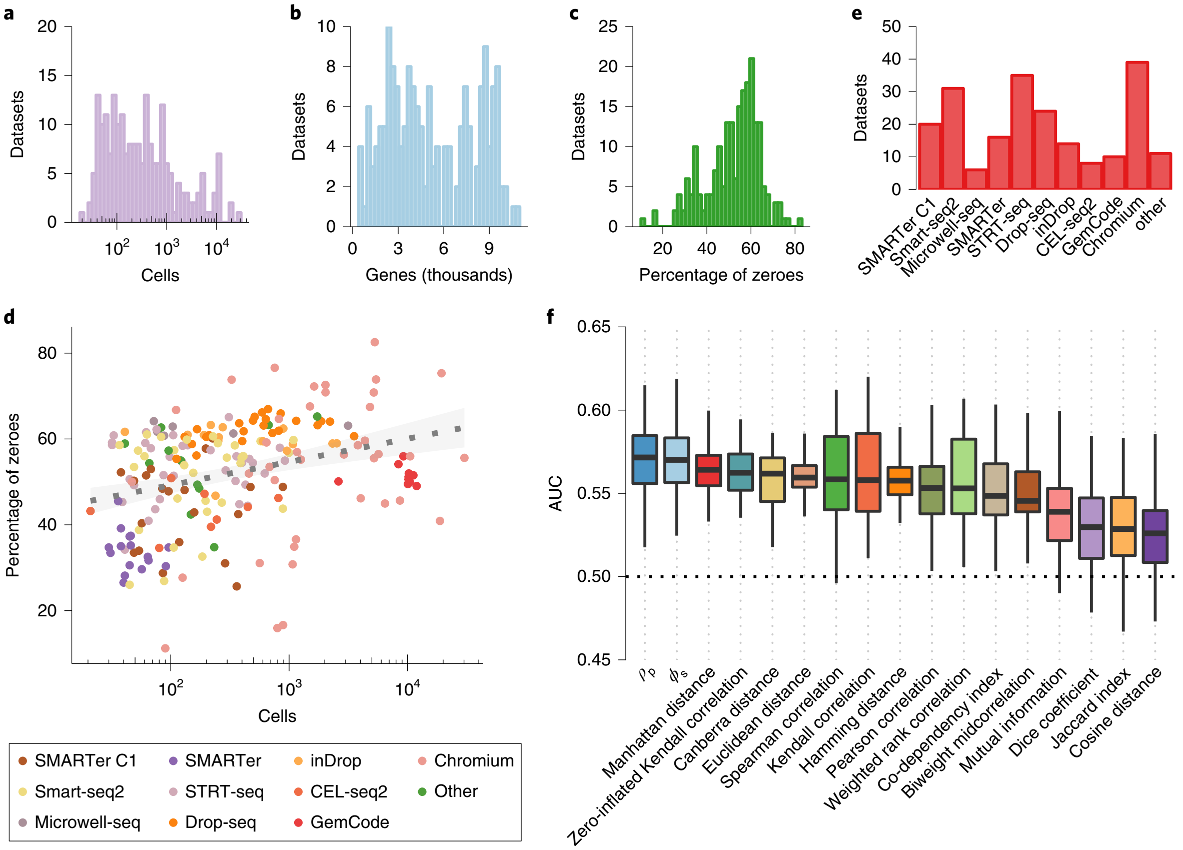

Fig. 1 | Single-cell transcriptomics compendium and functional coherence of single-cell gene coexpression networks. a–e, Properties of scRNA-seq datasets considered in this study. **a**, Number of cells in each dataset. **b**, Number of genes in each dataset, after filtering. **c**, Proportion of gene expression measurements across all cells that were zero (dropouts) in each dataset, after filtering. **d**, Relationship between the number of cells and the proportion of dropouts in each dataset. The gray dotted line shows ordinary least-squares regression. **e**, Number of datasets collected with each scRNA-seq protocol. **f**, Functional coherence of single-cell gene coexpression networks ($n = 43$ datasets, one per publication). Known gene functions were randomly withheld and predicted from the coexpression network in threefold cross-validation, and the AUC was calculated to quantify the degree to which genes with similar functions are coexpressed in networks constructed with each measure of association. Box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers).

estimator of Kendall's tau for zero-inflated continuous data[22]. Full details of our implementation and supporting source code are available in the Methods.

Functional coherence of single-cell gene coexpression networks. To quantify the functional coherence of each network, we first evaluated our ability to predict the biological properties of a gene (in this case, Gene Ontology terms) from those of its neighbors in the network[23]. We annotated each gene in the network with its known functions, then randomly withheld a subset of these labels and calculated the accuracy of gene function predictions made on the basis of coexpression alone in cross-validation (Methods). To mitigate the effects of larger studies on our results, we randomly sampled one dataset from each publication, and focused our further analysis on this subset (results from all 213 datasets are provided in Supplementary Fig. 1 and Supplementary Data 2). Across our single-cell transcriptome compendium, the two measures of proportionality, ρ_p and ϕ_s, consistently performed best, with median areas under the receiver operating characteristic curve (AUC) of 57.2% and 57.0%, respectively (Fig. 1; $P > 0.9$ for the comparison, Fisher integration of two-sided Brunner–Munzel tests). The next two top-performing metrics, the Manhattan distance and the zero-inflated Kendall correlation, were likewise not significantly different ($P = 0.089$), although both were significantly less accurate than either measure of proportionality (all $P \leq 1.9 \times 10^{-6}$). Notably, measures of gene co-occurrence performed relatively poorly, as did mutual information.

We next asked whether the functional coherence of single-cell coexpression networks varied with the number of cells profiled or the proportion of dropouts. Functional coherence was unrelated to the frequency of dropouts for all measures of association (Supplementary Fig. 2a; all $P \geq 0.10$, Spearman correlation). However, datasets with a larger number of cells consistently facilitated the reconstruction of more functionally coherent networks (Supplementary Fig. 2b, $P < 0.05$ for 7/17 measures of association after Benjamini–Hochberg correction), with the strongest effects for ρ_p, ϕ_s and the co-dependency index. This finding echoes recent analyses suggesting that, given a fixed sequencing capacity, profiling a greater number of cells at a shallower depth permits more accurate reconstruction of transcriptional programs[24]. The performance of each measure of association remained relatively stable across sequencing protocols and as a function of transcript coverage, with ρ_p, ϕ_s, zero-inflated Kendall correlation and the Manhattan distance consistently among the top-performing metrics (Supplementary Figs. 3 and 4). The choice of measure of association explained substantially more variation in the overall functional coherence of inferred networks than experimental factors did; however, among the latter, the total number of cells was more strongly associated

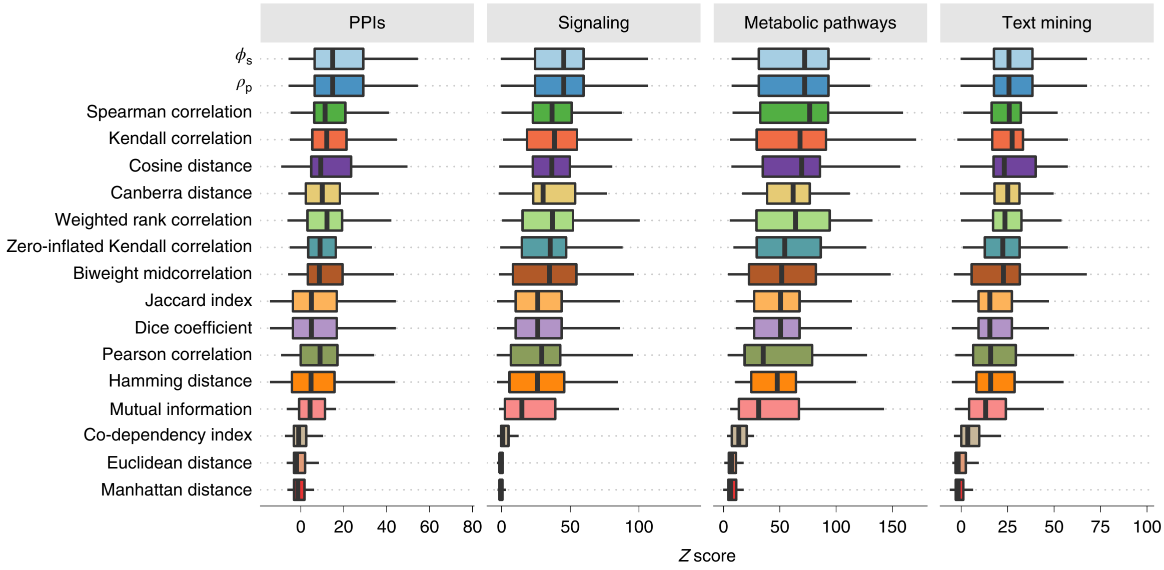

Fig. 2 | Overlap between single-cell gene coexpression networks and other biological networks. *Z* scores for each measure of association were calculated to quantify the statistical significance of the overlap between single-cell gene coexpression networks ($n = 43$ datasets, one per publication) and protein–protein interaction networks (PPIs), cellular signaling networks, metabolic pathway networks and literature co-association networks (text mining), relative to randomly rewired networks. Box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers).

with functional coherence than either biases in transcript coverage or the method of cell capture, whereas the proportion of dropouts was not significantly associated with functional coherence at all (Supplementary Fig. 5).

Convergence of macromolecular interactions. Gene coexpression networks that accurately capture cellular gene regulatory relationships should intersect with cellular networks inferred from other data, such as physical protein–protein interactions[25]. We therefore next explored the correspondence between single-cell gene coexpression networks and four other types of biological networks: protein–protein interaction networks, metabolic pathways, signaling networks and gene–gene associations inferred from text mining. We constructed unweighted gene coexpression networks by retaining the top 50,000 edges from each method, and assessed the significance of network overlap using a randomization procedure[26] (Methods). Both measures of proportionality (ρ_p and ϕ_s) yielded coexpression networks with highly significant intersections with other cellular networks, as did rank correlation coefficients (Fig. 2), and these trends were robust to the precise number of edges retained (Supplementary Fig. 6 and Supplementary Data 3). These results suggest that, in addition to ranking functionally related gene pairs above unrelated gene pairs in general, rank correlations and measures of proportionality also prioritize physically interacting proteins among the highest-ranked gene pairs.

Reproducibility. We next hypothesized that measures of association that prioritize biologically meaningful relationships should have the effect of increasing reproducibility across datasets, by better discriminating signal from noise. Consequently, we evaluated the reproducibility of gene coexpression networks inferred from each measure of association. To assess reproducibility, we analyzed five scRNA-seq datasets of alpha, beta and delta cells from the human pancreas, for a total of 30 pairwise comparisons, and quantified the degree of reproducibility of network pairs. Across all cell types, measures of proportionality yielded the most reproducible networks, both with median pairwise *Z* scores of 89.9 relative to permuted networks (Fig. 3). The Dice coefficient and Jaccard index, both of which quantify the degree to which pairs of genes are either expressed or not expressed within the same cells, likewise constructed highly reproducible networks. In contrast, coexpression networks constructed from rank correlations were less reproducible, although nearly all pairwise comparisons remained statistically significant (27/30 pairs with Bonferroni-corrected $P < 0.05$ for Spearman and Kendall correlations).

Cell-type clustering. Measures of association can also be used to define the similarities of pairs of cells on the basis of their transcriptome profiles—for instance, during unsupervised cell-type discovery, or when a cell population of interest is compared to a reference dataset of known cell types. To quantify the performance of each measure of association in identifying cell–cell relationships, we hierarchically clustered single-cell transcriptomes from seven human cell lines[27], using the adjusted Rand index to measure the correspondence between the observed clusters and the cell lines of origin. Measures of proportionality clustered cells with the greatest accuracy (Fig. 4 and Supplementary Fig. 7). The gene co-dependency index, which performed relatively poorly in coexpression network analysis, was also among the most accurate methods, whereas the zero-inflated Kendall correlation was less accurate in comparisons of cells, rather than genes. Notably, the gene co-dependency index produced the most accurate clustering when the analysis was restricted to two cell lines that were profiled in two different batches, which suggests that patterns of gene presence or absence may be more pronounced for cells of the same type across batches than the similarity of absolute expression values (Supplementary Fig. 8). We obtained similar results when using the normalized mutual information to evaluate clustering accuracy (Supplementary Fig. 9), or when applying the Louvain clustering algorithm to the shared nearest-neighbor graph[28] instead of hierarchical clustering (Supplementary Fig. 10).

Cell-type-specific disease gene networks. In bulk tissue, gene coexpression network analysis has led to insights into the pathobiology

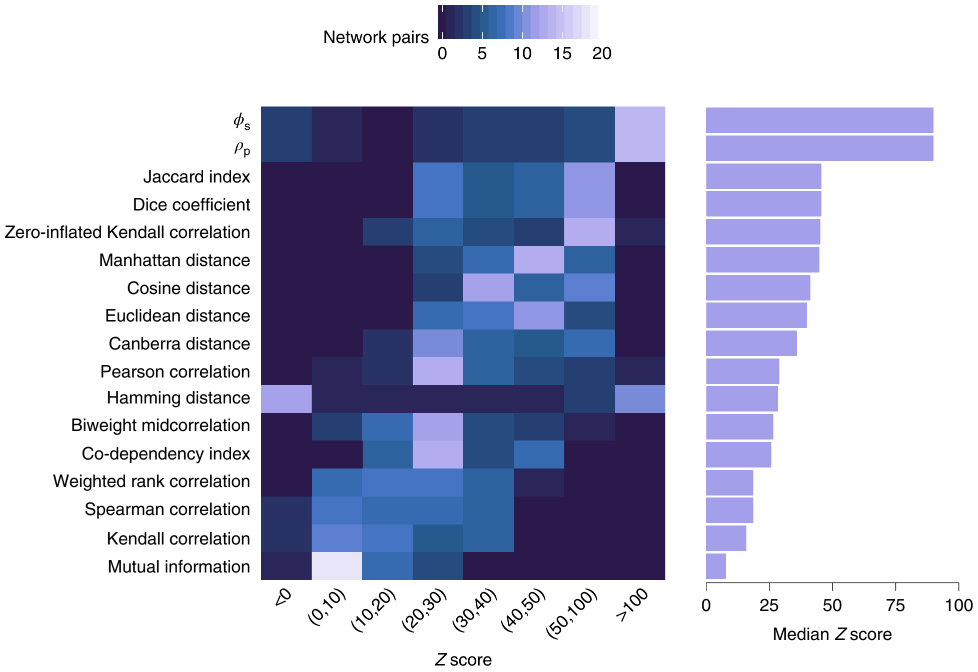

Fig. 3 | Reproducibility of single-cell gene coexpression networks of the human pancreas. Number of pairs of pancreatic coexpression networks in each bin of reproducibility *Z* score, as calculated using a permutation test, for each measure of association (left). Median reproducibility *Z* score for each measure of association across all pairs of pancreatic coexpression networks (right).

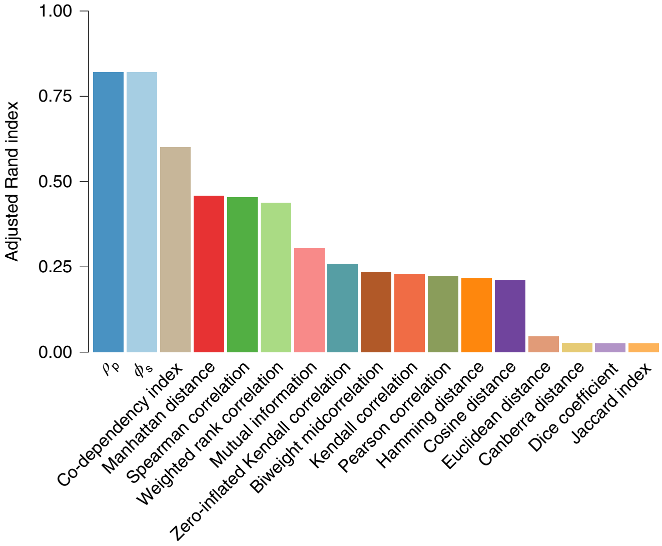

Fig. 4 | Accuracy of measures of association for clustering single-cell transcriptomes of known cell types. Hierarchical clustering of an scRNA-seq dataset composed of seven human cell lines was performed to quantify the ability of each measure of association to group single-cell transcriptomes of a common cellular origin, using the adjusted Rand index.

of neurological[29] and psychiatric[30,31] disease, and in keeping with these successes, a number of methods have been developed to identify interconnected modules of disease genes within molecular interaction networks[32–34]. The availability of cell-type-specific brain transcriptomes on a massive scale[5,35] offers an opportunity to identify cellular networks that drive disease pathogenesis at the level of individual cell types. To evaluate measures of association in this context, we constructed coexpression networks for 39 cell types in the mouse central nervous system (CNS) and asked which measures of association could most accurately predict genes implicated in neuropsychiatric disorders. Measures of proportionality again yielded the most accurate predictions, although the co-dependency index, Pearson correlation and rank correlations also performed well (Fig. 5a).

To highlight the power of cell-type-specific gene coexpression networks to identify particular cell types associated with disease etiology, we made use of an scRNA-seq atlas of the mouse CNS vasculature[36] to investigate whether genes associated with cerebrovascular disease (CVD) are selectively coexpressed in any subtype of brain vasculature cells in the CNS (Fig. 5b). Known CVD disease genes were coexpressed in cell types classically involved in CVD pathology and the integrity of the blood–brain barrier, including microvessel endothelial cells[37] and pericytes[38]. Surprisingly, however, CVD genes were also strongly coexpressed in a recently discovered subpopulation of perivascular fibroblast-like cells expressing the marker gene *Pdgfra*, which reside in the perivascular space and are proposed to be a key component of fluid transfer from the brain parenchyma to the cerebrospinal fluid[36]. The specific coexpression of CVD genes in these cells may reflect a previously unappreciated role in CVD pathophysiology.

Discussion

We report a large-scale evaluation of measures of association for single-cell transcriptomics. Across a large number of analytical tasks, two measures of proportionality, ρ_p and ϕ_s, consistently performed well, reconstructing gene networks with high functional coherence and significant overlap with other cellular interaction networks, yielding reproducible models of cellular organization across datasets generated using distinct experimental protocols and clustering transcriptomic profiles by their cell type of origin with high accuracy.

In contrast, several measures of association that are widely used in either single-cell or bulk transcriptomics, including Pearson correlation, Euclidean distance, mutual information and cosine distance, performed relatively poorly on one or more tasks. Notably,

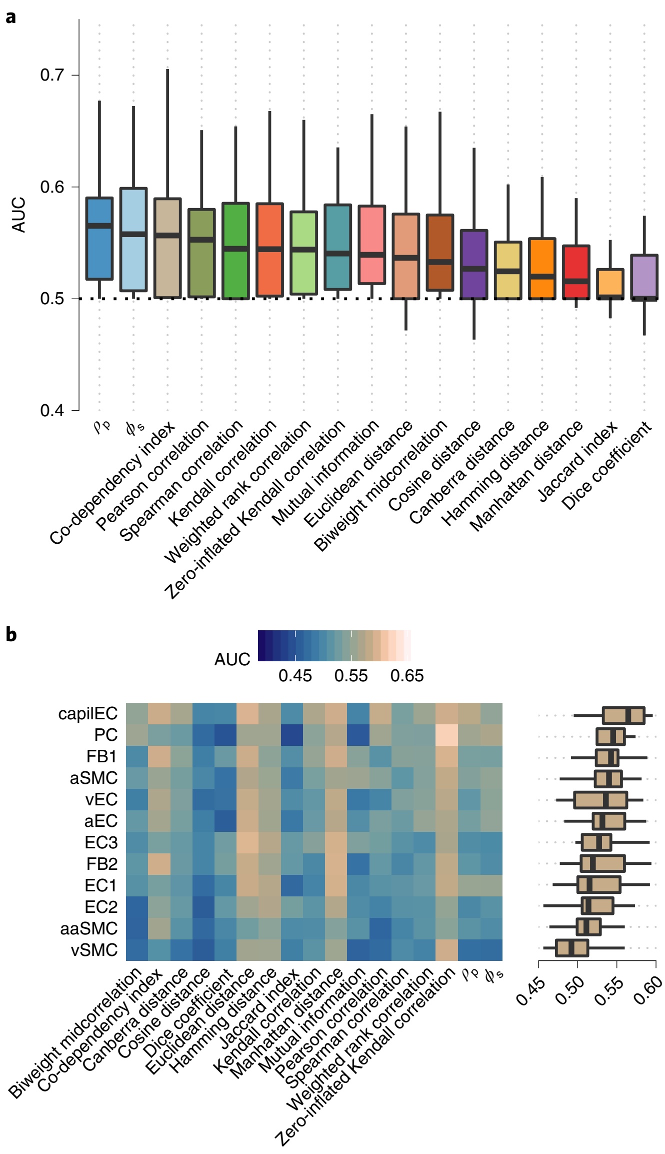

Fig. 5 | Disease gene prioritization through single-cell gene coexpression analysis. a, CNS disease gene prediction from single-cell gene coexpression networks. Known disease genes were randomly withheld and predicted from the coexpression networks of 39 CNS cell types in threefold cross-validation, and the AUC was calculated to quantify the degree to which genes implicated in the same neuropsychiatric diseases are coexpressed in networks constructed with each measure of association. **b**, Cell-type-specific disease gene coexpression in the CNS. Coexpression network connectivity of genes implicated in CVD in cell types of the mouse brain vasculature (left). Distribution of AUCs for each of 12 cell types (right). Box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers).

two measures of association designed specifically for single-cell transcriptomics, or zero-inflated continuous data more generally, performed well on only a subset of tasks. The gene co-dependency index[18] accurately predicted disease genes and clustered cell types, particularly across batches, but performed more poorly in gene network analyses. The zero-inflated Kendall correlation[22] yielded reproducible and functionally coherent gene coexpression networks, but performed only moderately well on the network overlap and clustering tasks.

The sparsity and overdispersion of scRNA-seq data pose a considerable challenge to robust network inference from single-cell transcriptomics, relative to bulk data[39]. Substitution of measures of proportionality for more broadly used measures of association yielded gene coexpression networks with a degree of functional coherence similar to that which has previously been reported for bulk RNA-seq[40]. However, in comparison with approaches for tissue- or cell-type-specific network reconstruction that integrate much larger compendia of data[41], the absolute degree of predictive power remained relatively low. Integration of single-cell gene expression with additional molecular phenotypes that can now be measured in high throughput, including the epigenome[42] and proteome[43], could provide a means to further increase the accuracy of cell-type-specific network inference. Furthermore, it is conceivable that only a subset of physiologically relevant functions can be accurately predicted for each cell type. It is also noteworthy in this respect that available 'gold standards' for evaluating network reconstruction, such as gene function annotations and physical protein–protein interactions, are based largely on biological relationships identified in bulk data, which raises the possibility that these are not reflected at the single-cell level.

The strong performance of ρ_p and ϕ_s can be rationalized on the basis that scRNA-seq captures only a small proportion of messenger RNA in a cell. Consequently, gene expression estimates must be interpreted as relative measures of abundance. In the setting of bulk RNA-seq, correlations between relative abundances can lead to conclusions at odds with those drawn from absolute quantifications[19]. Our analysis suggests that substituting conventional measures of association with measures of proportionality could lead to an increase in the accuracy, and therefore interpretability, of diverse computational analyses for scRNA-seq beyond those evaluated here, such as the reconstruction of intercellular signaling networks on the basis of patterns of receptor and ligand expression[44–46], or the analysis of temporally coupled gene expression measurements[47]. The computational time required to construct transcriptome-wide coexpression networks using measures of proportionality was comparable to that for other measures of association (Supplementary Fig. 11), which suggests that these methods are computationally efficient enough to scale with increasing numbers of cells profiled in individual experiments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0372-4.

Received: 1 October 2018; Accepted: 5 March 2019;
Published online: 08 April 2019

References

1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
2. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
3. Mahata, B. et al. Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep.* **7**, 1130–1142 (2014).
4. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
5. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
6. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
7. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
8. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies cell-type-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

9. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
10. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
11. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
12. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
13. La Manno, G. et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).
14. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).
15. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
16. Gerber, T. et al. Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. *Science* **362**, eaaq0681 (2018).
17. Zar, J. H. *Biostatistical Analysis* 5th edn (Prentice-Hall/Pearson, 2010).
18. Mohammadi, S., Davila-Velderrain, J., Kellis, M. & Grama, A. DECODE-ing sparsity patterns in single-cell RNA-seq. Preprint at https://www.biorxiv.org/content/10.1101/241646v2 (2018).
19. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11**, e1004075 (2015).
20. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* **7**, 16252 (2017).
21. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328 (2012).
22. Pimentel, R. S., Niewiadomska-Bugaj, M. & Wang, J.-C. Association of zero-inflated continuous variables. *Stat. Probabil. Lett.* **96**, 61–67 (2015).
23. Ballouz, S., Weber, M., Pavlidis, P. & Gillis, J. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics* **33**, 612–614 (2017).
24. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
25. Ramani, A. K. et al. A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. *Mol. Syst. Biol.* **4**, 180 (2008).
26. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
27. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
28. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
29. Zhang, B. et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
30. Parikshak, N. N. et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).
31. Gulsuner, S. et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
32. Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
33. Huang, J. K. et al. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **6**, 484–495 (2018).
34. Choobdar, S. et al. Open community challenge reveals molecular network modules with key roles in diseases. Preprint at https://www.biorxiv.org/content/10.1101/265553v1 (2018).
35. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
36. Vanlandewijck, M. et al. A molecular atlas of cell types and zonation in the brain vasculature. *Nature* **554**, 475–480 (2018).
37. Zhao, Z., Nelson, A. R., Betsholtz, C. & Zlokovic, B. V. Establishment and dysfunction of the blood-brain barrier. *Cell* **163**, 1064–1078 (2015).
38. Lindahl, P., Johansson, B. R., Levéen, P. & Betsholtz, C. Pericyte loss and microaneurysm formation in PDGF-B-deficient mice. *Science* **277**, 242–245 (1997).
39. Chen, S. & Mar, J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232 (2018).
40. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).
41. Yao, V. et al. An integrative tissue-network approach to identify and test human disease genes. *Nat. Biotechnol.* **36**, 1091–1099 (2018).
42. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
43. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).
44. Camp, J. G. et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).
45. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
46. Cohen, M. et al. Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell* **175**, 1031–1044 (2018).
47. Qiu, X. et al. Towards inferring causal gene regulatory networks from single cell expression measurements. Preprint at https://www.biorxiv.org/content/10.1101/426981v1 (2018).

Acknowledgements

We thank E. Su and M. Hirst for computational assistance and P. Pavlidis for comments on an earlier version of the manuscript. This work was supported by Genome Canada and Genome British Columbia (to L.J.F.; project 214PRO) and enabled in part by support provided by WestGrid and Compute Canada. M.A.S. is supported by a CIHR Vanier Canada Graduate Scholarship, an Izaak Walton Killam Memorial Pre-Doctoral Fellowship, a UBC Four Year Fellowship and a Vancouver Coastal Health–CIHR–UBC MD/PhD Studentship. J.W.S. is supported by an Izaak Walton Killam Memorial Pre-Doctoral Fellowship, a UBC Four Year Fellowship and a Vancouver Coastal Health–CIHR–UBC MD/PhD Studentship.

Author contributions

M.A.S., J.W.S. and L.J.F. designed experiments. M.A.S. and J.W.S. performed experiments. M.A.S. wrote the first draft of the manuscript, which was edited by J.W.S. and L.J.F.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0372-4.

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

Correspondence and requests for materials should be addressed to M.A.S. or L.J.F.

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

Methods

Datasets. We collected a total of 213 scRNA-seq transcriptomic datasets from 43 different publications (Supplementary Data 1). Of these, 164 datasets were obtained from the Gene Expression Omnibus, 10 were obtained from the 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets) and the remaining 39 were obtained from http://mousebrain.org (ref. [35]) (taxonomy level 3, file 'l6_r3.loom') and exported by cell type from the loom file format to CSV using the 'loompy' Python package (http://loompy.org). Whenever permitted by sample size, we separated different cell types according to the authors' annotations to construct cell-type-specific coexpression networks. Human and mouse gene identifiers were mapped to Ensembl accessions using Bioconductor (packages org.Hs.eg.db and org.Mm.eg.db). Datasets were filtered to exclude non-protein-coding genes, as annotated in v.91 of the Ensembl human and mouse genome annotations. Datasets obtained from http://mousebrain.org were subsequently filtered to include only the 2,000 genes expressed in the largest number of cells, whereas datasets obtained from GEO and 10X Genomics were filtered to exclude genes with an expression of zero in more than 80% of samples. We additionally evaluated the effect of thresholds between 50% and 95%, and found that our results were largely similar across thresholds (Supplementary Fig. 12), but that rank correlations generally performed well only in datasets with a low proportion of dropouts, whereas measures of gene co-occurrence displayed the opposite trend (Supplementary Fig. 13).

Measures of association. We evaluated 17 measures of association in the context of single-cell transcriptomics, ranging from measures that are broadly used in analysis of gene expression data to measures that have not, to our knowledge, previously been applied in this setting. We provide a brief overview of each measure of association in the Supplementary Note. Pearson and Spearman correlations were calculated using the base R 'stats' package, as were the Euclidean, Canberra and Manhattan distances. Mutual information and biweight midcorrelation matrices were calculated using functions from the 'WGCNA' R package[48] (v.1.66). Kendall correlation matrices were calculated using the fast Kendall's tau function from the R package 'pcaPP' (v.1.9–73). Cosine distances were calculated in the 'lsa' package (v.0.73.1). The Sørensen–Dice coefficient was calculated in the 'arules' package[49] (v.1.6–1). Both measures of proportionality, $\rho_{\rm p}$ and $\phi_{\rm s}$, were calculated using the 'propr' package[20] (v.4.0.0). We additionally implemented custom R code to calculate the Hamming distance and Jaccard index on the basis of gene presence or absence; a weighted rank correlation[17]; the gene co-dependency index, which models the probability of gene co-occurrence under the binomial distribution[18]; and a recently described estimator of Kendall's tau for zero-inflated data[22]. Matrices that were not naturally bounded by the range (−1, 1) were scaled to this range before further analysis. We implemented a single interface to access all 17 measures of association in the R package 'dismay' (Supplementary Software 1).

Functional coherence. We assessed the functional coherence of each network by evaluating the degree to which the functional properties of a gene (in this case, Gene Ontology terms) can be predicted from those of its neighbors in the network. In this analysis, networks are determined to have greater functional coherence if a gene's functional annotations can be more accurately predicted by those of its neighbors, based on the principle of guilt by association[40]. Gene Ontology annotations were obtained from the UniProt-GOA database[50], and the complete ontology was retrieved from the Gene Ontology website (file 'go-basic.obo'), both downloaded 27 April 2017. UniProt accessions were mapped to Ensembl gene identifiers. Annotations supported by one or more of the evidence codes ND (no biological data available), IEA (inferred from electronic annotation), IPI (inferred from physical interaction) and NAS (non-traceable author statement), or associated with the qualifier NOT, were removed, and the remaining annotations were propagated up the ontology. Functional connectivity analyses were performed using EGAD[23] (v.1.8.0), which uses a neighbor-voting algorithm to predict the functions of left-out genes in cross-validation. We constructed dense weighted gene coexpression networks, using each measure of association in turn to assign a weight to each gene–gene pair, using code adapted from the EGAD function 'build_coexp_network'. This approach ensures that for a given dataset, each measure of association produces the same number of connections, differing only in how they are ranked. These ranks are used to predict the functions of the subset of genes with annotations withheld. Threefold cross-validation was performed, and the mean area under the receiver operating characteristic curve was retained for each gene ontology term. Statistical comparisons of functional connectivity were performed using the Brunner–Munzel test, a non-parametric test robust to differences in the shape of distributions, as implemented in the R package 'lawstat' (v.3.2). Univariate analyses presented in Supplementary Fig. 5 were performed on the subset of protocols used in at least two different publications.

Molecular interaction network overlap. Human protein–protein interaction and signaling networks were obtained from HIPPIE[51] and OmniPath[52], respectively, and mapped to mouse using one-to-one orthologs downloaded from Ensembl BioMart (downloaded 7 April 2018). We obtained human and mouse metabolic pathways from Reactome[53] and converted them into a metabolic pathway co-membership network by linking each pair of proteins that was present in a common pathway. We constructed human and mouse literature co-occurrence networks by filtering the STRING interaction database on the basis of the text-mining channel, requiring a minimum score of 500. For quantification of the overlap between coexpression networks and each macromolecular interaction network, the top-ranked 50,000 edges from each dense coexpression matrix constructed in the functional coherence analysis were retained to construct a sparse unweighted coexpression network; we additionally performed the same analysis with the top 20,000 or 100,000 edges retained from each dense coexpression network. Each target network was rewired 100 times using a degree-preserving algorithm[26], as implemented in the 'igraph' R package (v.1.2.2), with the number of iterations set to ten times the number of edges in the network, and the significance of the coexpression network overlap was assessed on the basis of the Z score for the number of intersecting edges relative to the randomized networks.

Reproducibility. To assess reproducibility, we extracted cell-type-specific coexpression networks for alpha, beta and delta cells from five scRNA-seq studies of the human pancreas[54–58] (accessions E-MTAB-5061, GSE84133, GSE81547, GSE81608 and GSE85241). For each pair of coexpression networks derived from the same cell type, we assessed reproducibility by using a permutation test in which we randomly permuted the rows and columns of the first distance matrix using the 'vegan' R package[59] (v.2.5–2), and calculating the Z score of the Spearman correlation between the two matrices relative to permuted matrices[60].

Cell clustering. Cell-type clustering accuracy was evaluated using an scRNA-seq dataset of 561 cells derived from seven cell lines, two of which were sequenced in two batches[27], in which the ground truth (that is, the cell line of origin) was known for each transcriptome profile. We additionally calculated clustering accuracy for distance matrices limited to the two cell lines that were sequenced in two separate batches, to specifically evaluate the impact of batch effects. Clustering was performed using two different methods: hierarchical clustering and Louvain clustering of the shared nearest-neighbors graph. The former provides a broadly accurate[61] and parameter-free benchmark (apart from the number of clusters, which is known a priori from the experiment design in this case), whereas the second method more closely reflects approaches adopted in widely used scRNA-seq software[28]. Hierarchical clustering was implemented with the base R function 'hclust', with the number of clusters set to seven (two in the batch experiments). The shared nearest-neighbor graph was calculated using code adapted from the 'scran' R package (v.1.10.1)[62], with Louvain clustering performed in the 'igraph' R package (v.1.2.2)[63], and the parameter k in k-nearest-neighbor graph construction in the range (2, 5, 10, 20, 50). Accuracy was quantified using the adjusted Rand index and normalized mutual information, calculated using the R packages 'mclust' (v.5.4.1) and 'ClusterR' (v.1.1.6), respectively. Dendrograms were visualized with the R package 'ggtree' (v.1.14.4)[64].

Disease gene analysis. We obtained disease genes for neuropsychiatric disorders from Phenopedia[65] by obtaining all children of the term D001523 in the MeSH hierarchy. Disease gene connectivity in cell-type-specific coexpression networks was evaluated using EGAD[23], as described above. CVD genes were obtained from Phenopedia using the NCI Metathesaurus term C0007820.

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

Data availability

The data that support the findings of this study are available from the following GitHub repository: https://github.com/skinnider/SCT-MoA. Raw data are available from the Gene Expression Omnibus, http://mousebrain.org, or https://support.10xgenomics.com, as detailed in the Methods; dataset identifiers are provided in Supplementary Data 1.

Code availability

The 'dismay' R package is available as Supplementary Software 1 and from the following GitHub repository: https://github.com/skinnider/dismay. R code used to reproduce the analysis and figures is available from the following GitHub repository: https://github.com/skinnider/SCT-MoA.

References

48. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
49. Hahsler, M., Chelluboina, S., Hornik, K. & Buchta, C. The arules R-Package ecosystem: analyzing interesting patterns from large transaction datasets. *J. Mach. Learn. Res.* **12**, 2021–2025 (2011).
50. Dimmer, E. C. et al. The UniProt-GO annotation database in 2011. *Nucleic Acids Res.* **40**, D565–D570 (2012).
51. Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* **45**, D408–D414 (2017).

52. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
53. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
54. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394 (2016).
55. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
56. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330 (2017).
57. Xin, Y. et al. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* **24**, 608–615 (2016).
58. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
59. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
60. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
61. Wiwie, C., Baumbach, J. & Röttger, R. Comparing the performance of biomedical clustering methods. *Nat. Methods* **12**, 1033–1038 (2015).
62. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
63. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* **1695**, 1–9 (2006).
64. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
65. Yu, W., Clyne, M., Khoury, M. J. & Gwinn, M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* **26**, 145–146 (2010).

nature research

Corresponding author(s): Michael Skinnider, Leonard Foster

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☒	□	An indication of 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 statistics 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
☒	□	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The 'dismay' R package is available as Supplementary Software 1.
Data analysis	R code used to reproduce the analysis and figures is available from the GitHub repository accompanying the article at https://github.com/skinnider/SCT-MoA. This code makes use of the following R packages: \`WGCNA\` (version 1.66); \`pcaPP\` (version 1.9-73); \`lsa\` (version 0.73.1); \`arules\` (version 1.6-1); \`propr\` (version 4.0.0); \`EGAD\` (version 1.8.0); \`lawstat\` (version 3.2); \`igraph\` (version 1.2.2); \`vegan\` (version 2.5-2); \`scran\` (version 1.10.1); \`mclust\` (version 5.4.1); \`ClusterR\` (version 1.1.6); and \`ggtree\` (1.14.4).

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 upon request. 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

The data that support the findings of this study are available from the GitHub repository accompanying the article at https://github.com/skinnider/SCT-MoA.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	Sample sizes were determined by the original authors of the studies whose data was reanalyzed here.
Data exclusions	Data were excluded from the analysis when the authors provided cell type classifications and insufficient cells of a given type were available to perform coexpression network analysis.
Replication	Coexpression network analyses were performed for over 200 different datasets. Conclusions are based on aggregate trends over the entire dataset compendium. No further attempts at replication were performed.
Randomization	Randomization was not performed as the study involved reanalysis of public data.
Blinding	Blinding was not performed as the study involved reanalysis of public data.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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