Supplementary discussion

Unlike other tools which identify cell type specific markers, e.g. e.g. Seurat v3, Comet and Panglao, scfind is distinct in five ways: 1) it adopts a novel compression algorithm for fast data retrieval in very large datasets. 2) it adopts a novel gene set optimisation routine. 3) it has the ability to conduct searches of cell type specific markers in a multi-omic manner involving different omes simultaneously. 4) it retains single cell resolution for in silico cell gating with logical operators. 5) It is the first single cell analysis tool that adapts natural language technique word2vec combined with resources such as the gene ontology annotation, the GWAS catalog and the entire collection of PubMed abstracts for biologically and clinically relevant free text searches.

One of the important features compared to existing interactive scRNAseq databases is that scfind retains information about individual cells and not just pre-defined clusters. Since scfind is freely available as an R package as part of Bioconductor with source code available under the MIT licence at https://github.com/hemberg-lab/scfind, it is easy for users to build their own references and carry out searches based on groupings other than the ones provided by the original authors. Because of its performance, scfind can be used for tasks that otherwise would have been prohibitive, e.g. running searches of all genes to be able to identify marker genes, similar genes, housekeeping genes, and cell type specific genes. These global evaluations can be carried out in seconds, making it possible to continuously update information on expression profiles, even for very large collections of data.

Apart from the demonstrated use cases, scfind's free text search feature provides a possibility of enhancing biomedical research. For instance, the immunosuppressant dexamethasone has recently been discovered to be effective for treating COVID-19 patients who are under critical conditions [59]. With the recently published Lung Cell Atlas [60] as well as the single cell datasets generated from the healthy and patient donors on the COVID-19 Cell Atlas (https://www.covid19cellatlas.org/) [61], one could easily identify the cells in the lung that are response to pathways triggered by dexamethasone in both datasets simultaneously. With the function `findSimilarGenes()`, researchers are allowed to identify not only genes but also the corresponding transcription factors that share similar expression patterns with the genes of interest.

Supplementary tables

These are large and provided as separate files.

Supplementary Table 1: Precision, recall and F1 scores for all cell types in the atlases considered.
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S1.tsv

Supplementary Table 2: Information about the total number of marker genes and the precision and F1 scores that they provide for each cell type.
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S2.tsv

Supplementary Table 3: Cell type specificity for the genes found in the MCA and the two Tabula Muris datasets.
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S3.tsv

Supplementary Table 4: Number of maximal marker genes for each cell type in the MCA and the two Tabula Muris datasets.
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S4.tsv

Supplementary Table 5: Number of cell type specific genes for each cell type in the MCA and the two Tabula Muris datasets.
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S5.tsv

Supplementary Table 6: Best matches of the TM, FACS dataset from queries generated by sample variants from the index created from PubTator.
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S6.xlsx

Supplementary Table 7: Best matches of the TM, FACS dataset from queries generated by sample diseases names/MeSH/OMIM IDs
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S7.xlsx

Supplementary Table 8: Best matches of the TM, FACS dataset from queries generated by sample chemical names and their corresponding IDs
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S8.xlsx

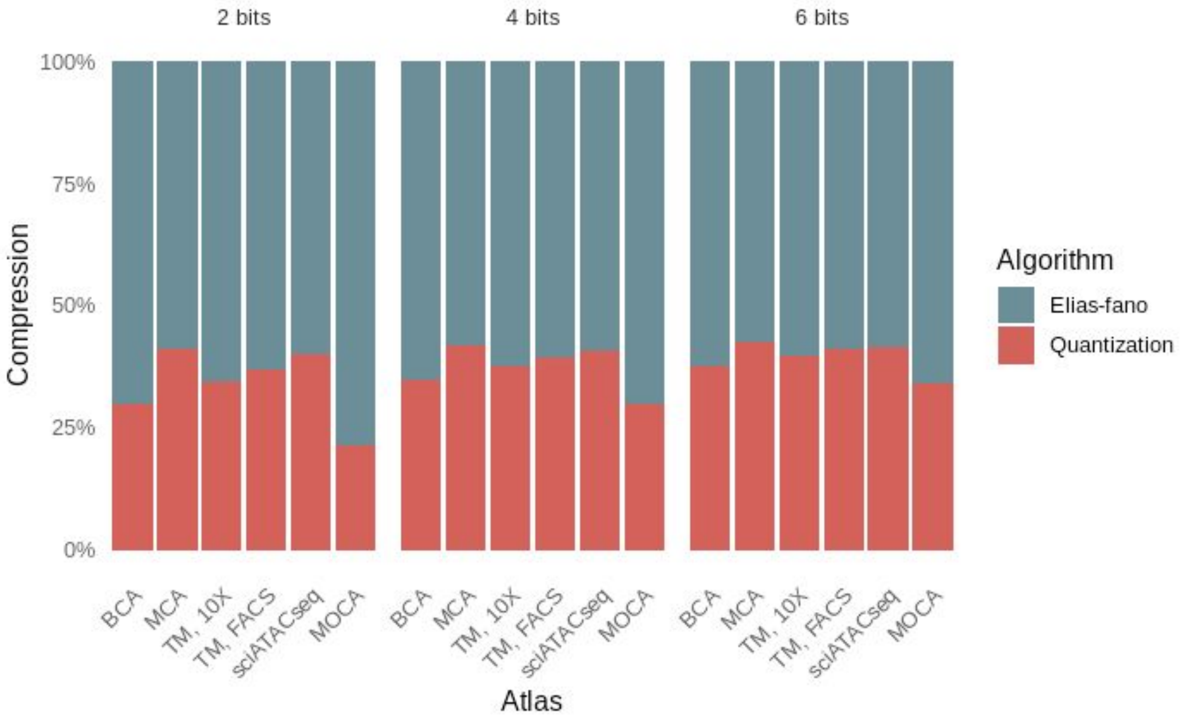
Supplementary Table 9: Best matches of the TM, FACS dataset from queries generated by sample phrases from the dictionary from the PubMed
https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S9.xlsx

Supplementary Table 10: Cell type specificity of super enhancers https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S10.tsv

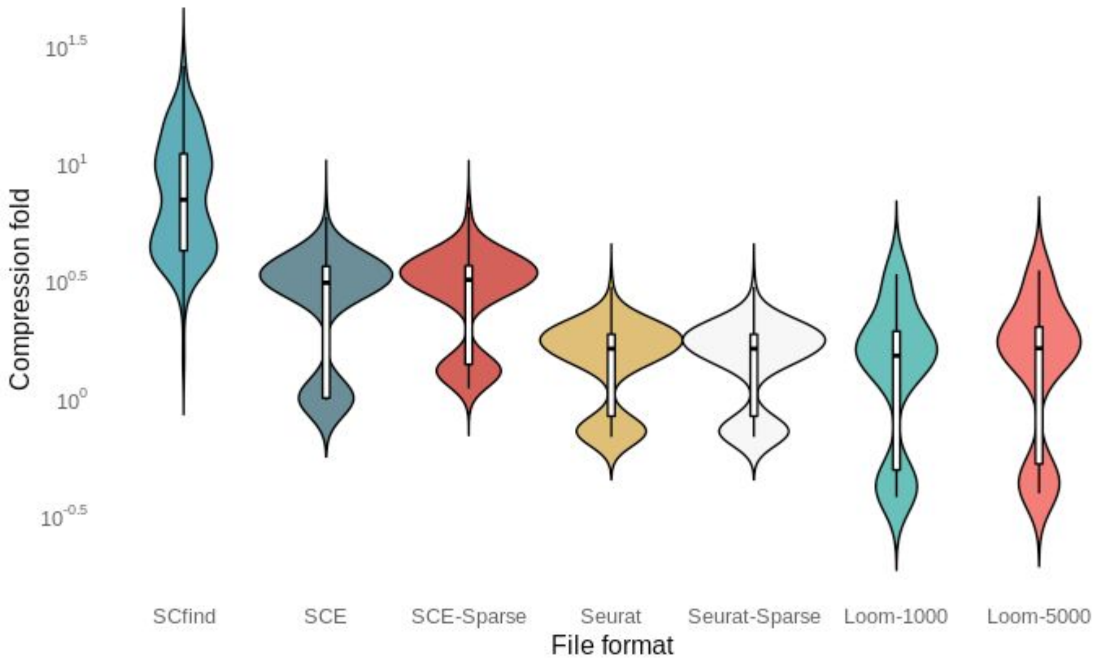
Supplementary Table 11: Cell type specific enhancer-gene pairs https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S11.tsv

Supplementary Table 12: Top 20 and 30 marker genes in the 3 batch correction methods https://raw.githubusercontent.com/hemberg-lab/scfind-paper-figures/master/data/S12.tsv

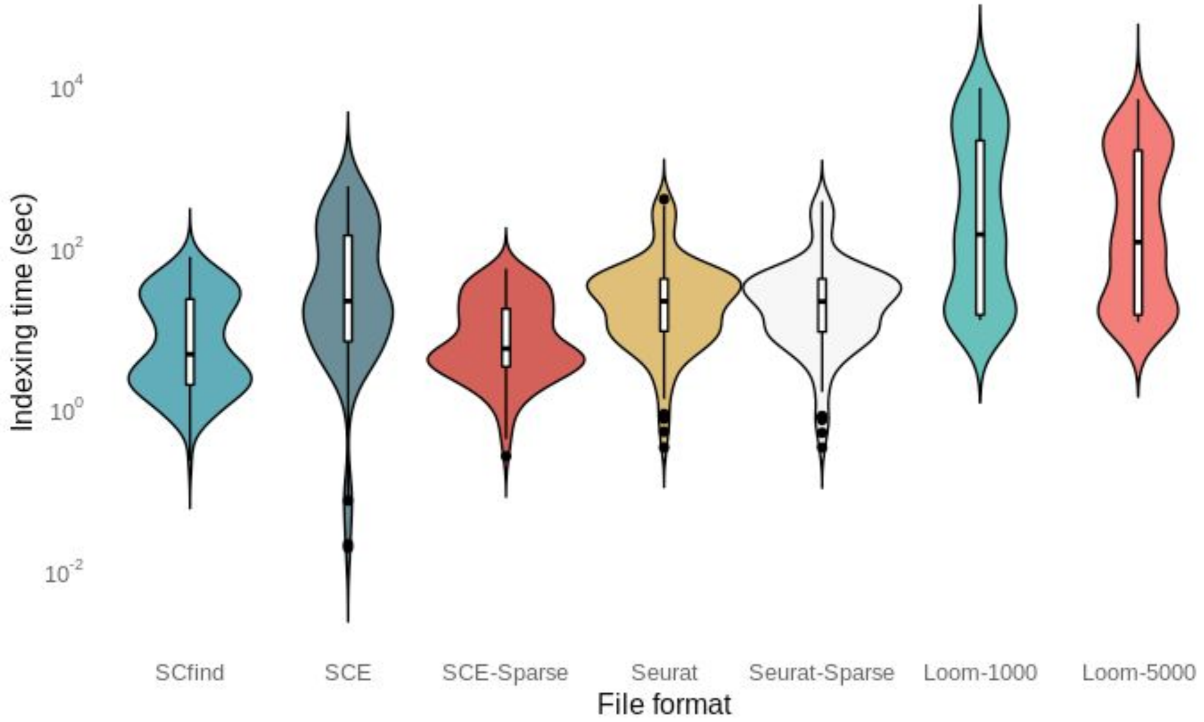
Supplementary figures



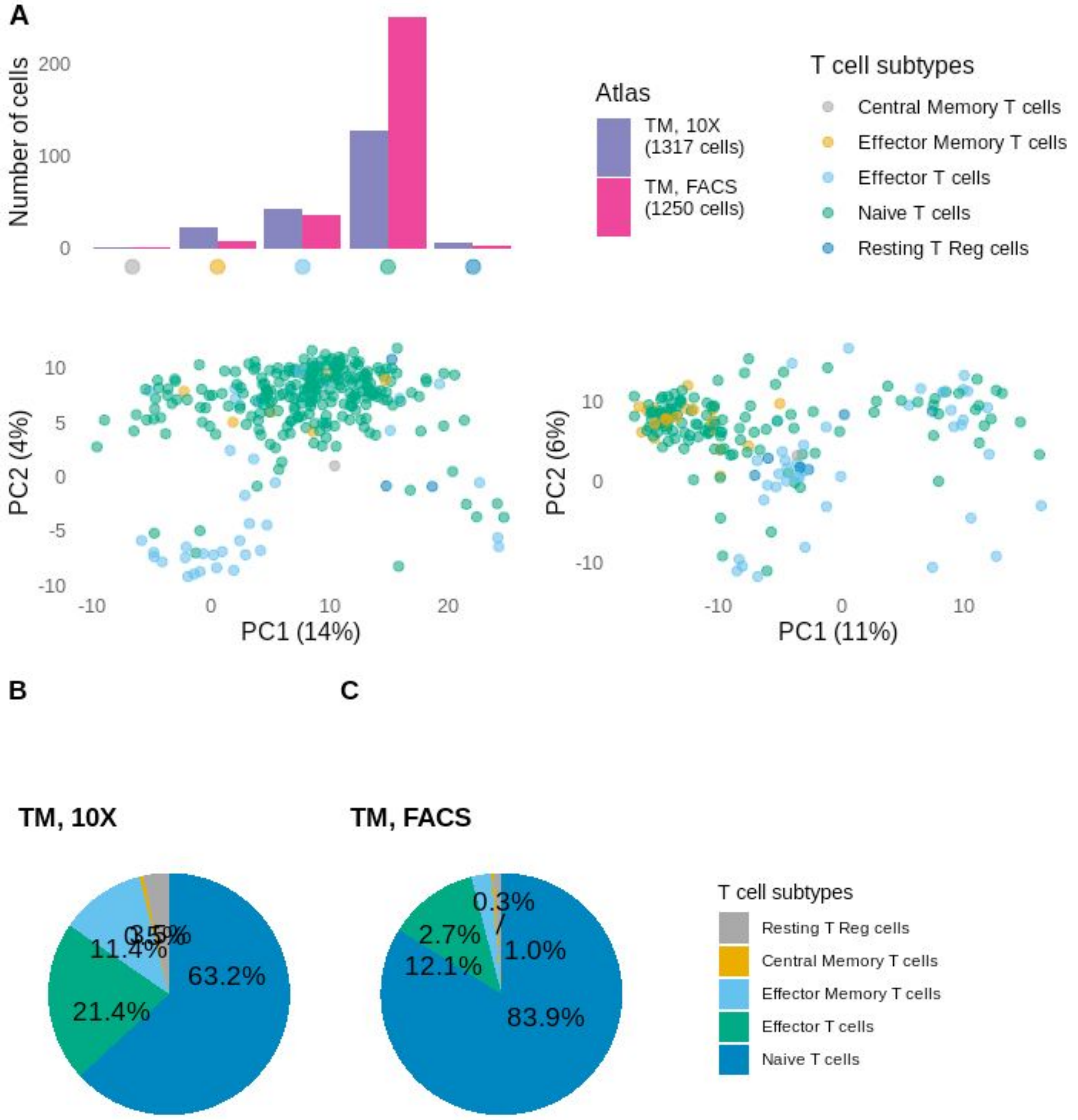
Supplementary Figure 1: Relative contribution to compression from Elias-Fano coding and quantization.



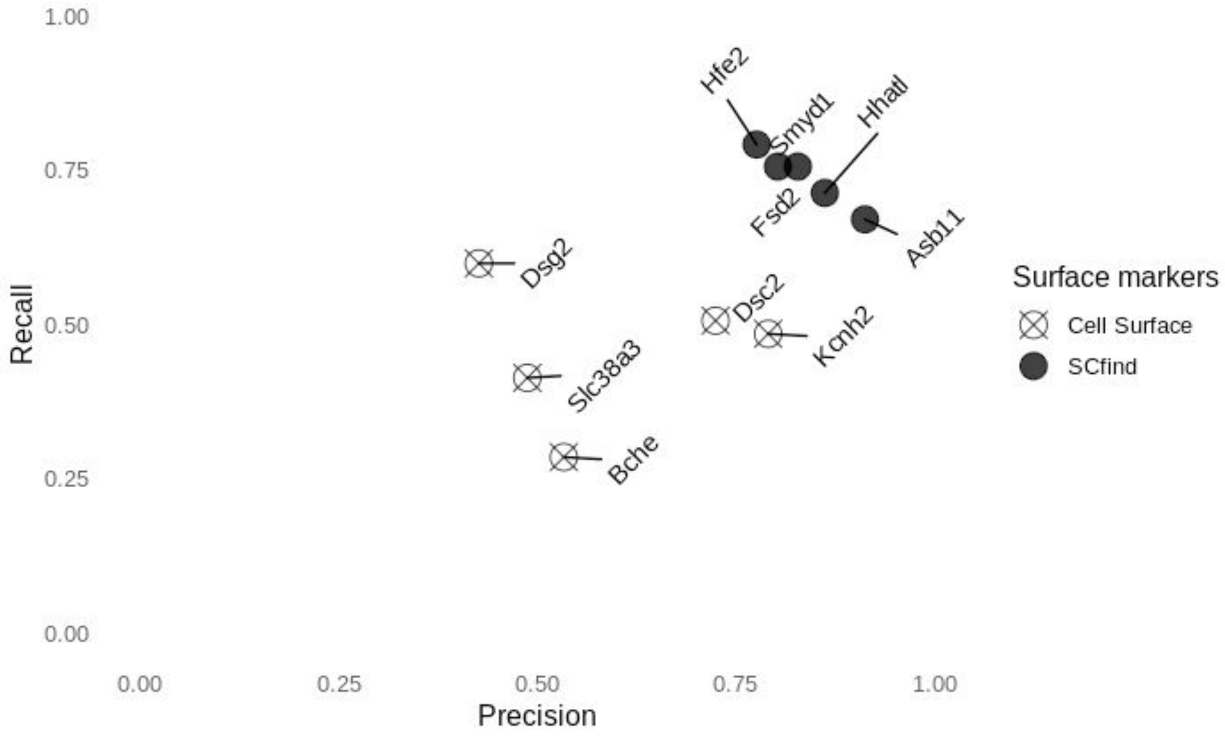
Supplementary Figure 2: Compression ratios for different file formats, the violin plots represent the density of folds relative to the uncompressed expression matrix for all tissues (n=132 tissues) in the six atlases in figure 1. Violin plots show the density (width), median (center line), interquartile range (hinges) and 1.5 times the interquartile range (adjacent lines); outlier data beyond this range are plotted as individual points.



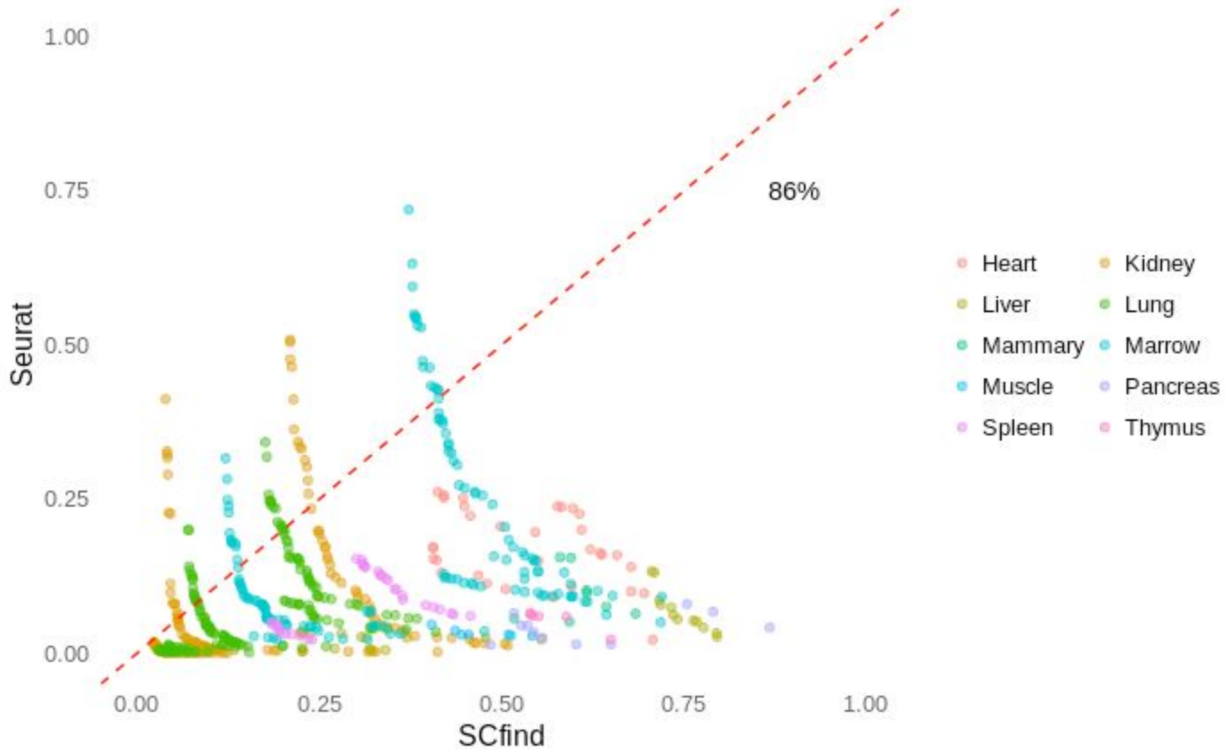
Supplementary Figure 3: Build times for scfind indexes and other file formats, the violin plots represent the indexing time for the six atlases (n=132 tissues) with different file formats in figure 1. Violin plots show the density (width), median (center line), interquartile range (hinges) and 1.5 times the interquartile range (adjacent lines); outlier data beyond this range are plotted as individual points.



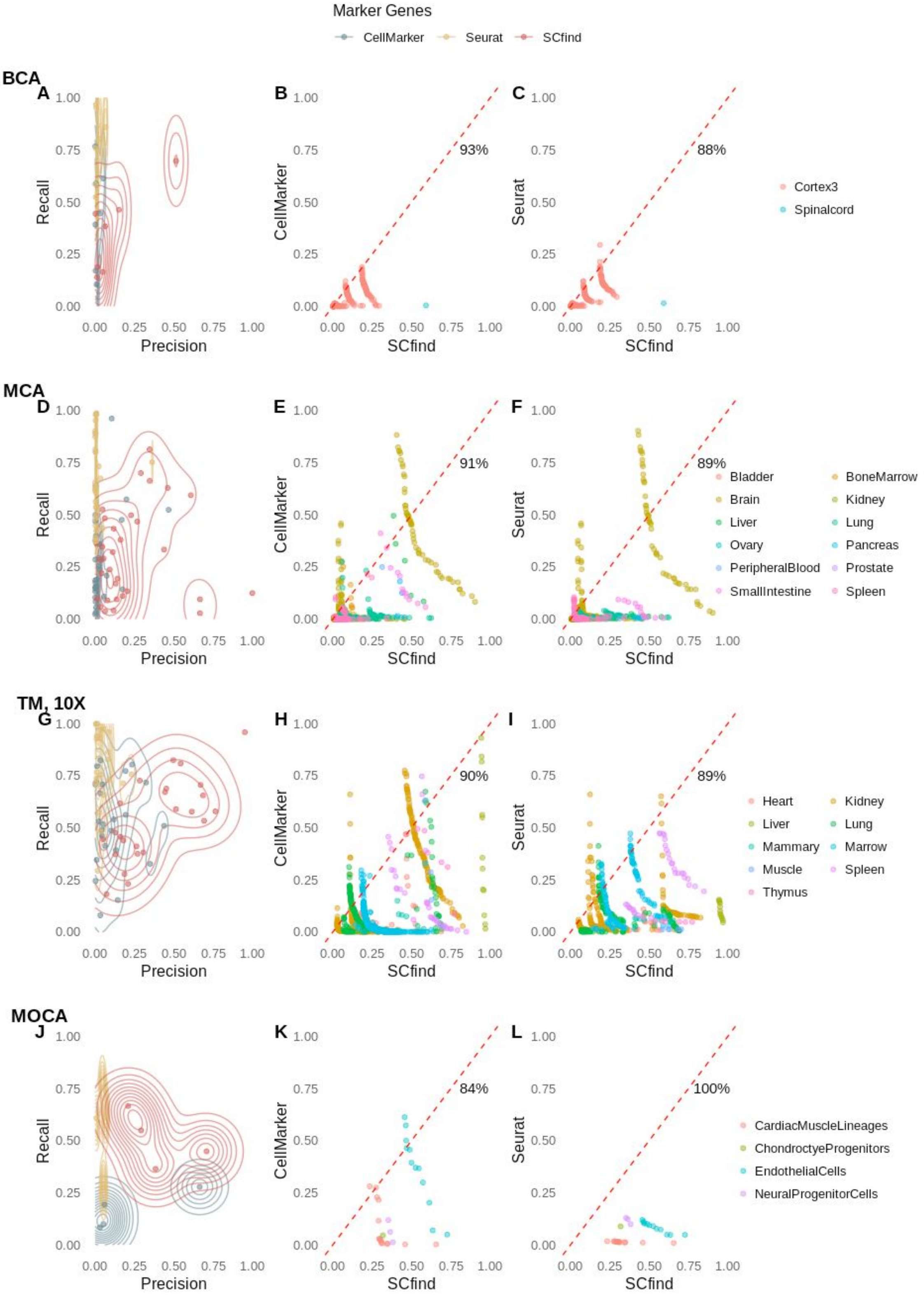
Supplementary Figure 4: (a) Number of cells found for subsets of T-cells in the thymus based on combinations of *Il2ra*, *Ptprc*, *Il7r* and *Ctla4* (top) and PCA projection of the T cells from the thymus shows good separation between naive T cells and resting T regulatory cells for both TM FACS (bottom, left) and 10X (bottom, right). (b-c) Percentage of T cell subtypes in thymus tissue of the Tabula Muris datasets



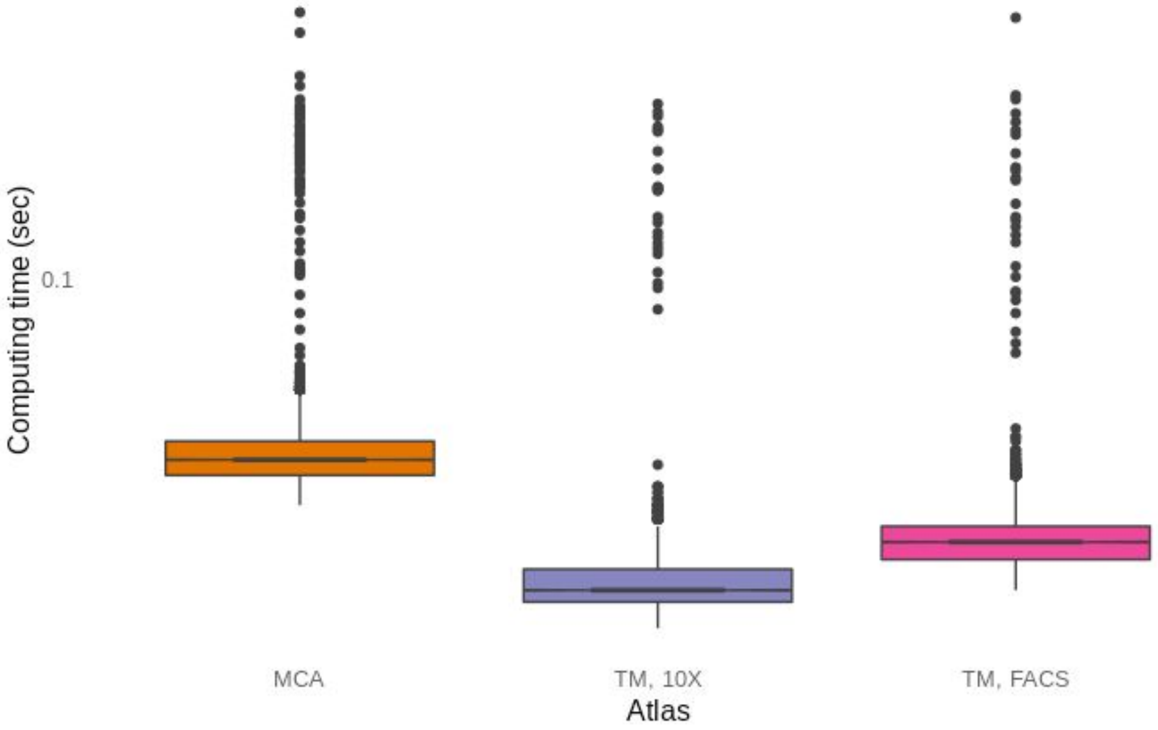
Supplementary Figure 5: Precision and recall for the five best cardiomyocyte surface markers



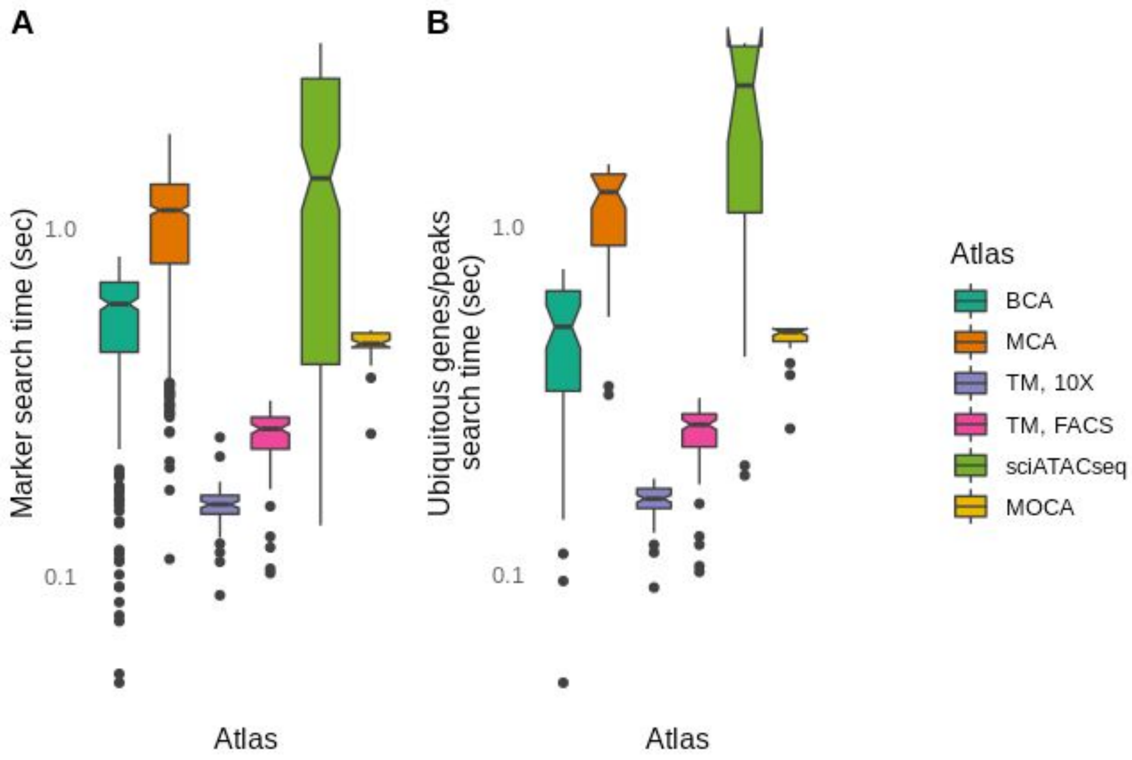
Supplementary Figure 6: Distribution of F1 scores (harmonic mean of the precision and recall scores from Figure 2B) of marker genes identified by Seurat and scfind for the Tabula Muris FACS dataset



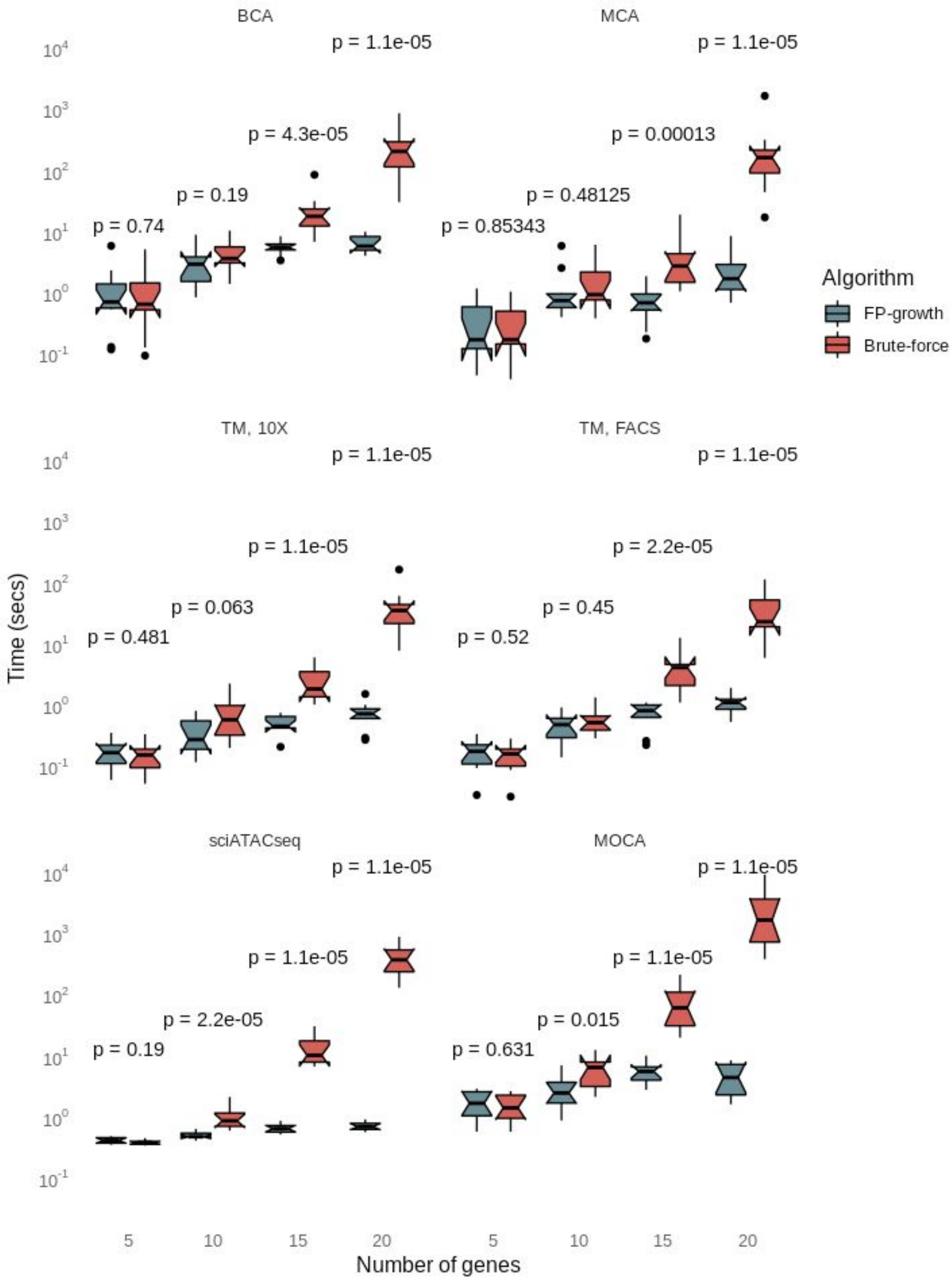
Supplementary Figure 7: A comprehensive comparison of all matched cell types in the CellMarkers database and cell type specific markers identified by Seurat against all five atlases



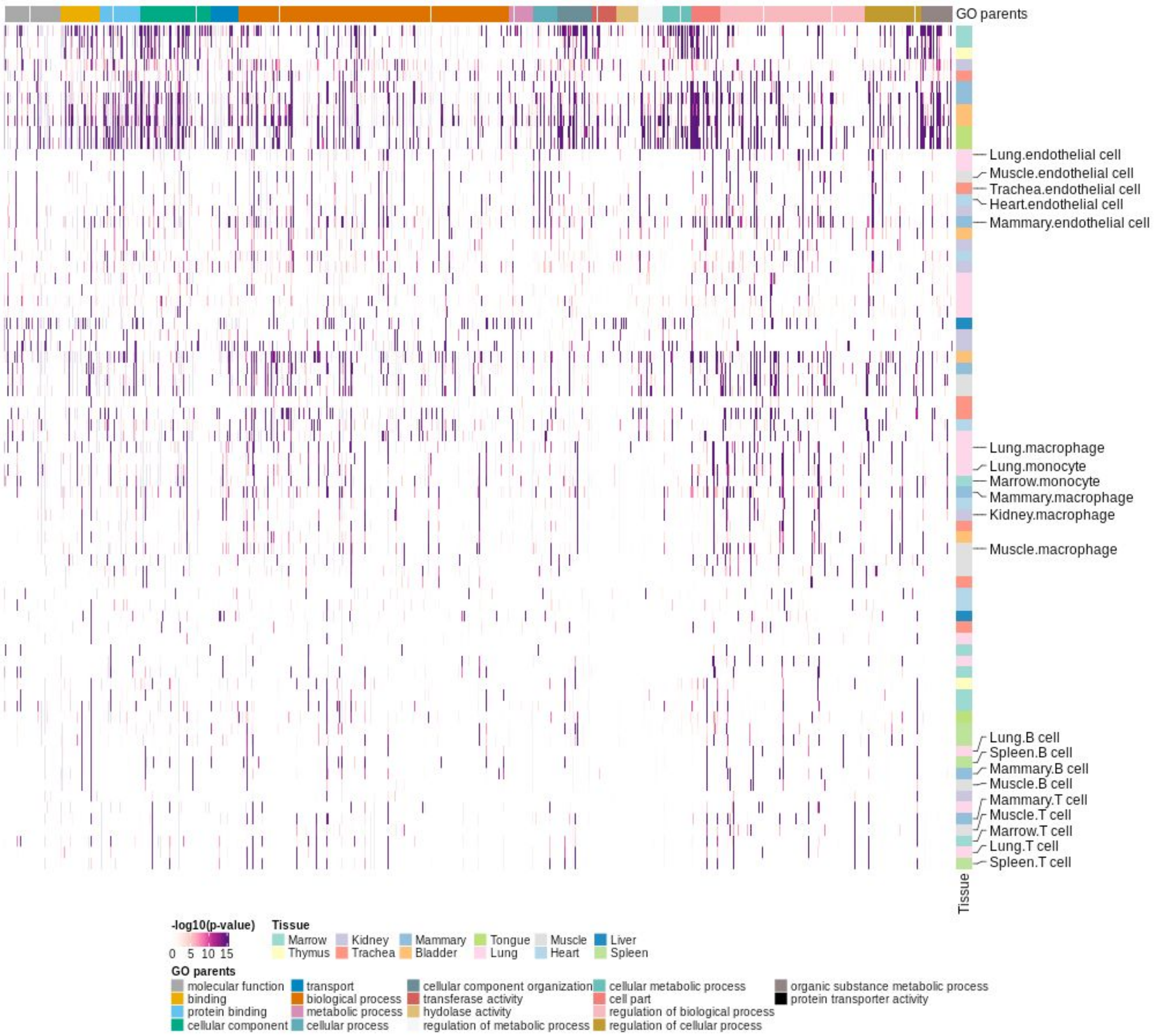
Supplementary Figure 8: Times to calculate cell type specificity for the genes found in the MCA and the two Tabula Muris datasets (MCA n=774, TM, 10X n=75 and TM, FACS n=110 cell types). Box plots show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points.



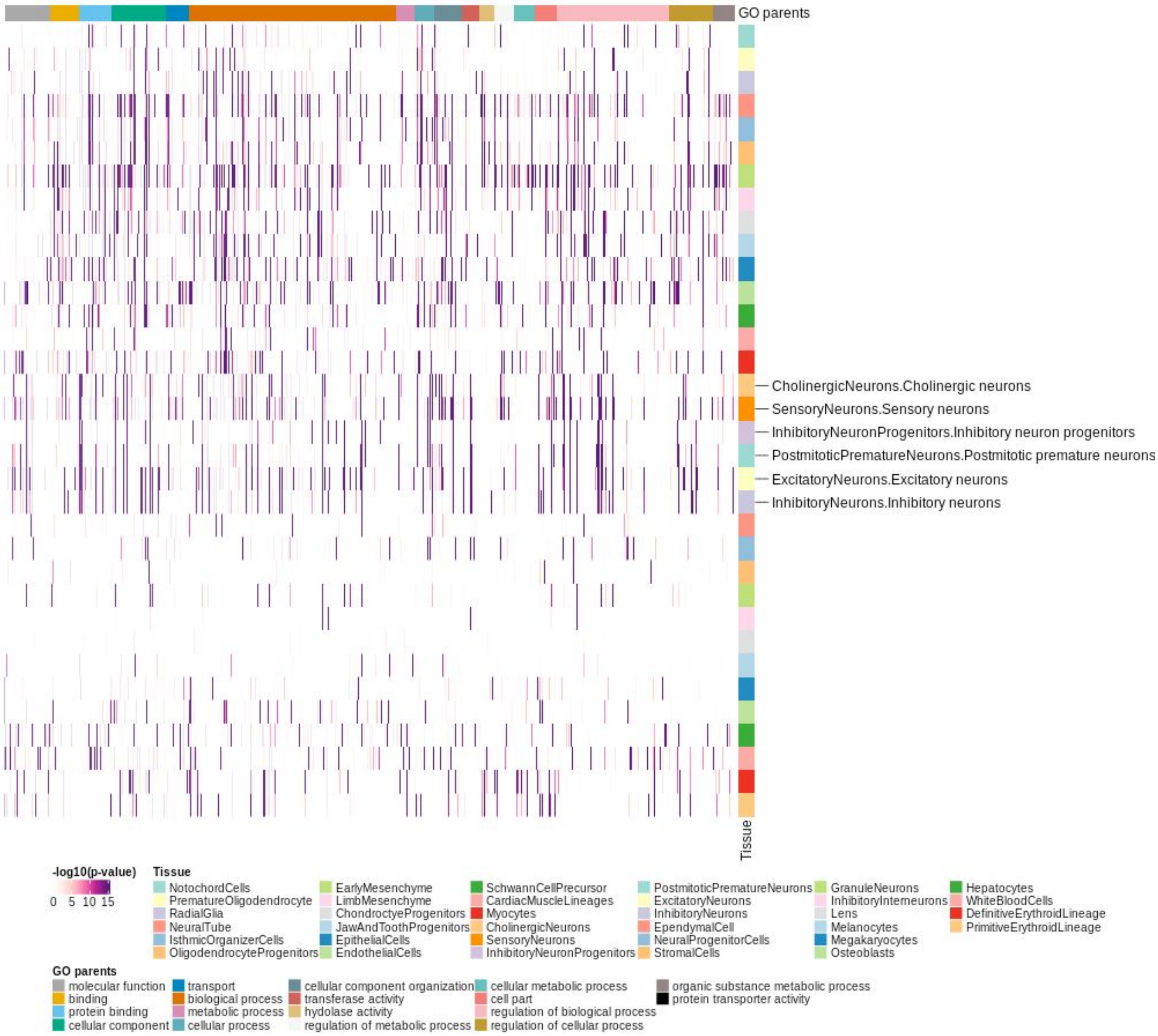
Supplementary Figure 9: (a) Search times for marker genes/peaks and (b) evaluation of number of cell types where a gene is found. Box plots show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points.



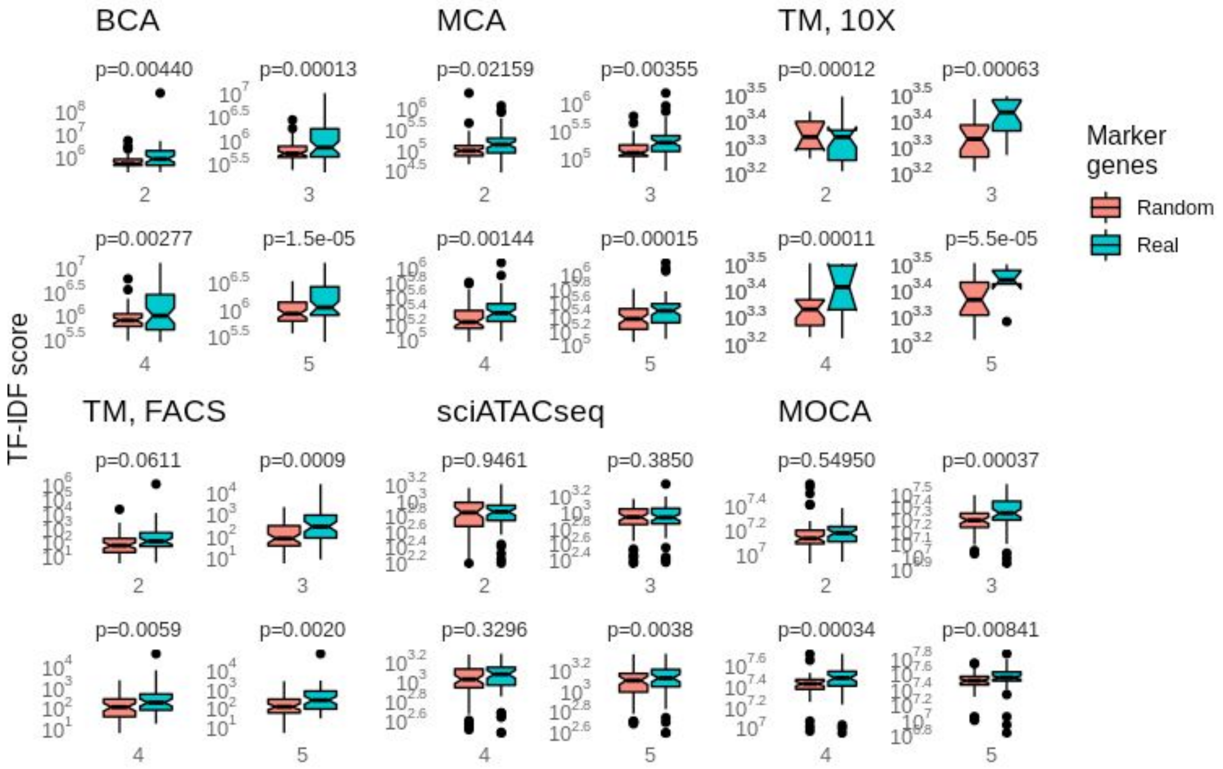
Supplementary Figure 10: Run times for subquery optimization with FP-growth or brute-force algorithms for gene sets containing between 5 and 20 genes. The mean run time (n=10) are presented by boxplots and the difference is assessed using a Wilcoxon test. Box plots show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points. Unpaired Wilcoxon Test was used.



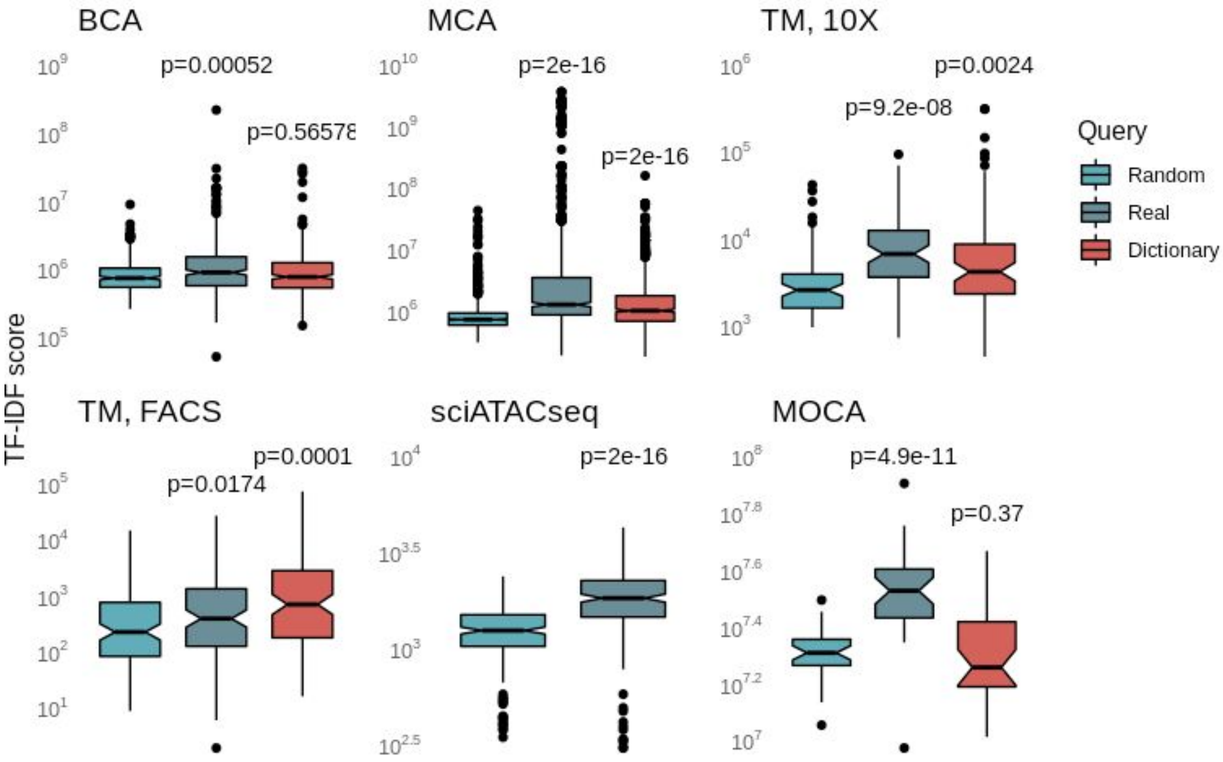
Supplementary Figure 11: Heatmap showing the enrichment of cell types from the TM 10X data for GO terms with between 5 and 25 genes. For the result of each gene set, one-tailed hypergeometric test with Holm adjustment for multiple comparison was used.



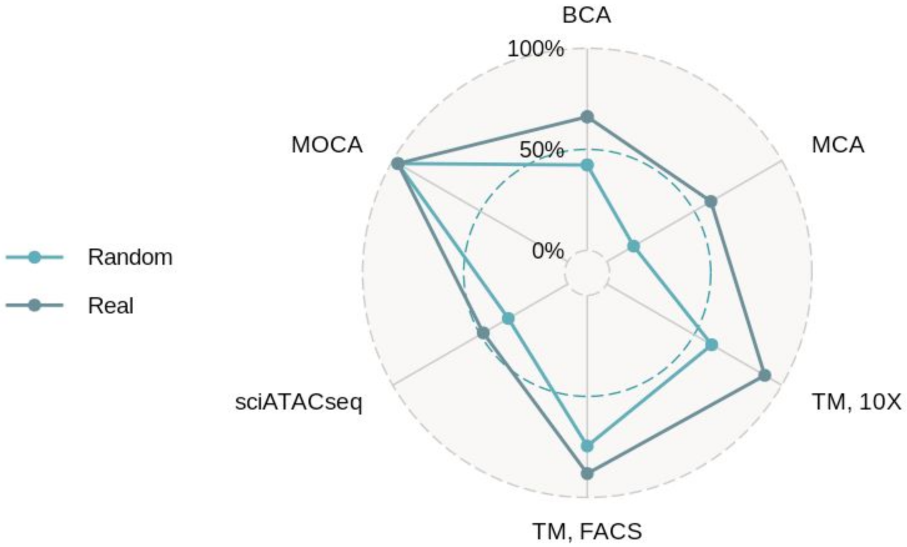
Supplementary Figure 12: Heatmap showing the enrichment of cell types from the MOCA data for GO terms with between 5 and 25 genes. For the result of each gene set, one-tailed hypergeometric test with Holm adjustment for multiple comparison was used.



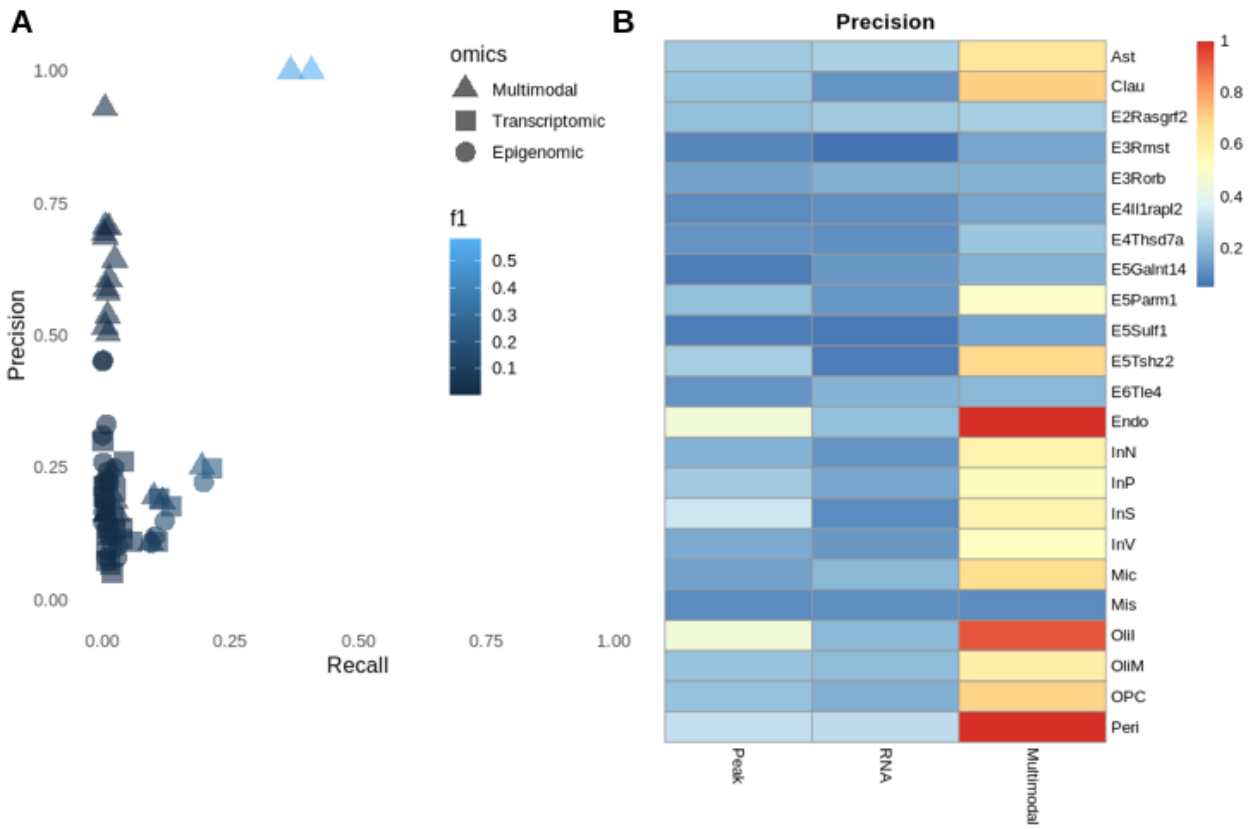
Supplementary Figure 13: Comparison of TF-IDF score between queries based on the top 20 (Real) and top 100 (Random) marker genes for the Mouse Brain Atlas, Mouse Cell Atlas, Tabula Muris (FACS and 10X), sciATACseq and MOCA datasets . Fifty sets of real gene queries and random gene queries with up to 5 genes from each dataset were generated. The highest TF-IDF scores from the best queries are presented by boxplots and the difference is assessed using a Wilcoxon test (n=50 gene sets). Box plots show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points. Paired Wilcoxon Test was used.



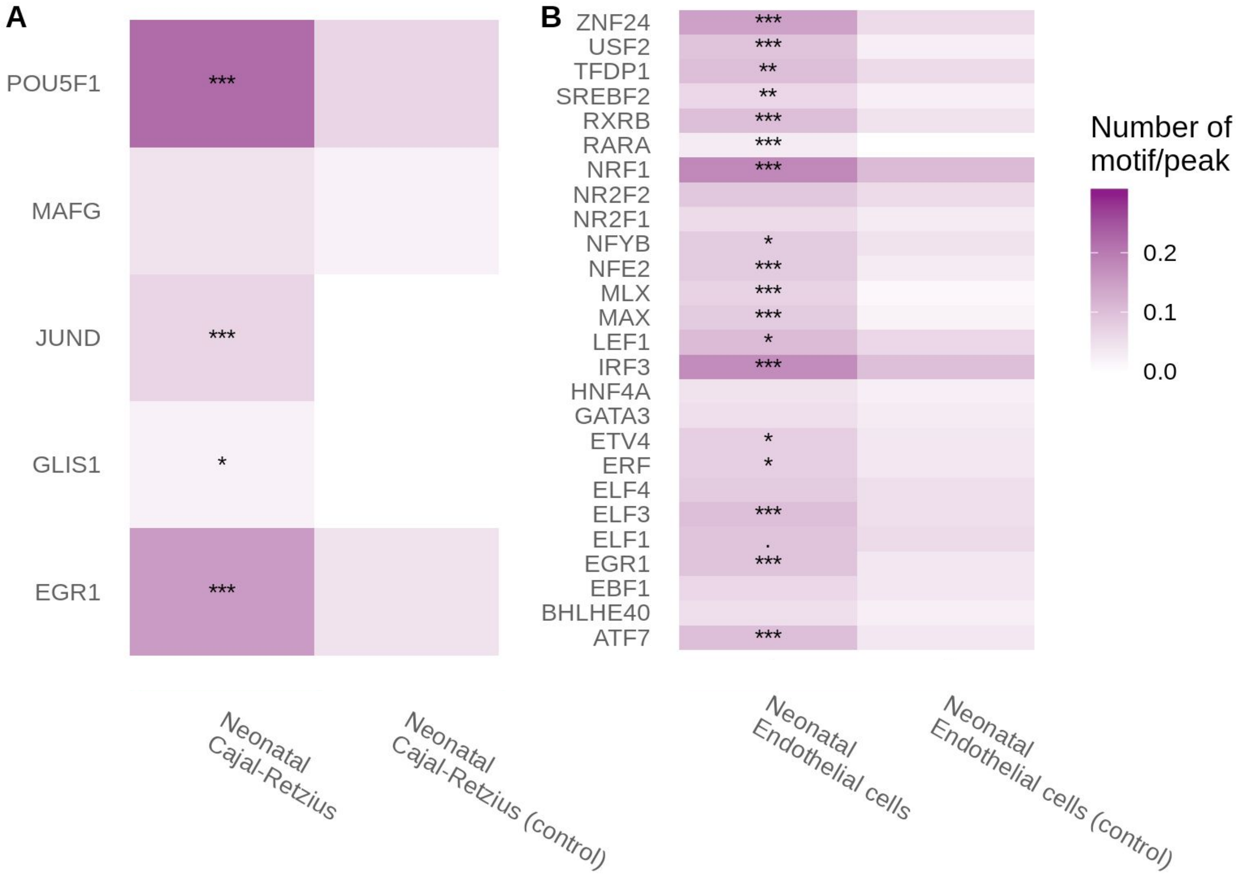
Supplementary Figure 14: Comparison of TF-IDF score between queries with 10 gene names sampled from each of the top 20 (Real) marker genes, top 1000 (Random) marker genes and phrases to gene names dictionaries (Dictionary) for the Mouse Brain Atlas, Mouse Cell Atlas, Tabula Muris (FACS and 10X), sciATACseq, MOCA. 1000 sets of gene queries per each group were generated. The highest TF-IDF scores from the best queries are presented by boxplots and the difference is assessed using a Wilcoxon test (n=1000 gene sets). Box plots show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points. Paired Wilcoxon Test was used.



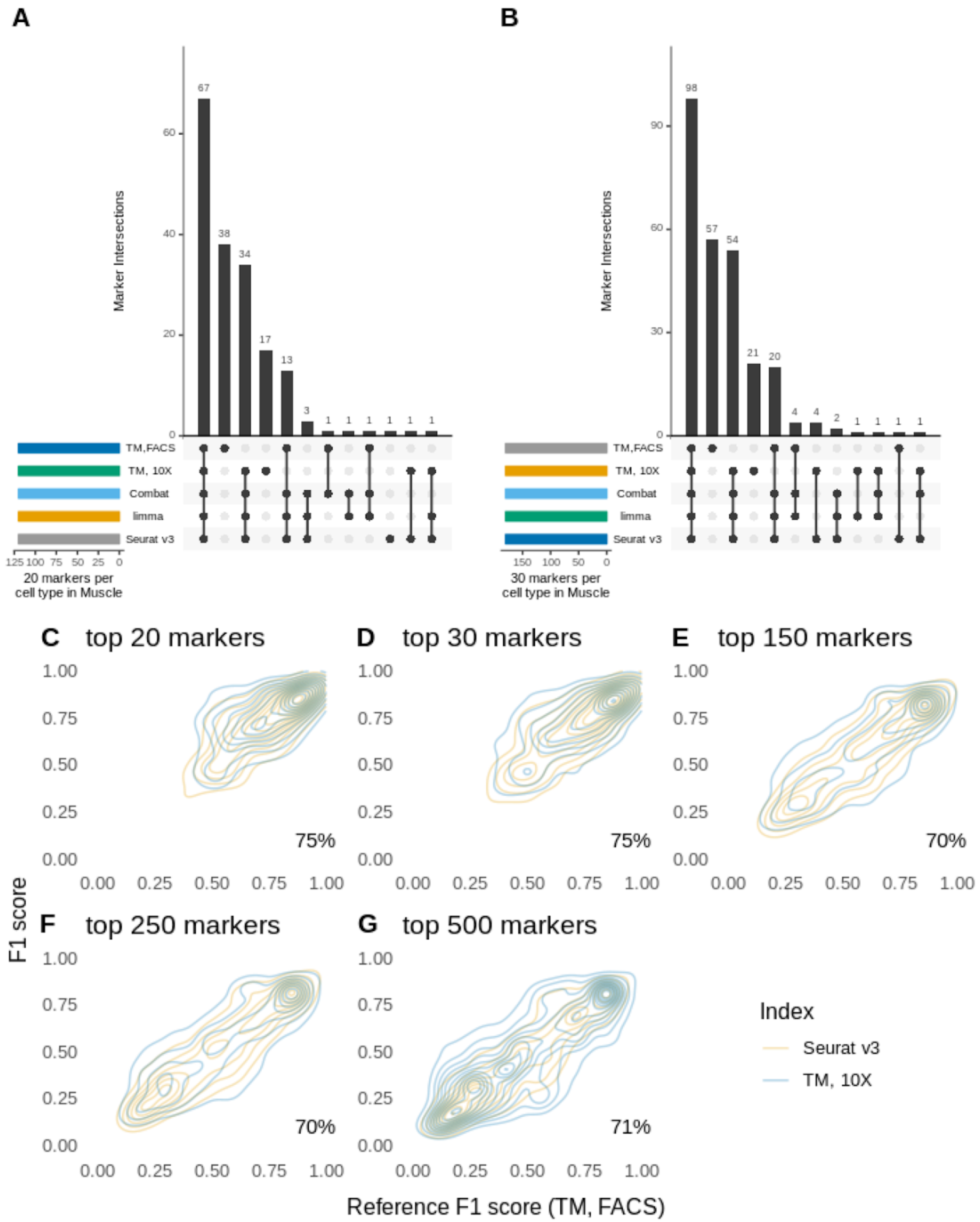
Supplementary Figure 15: Fraction of searches using the top query from **Supplementary Figure** 8 resulting in the desired cell type as the top query.



Supplementary Figure 16: Precision, recall and F1 values for cells of each cell type identified with multimodal, transcriptomic and epigenomic queries (a). A heatmap illustrating the result of precision values (b).



Supplementary Figure 17: Motif enrichment in putative distal enhancers that are specific to (a) neonatal Cajal-Retzius and (b) endothelial cells in the SNARE-seq dataset. Two-tailed Fisher's exact test with Benjamini and Hochberg adjustment for multiple comparison was used. (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$).



Supplementary Figure 18: Effect of batch correction on scfind marker gene search

A) Upset plot showing the overlap of top 20 marker genes from all 6 cell types in the Muscle dataset after batch correction against TM, 10X and TM, FACS. B) Upset plot showing the overlap of top 30 marker genes from all 6 cell types in the Muscle dataset after batch correction against TM, 10X and TM, FACS. Plot of shared top 20 after batch correction against TM, 10X and TM, FACS. C-G) Contour plots of F1 scores of top 20-500 marker genes after Seurat v3 batch correction against TM, 10X and TM, FACS with percentage of change against TM, FACS by <0.1 F1 scores. The curves show the density of genes and points close to the diagonal indicate that there is little difference in F1 scores between the two datasets.

Bibliography

1. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
2. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172,** 1091-1107.e17 (2018).
3. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174,** 1015-1030.e16 (2018).
4. Zeisel, A. *et al.* Molecular architecture of the mouse nervous system. *Cell* **174,** 999-1014.e22 (2018).
5. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174,** 1309-1324.e18 (2018).
6. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).
7. Regev, A. *et al.* The human cell atlas. *elife* **6,** (2017).
8. Howick, V. M. *et al.* The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science* **365,** (2019).
9. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
10. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
11. Sewell, W. Medical Subject Headings in MEDLARS. *Bull Assoc Med Libr* **52,** 164–170 (1964).
12. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862-8 (2016).

13. Cariaso, M. & Lennon, G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* **40,** D1308-12 (2012).

14. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019,** (2019).

15. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47,** D853–D858 (2019).

16. Athar, A. *et al.* ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47,** D711–D715 (2019).

17. Srivastava, D., Iyer, A., Kumar, V. & Sengupta, D. CellAtlasSearch: a scalable search engine for single cells. *Nucleic Acids Res.* **46,** W141–W147 (2018).

18. Sato, K., Tsuyuzaki, K., Shimizu, K. & Nikaido, I. CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing. *Genome Biol.* **20,** 31 (2019).

19. Vigna, S. Quasi-succinct indices. in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13* 83 (ACM Press, 2013). doi:10.1145/2433396.2433409

20. Schaum, N. *et al.* Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a *Tabula Muris*. *BioRxiv* (2017). doi:10.1101/237446

21. Golubovskaya, V. & Wu, L. Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy. *Cancers (Basel)* **8,** (2016).

22. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47,** D721–D728 (2019).

23. Bausch-Fluck, D. *et al.* A mass spectrometric-derived cell surface protein atlas. *PLoS ONE* **10,** e0121314 (2015).

24. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172,** 650–665 (2018).

25. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457-62 (2016).

26. Ju, W. *et al.* Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* **23,** 1862–1873 (2013).

27. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–574 (2013).

28. Han, J., Pei, J., Yin, Y. & Mao, R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Min. Knowl. Discov.* **8,** 53–87 (2004).

29. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28,** 11–21 (1972).

30. Piccini, I., Rao, J., Seebohm, G. & Greber, B. Human pluripotent stem cell-derived cardiomyocytes: Genome-wide expression profiling of long-term in vitro maturation in comparison to human heart tissue. *Genom. Data* **4,** 69–72 (2015).

31. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **47,** D23–D28 (2019).

32. Wei, C.-H., Kao, H.-Y. & Lu, Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **41,** W518-22 (2013).

33. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).

34. Manica, M., Mathis, R., Cadow, J. & Rodríguez Martínez, M. Context-specific interaction networks from vector representation of words. *Nat. Mach. Intell.* **1,** 181–190 (2019).

35. Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41,** D456-63 (2013).

36. Kim, S., Yeganova, L., Comeau, D. C., Wilbur, W. J. & Lu, Z. PubMed Phrases, an open set

of coherent phrases for searching biomedical literature. *Sci. Data* **5,** 180104 (2018).

37. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. *NIPS* 3111–3119 (2013).

38. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T. & Ananiadou, S. Distributional Semantics Resources for Biomedical Text Processing. *Languages in Biology and Medicine* (2013).

39. Alfares, A. A. *et al.* Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. *Genet. Med.* **17,** 880–888 (2015).

40. Flavigny, J. *et al.* Identification of two novel mutations in the ventricular regulatory myosin light chain gene (MYL2) associated with familial and classical forms of hypertrophic cardiomyopathy. *J. Mol. Med.* **76,** 208–214 (1998).

41. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).

42. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci USA* **110,** 17921–17926 (2013).

43. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44,** D164-71 (2016).

44. Joo, M. S., Koo, J. H., Kim, T. H., Kim, Y. S. & Kim, S. G. LRH1-driven transcription factor circuitry for hepatocyte identity: Super-enhancer cistromic analysis. *EBioMedicine* **40,** 488–503 (2019).

45. Thomas, G. D. *et al.* Deleting an Nr4a1 Super-Enhancer Subdomain Ablates Ly6Clow Monocytes while Preserving Macrophage Gene Function. *Immunity* **45,** 975–987 (2016).

46. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics

approaches for enhancer identification. *Brief. Bioinformatics* **17,** 967–979 (2016).

47. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37,** 1452–1457 (2019).
48. Bradshaw, A. D. & Sage, E. H. SPARC, a matricellular protein that functions in cellular differentiation and tissue response to injury. *J. Clin. Invest.* **107,** 1049–1054 (2001).
49. Callaham, M. L., Wears, R. L., Weber, E. J., Barton, C. & Young, G. Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA* **280,** 254–257 (1998).
50. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20,** 273–282 (2019).
51. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing . *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300 (1995).
52. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *JOSS* **3,** 861 (2018).
53. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16,** 479–487 (2019).
54. Chazarra-Gil, R., Hemberg, M., Kiselev, V. Y. & van Dongen, S. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *BioRxiv* (2020). doi:10.1101/2020.05.22.111211
55. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
56. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888-1902.e21 (2019).

57. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46,** D260–D266 (2018).

58. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32,** 1555–1556 (2016).

59. Matthay, M. A. & Thompson, B. T. Dexamethasone in hospitalised patients with COVID-19: addressing uncertainties. *Lancet Respir. Med.* (2020). doi:10.1016/S2213-2600(20)30503-8

60. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* (2020). doi:10.1038/s41586-020-2922-4

61. Sungnak, W. *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* **26,** 681–687 (2020).