Title

Methods and applications of integrating single nucleus and bulk tissue RNA sequencing

Author

Alvarez, Marcus Fernando

Publication Date

2022

UNIVERSITY OF CALIFORNIA

Los Angeles

Methods and applications of integrating single nucleus and bulk tissue RNA sequencing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Human Genetics

by

Marcus Fernando Alvarez

2022

ABSTRACT OF THE DISSERTATION

Methods and applications of integrating single nucleus and bulk tissue RNA sequencing

by

Marcus Fernando Alvarez
Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2022
Professor Päivi E. Pajukanta, Chair

Obesity typically precedes and accompanies the development of cardiometabolic diseases (CMD) that lead to increased morbidity and mortality. One of these disorders is non-alcoholic fatty liver disease (NAFLD), which encompasses a spectrum of varying degrees of fat accumulation and inflammation in the liver. More severe forms of NAFLD, such as non-alcoholic steatohepatitis (NASH), lead to a higher risk of developing hepatocellular carcinoma (HCC), the most prevalent form of liver cancer. Adipose tissue dysfunction in obesity can lead to increased circulating free fatty acids, and thus to ectopic lipid deposition in the liver. Left unchecked, lipotoxicity in the liver can result in inflammation, cell death, fibrosis, and ultimately the development of HCC. In both adipose and liver tissues, non-parenchymal cells, such as vascular and immune cell-types, play important roles in the normal function of these tissues and the pathophysiology of obesity, NAFLD, and HCC. A holistic approach to studying cell-types in a global manner would therefore greatly enhance our understanding of these common obesity-related diseases.

Single-cell technologies, such as single-cell RNA-sequencing (scRNA-seq), assay individual cells and provide an excellent tool to study cell-type changes. While these approaches provide high resolution, they are currently costly and low-throughput. Traditional methods that measure

molecular phenotypes at the tissue level are therefore still more practical. These assess a composite sum of cells present in the sample or biopsy, leading to inherent uncertainty in whether observed results are due to changes at the compositional level, cellular level, or both. Given these limitations, I aimed to integrate bulk-tissue RNA-sequencing (RNA-seq) and scRNA-seq data to leverage larger sample sizes in bulk RNA-seq and higher resolution in scRNA-seq.

The application of single-cell technologies is especially promising for biobanks, as they can contain multiple levels of data on participants to uncover novel associations. Tissues are typically stored frozen, however, and this usually requires nuclei suspensions for single-nucleus RNA-seq (snRNA-seq), whereas whole cells would typically be used for scRNA-seq. This presents challenges for current droplet-based technologies. RNA from the ambient pool of lysed cells and nuclei can encapsulate into droplets, confounding results. In Chapter 2, I present a computational method to remove empty droplets from gene expression data (Alvarez et al. 2020). This allows for cleaner downstream data analysis by ensuring that only droplets with nuclei or cells are used.

As current scRNA-seq technologies are low-throughput, their application to population-based studies and cohorts are limited. Present scRNA-seq technologies have lower throughput compared to bulk-tissue RNA-seq, which are typically available in higher sample sizes. In Chapter 3, I developed a method to help address this methodological gap. This approach, called Bisque (Jew et al. 2020), estimates cell-type composition in bulk RNA-seq data sets using single cell level reference data from the same tissue. The estimated cell-type proportions can be associated with sample-level data to uncover relevant cell-types, or they can be included as covariates in a model to reduce confounding caused by cell-type heterogeneity. One advantage of our method is that it requires only a minimum amount of information in the form of cell-type markers. This makes it attractive for existing data sets, which may not have accompanying single-cell level RNA-seq data.

In the fourth chapter of this dissertation, I present our application of snRNA-seq to HCC. Carcinomas, such as HCC, are typically characterized by high amounts of tissue heterogeneity. Larger scale cancer cohorts usually lack single-cell level data, making interpretation of bulk-tissue results challenging. Here, I integrated HCC single-cell level experiments with relatively large

HCC case-control bulk RNA-seq cohorts. The results from these analyses highlighted the role that proliferating cells play in HCC (Alvarez et al. 2022). These cycling cells were highly enriched in cancer tissue, as expected, and were prognostic of poor survival outcomes consistently in two independent cohorts. Furthermore, we observed that individuals with TP53 mutations have higher levels of these proliferating cells. Thus, our integration helped to interpret tumor gene expression changes as cell-type composition changes.

In the fifth chapter, I present our human adipose tissue snRNA-seq results, showing changes in obesity and insulin resistance (Alvarez et al. manuscript in preparation). We applied multiplexing to increase our snRNA-seq sample size to roughly 100 subcutaneous adipose samples and over 100,000 nuclei, providing unprecedented resolution of human adipose tissue. This allowed us to identify finer resolution subcell-types, or cell states, which are more challenging to study as they are lower in frequency and exhibit more subtle differences. In addition to substantiating previous findings, we identified subcell-types associated with CMD. Then, we apply integrative approaches to corroborate these cell state changes in adipose bulk RNA-seq. Overall, our results show that both main cell-type and subcell-type variations are associated with metabolic traits.

In summary, this dissertation presents my work on the integration of snRNA-seq and bulk-tissue RNA-seq to leverage distinct advantages provided by each. This has allowed us to gain a better understanding of the origin of gene expression changes in CMD.

The dissertation of Marcus Fernando Alvarez is approved.

M. Luisa Iruela-Arispe

Janet S. Sinsheimer

Joseph R. Pisegna

Päivi E. Pajukanta, Committee Chair

University of California, Los Angeles

2022

To my father, mother, and brother

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CMD . . . . . cardiometabolic disease

CVD . . . . . cardiovascular disease

DIEM . . . . . debris identification using expectation maximization

DE . . . . . differential expression

FFA . . . . . free fatty acid

HCC . . . . . hepatocellular carcinoma

LCI . . . . . liver cancer institute

NAFLD . . . . . non-alcoholic fatty liver disease

NASH . . . . . non-alcoholic steatohepatitis

OS . . . . . overall survival

PBMC . . . . . peripheral blood mononuclear cell

PFI . . . . . progression free interval

RNA-seq . . . . . RNA sequencing

scRNA-seq . . . . . single-cell RNA-seq

snRNA-seq . . . . . single-nucleus RNA-seq

T2D . . . . . type 2 diabetes

TCGA . . . . . the cancer genome atlas

TG . . . . . triglyceride

UMI . . . . . unique molecular index

UMAP . . . . . uniform manifold approximation and projection

ACKNOWLEDGMENTS

I would like to thank Dr. Päivi Pajukanta for guiding and supporting me throughout my PhD. Your unwavering commitment to education has allowed my career to grow unimpeded. In our scientific endeavors you have managed to provide a careful balance of direction and independence that I find rare. Your enthusiasm for research, no matter how demanding it can become, truly stands out.

I would like to thank my committee members Drs. Janet Sinsheimer, Joseph Pisegna, and Luisa Iruela-Arispe. Your support and insight have critically pushed me in the right direction with statistical and experimental design and interpretation. I have seen how important education is to each of you, and your support for your students is immeasurable and extensive. Joe, I would like to thank you, Patrizia, and your lab for welcoming me to work in your lab throughout these years.

I would like to thank the current and past members of the Pajukanta lab for providing both a hospitable and academic environment. It would be impossible to fit a complete acknowledgment for all of you, but I would like to thank Elina Nikkola, David Pan, and Zong Miao. Kristina Garske and Jihane Benhammou, it has been a pleasure working together as colleagues and sharing unforgettable memories as friends. Tony Lee and Sankha Subhra Das, it has been wonderful working with you in these recent years.

I would like to thank the Halperin lab for working with us on our single-cell projects. Your collaborations have produced exceptional research. Thank you, Dr. Eran Halperin, Elior Rahmani, Brandon Jew, Johnson Chen, and Oren Avram. Working together with you has taught me the computational and methodological fields of research that I would never have been able to learn on my own.

I would like to thank the students and mentees of our lab. Elliot Kim, Yash Bhagat, Caroline Comenho, Sandhya Rajkumar, and Niko Darci-Maher, it has been rewarding to not just help but to work together with you on the lab's research projects. Teaching and mentoring are mutually

beneficial endeavors, and I am excited to see where your paths lead you.

I would like to thank our external collaborators. Dr. Kirsi Pietiläinen, you have been such an important associate that my research projects have almost completely depended on your support. I would like to thank you and your lab for welcoming me to Helsinki during my visits. Dr. Carlos Aguilar-Salinas, it has been a pleasure working with you and Dr. Teresa Tusie-Luna. I would like to thank you both for your supportive collaborations and for warmly accommodating me in my visit to the UNAM.

Chapter 2 was originally published as "Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM" in *Scientific Reports* in volume 10 in the year 2020 under doi 10.1186/s13073-022-01055-5. The authors that contributed to this study were Marcus Alvarez, Elior Rahmani, Brandon Jew, Kristina M. Garske, Zong Miao, Jihane N. Benhammou, Chun Jimmie Ye, Joseph R. Pisegna, Kirsi H. Pietiläinen, Eran Halperin and Päivi Pajukanta. Elior Rahmani and I are co-first authors on this publication. Elior Rahmani, Eran Halperin, Päivi Pajukanta and I conceived and designed the study. Kristina M. Garske, Jihane Benhammou and I designed and performed experiments to collect the snRNA-seq data. Elior Rahmani, Brandon Jew, Zong Miao, Kristina M. Garske, Jihane Benhammou, and I analyzed and interepreted this data. Elior Rahmani, Eran Halperin, Päivi Pajukanta and I designed the DIEM approach. Joseph R. Pisegna, Chun Jimmie Ye, Kirsi H. Pietiläinen, and Päivi Pajukanta designed and supervised the snRNA-seq experiments. Elior Rahmani, Brandon Jew, Zong Miao, Eran Halperin, Päivi Pajukanta and I wrote the manuscript.

Chapter 3 was originally published as "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information" in *Nature Communications* in the year 2020 in volume 11 under doi 10.1038/s41467-020-15816-6. The authors that contributed to this study were Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, and Eran Halperin. Brandon Jew and I were co-first authors on this publication. Eran Halperin and Päivi Pajukanta conceived and supervised the study. Brandon Jew and I developed the method, software, and conducted experiments. Kirsi H. Pietiläinen

provided biopsies for snRNA-seq experiments. Kristina M. Garske and I developed and performed the snRNA-seq of frozen adipose. Jae Hoon Sul provided guidance on analysis of the brain data. Elior Rahmani, Zong Miao, and Arthur Ko provided statistical and biological insight. Brandon Jew, Päivi Pajukanta, Eran Halperin and I wrote the manuscript with support from all listed authors.

Chapter 4 was originally published as "Human liver single nucleus and single cell RNA sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival" in *Genome Medicine* in the year 2022 in volume 14 under doi 10.1186/s13073-022-01055-5. The authors that contributed to this study were Marcus Alvarez, Jihane N. Benhammou, Nicholas Darci-Maher, Samuel W. French, Steven B. Han, Janet S. Sinsheimer, Vatche G. Agopian, Joseph R. Pisegna and Päivi Pajukanta. Jihane N. Benhammou and I were co-first authors on this publication. Jihane N. Benhammou, Joseph R. Pisegna, Päivi Pajukanta and I conceived the study. Jihane N. Benhammou and I performed the snRNA-seq experiments. Jihane N. Benhammou, Niko Darci-Maher and I performed analysis and interpretation of data. Samuel W. French provided histology review. Janet S. Sinsheimer contributed statistical methodology and interpretation. Vatch G. Agopian provided NAFLD-HCC patients. Jihane N. Benhammou, Joseph R. Pisegna, Päivi Pajukanta and I wrote the manuscript with support from all authors. Joseph R. Pisegna and Päivi Pajukanta supervised the experiments and analysis.

Chapter 5 titled "Human adipose single nucleus RNA-seq reveals an adipocyte axis associated with cardiometabolic disease" is a manuscript currently in preparation. The authors that contributed to this study were Marcus Alvarez, Elior Rahmani, Zeyuan Chen, Oren Avram, Birgitta W. van der Kolk, Niko Darci-Maher, Karen L. Mohlke, Kirsi H. Pietiläinen, Eran Halperin, Markku Laakso, and Päivi Pajukanta. Päivi Pajukanta, Eran Halperin, and I conceived and designed the study. Kirsi H. Pietiläinen and Markku Laakso provided adipose tissue biopsies. Birgitta W. van der Kolk and I processed adipose sampes for snRNA-seq. Elior Rahmani, Zeyuan Chen, Oren Avram, Niko Darci-Maher and I performed analysis. Karen L. Mohlke provided insight into experimental design and data interpretation. Päivi Pajukanta and Eran Halperin supervised the study.

I would like to thank my sources of funding. This dissertation was supported by the NIH T32HG002536, NIH-NHLBI F31 HL144078, and HHMI Gilliam Fellowship grants.

I would like to thank the patients and individuals of the cohorts that participated in these studies. These experiments were only made possible by the contribution of their time and biological samples.

VITA

Education

Sept 2011–June 2015 **University of California, Los Angeles**
B.S., Microbiology, Immunology, & Molecular Genetics

Honors and Awards

2019	ASHG Epstein Semifinalist
2018	HHMI Gilliam Fellowship
2016	NSF Honorable Mention
2015	UCLA Eugene V. Cota Robles Scholar
2015	UCLA NIH T32 Genomic Analysis Training Program Grant

Selected Publications

1. **Alvarez, M.***, Benhammou, J.N.*, Darci-Maher, N., French, S.W., Han, S.B., Sinsheimer, J.S., Agopian, V.G., Pisegna, J.R., Pajukanta, P. Human liver single nucleus and single cell RNA sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival. Genome Medicine **14**, 50 (2022).

2. Rao, S., Yang, X., Ohshiro, K., Zaidi, S., Wang, Z., Shetty, K., Xiang, X., Hassan, M.I., Mohammad, T., Latham, P.S., Nguyen, B., Wong, L., Yu, H., Al-Abed, Y., Mishra, B., Vacca, M., Guenigault, G., Allison, M.E.D., Vidal-Puig, A., Benhammou, J.N., **Alvarez, M.**, Pajukanta, P., Pisegna, J.R., Mishra, L. β2-spectrin (SPTBN1) as a therapeutic target for diet-induced liver disease and preventing cancer development. Science Translational Medicine **13**, eabk2267 (2022).

3. Pan, D.Z., Miao, Z., Comenho, C., Rajkumar, S., Koka, A., Lee, S.H.T., **Alvarez, M.**, Kaminska, D., Ko, A., Sinsheimer, J.S., Mohlke, K.L., Mancuso, N., Muñoz-Hernandez, L.L., Herrera-Hernandez, M., Tusié-Luna, M.T., Aguilar-Salinas, C., Pietiläinen, K.H., Pihlajamäki, J., Laakso, M., Garske, K.M., Pajukanta, P. Identification of TBX15 as an adipose master trans regulator of abdominal obesity genes. Genome Medicine **13**, 123 (2021).

4. van der Kolk, B.W., Muniandy, M., Kaminska, D., **Alvarez, M.**, Ko, A., Miao, Z., Valsesia, A., Langin, D., Vaittinen, M., Pääkkönen, M., Jokinen, R., Kaye, S., Heinonen, S., Virtanen, K.A., Andersson, D.P., Männistö, V., Saris, W.H., Astrup, A., Rydén, M., Blaak, E.E., Pajukanta, P., Pihlajamäki, J., Pietiläinen, K.H. Differential Mitochondrial Gene Expression in Adipose Tissue Following Weight Loss Induced by Diet or Bariatric Surgery. The Journal of Clinical Endocrinology & Metabolism **106**, 1312-1324 (2021).

5. Karunakaran, D., Turner, A.W., Duchez, A., Soubeyrand, S., Rasheed, A., Smyth, D., Cook, D.P., Nikpay, M., Kandiah, J.W., Pan, C., Geoffrion, M., Lee, R., Boytard, L., Wyatt, H., Nguyen, M., Lau, P., Laakso, M., Ramkhelawon, B., **Alvarez, M.**, Pietiläinen, K.H., Pajukanta, P., Vanderhyden, B.C., Liu, P., Berger, S.B., Gough, P.J., Bertin, J., Harper, M., Lusis, A.J., McPherson, R., Rayner, K.J. RIPK1 gene variants associate with obesity in humans and can be therapeutically silenced to reduce obesity in mice. Nature Metabolism **2**, 1113-1125 (2020).

6. Miao, Z., **Alvarez, M.**, Ko, A., Bhagat, Y., Rahmani, E., Jew, B., Heinonen, S., Muñoz-Hernandez, L.L., Herrera-Hernandez, M., Aguilar-Salinas, C., Tusie-Luna, T., Mohlke, K.L., Laakso, M., Pietiläinen, K.H., Halperin, E., Pajukanta, P. The causal effect of obesity on pre-diabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance. PLOS Genetics **16**, e1009018 (2020).

7. **Alvarez, M.***, Rahmani, E.*, Jew, B., Garske, K.M., Miao, Z., Benhammou, J.N., Ye, C.J., Pisegna, J.R., Pietiläinen, K.H., Halperin, E., Pajukanta, P. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. Scientific Reports **10**, 11019 (2020)

8. Jew, B.*, **Alvarez, M.***, Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., Halperin, E. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature Communications **11**, 1971 (2020).

9. Miao, Z., **Alvarez**, M., Pajukanta, P., Ko, A. ASElux: an ultra-fast and accurate allelic reads counter. Bioinformatics **34**, 1313-1320 (2018).

10. Pan, D.Z., Garske, K.M., **Alvarez, M.**, Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek, M., Glastonbury, C.A., Small, K.S., Boehnke, M., Lusis, A.J., Sinsheimer, J.S., Mohlke, K.L., Laakso, M., Pajukanta, P., Ko, A. Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. Nature Communications **9**, 1512 (2018).

* indicates shared first authorship

CHAPTER 1

Introduction and background

1.1 Obesity-related cardiometabolic disorders

Obesity afflicts a large percentage of the population and predisposes individuals to more threatening comorbidities, such as type 2 diabetes (T2D) and cardiovascular disease (CVD). Obesity has risen in prevalence globally due to fat and sugar rich diets combined with physical inactivity. Current studies estimate the rate of obese adults in the U.S. at around 34% [1], and some studies predict nearly 1 in 2 adults will be obese by 2050 [2]. As a high BMI level is a substantial risk factor for these common comorbidities, obesity also imposes vast economic costs on the society [3]. Obtaining more refined cell-type and tissue level knowledge on the pathophysiology of obesity will ultimately help better address this health epidemic.

Given the widespread nature of obesity, we have a general understanding of its risk factors. Two factors, in varying degrees, interact to give a higher obesity predisposition: environment and genetics. The environment of an individual includes contexts and behaviors that lead to higher energy intake compared to expenditure, such as calorie-rich diets and physical inactivity [4]. Genetic factors encompass a spectrum of rare and common variation that typically, but not exclusively and deterministically, affect appetite control in humans [5]. While numerous genetic studies of obesity and BMI have revealed hundreds of associations, many loci maintain an elusive mechanism by which they act. For example, the association at the FTO locus is one of the strongest and most validated for BMI, yet studies have shown distinct mechanisms through which this gene acts [6]. Therefore, we still have an incomplete understanding of the pathogenesis of aberrant weight gain.

The principal site of action in obesity is adipose tissue, the fat-storing organ of the body. Excess energy intake leads to fat storage in the form of triglycerides (TGs), which are energy-dense molecules that can be efficiently packed into lipid droplets. Adipose tissue expands and remodels to accommodate increasing amounts of these TGs [7]. Adipocytes, the parenchymal fat-storing cells of adipose, can expand in size (hypertrophy) and increase in number (hyperplasia) to drive this volume expansion [8]. While fat storage is a physiologically normal process, its excess can be accompanied by unhealthy abnormalities. A major feature of obesity is systemic inflammation characterized by inflammatory cytokines in the blood [9]. This is paralleled by local immune cell infiltration in the adipose tissue. Macrophages were one of the first immune cells found enriched in obese fat depots [10]. Additionally, T cells have been found to increase with weight gain [11]. Overall, fat tissue expansion is accompanied by complex interactions with different organs and between various cells within adipose tissue itself.

One of the key interacting organs in obesity is the liver. Adipose releases free fatty acids (FFAs) from hydrolysis of TG and the liver takes them in from the blood. These FFAs are stored again as TGs in lipid droplets, which can later be secreted in VLDL particles or degraded. In obesity, and especially in insulin resistant states, excessive FFA release and liver uptake can lead to an imbalance favoring lipid droplet formation. This excess of lipid droplets in the liver is known as hepatic steatosis, which is part of a spectrum of non-alcoholic fatty liver disease (NAFLD). The positive connection between obesity and NAFLD has been shown in epidemiological studies [12]. Furthermore, the presence of additional metabolic disorders, such as T2D and hypertension, is associated with more severe forms of liver disease [13]. These include the presence of inflammation, known as non-alcoholic steatohepatitis (NASH), and liver fibrosis and cirrhosis. While simple steatosis is usually benign, the progression of NASH and fibrosis can lead to hepatocellular carcinoma (HCC), the most common form of liver cancer with a high mortality rate [14]. This involves multiple mechanisms and cell-types that include the DNA damage response, metabolic stress, stellate cell activation, and inflammation [14].

Obesity, its associated CMDs, NAFLD/NASH, and HCC are characterized by complex pro-

cesses involving changes in parenchymal cells and their interactions with non-parenchymal cells, such as vascular and immune cells. While progress has been made in characterizing these changes, much is still unknown. For example, adipocyte morphology changes in obesity and insulin resistance have been studied [15] but the precise changes of molecular pathways *in vivo* are still unknown. Detailed investigations of these processes in humans could therefore elucidate many additional mechanisms involved in the pathogenesis of these CMDs.

1.2 Single-cell technologies

Gene expression in cells or tissues can provide valuable insights into biological processes. The encoded proteins serve specific functions, and they typically act in concerted networks to form pathways involved in a myriad of cellular processes [16]. Gene expression can therefore reflect the state of a cell. The comparison of gene expression across two contexts, i.e. differential gene expression, is also useful to discover changes in cellular pathways.

RNA sequencing (RNA-seq) is a powerful approach to measure gene expression and carry out differential gene expression [17]. RNA is isolated from a sample of interest, converted to cDNA, and adapted into a library for sequencing [17]. A distinguishing feature of RNA-seq is the ability to profile transcription genome-wide, permitting hypothesis-free discoveries. Numerous studies across various fields have successfully applied RNA-seq to perform differential gene expression in human tissues [18, 19]. Although many insights can be derived from the application of bulk-tissue RNA-seq, the measure is a composite sum of all cells present in the sample. Consequently, the source of observed changes in gene expression is uncertain. Variation in cell-type proportion, cell-type-specific expression, or both could cause differences in expression [19].

The limitations of bulk-tissue approaches have brought forth the field of single-cell genomics. The overarching goal of this field is to provide measurements on individual cells from a sample. Single-cell RNA-seq can thus provide insight into cell-types present in heterogeneous tissues. Earlier methods accomplished cell separation using flow sorting or micropipettes [20]. These

plate-based methods typically work in the order of hundreds of cells. Later developments using microfluidics have allowed experiments to scale to thousands of cells simultaneously [21, 22]. Various adaptations in technologies and sequencing methods have provided assays to measure diverse molecular phenotypes, such as the epigenome [23]. The use of nuclei instead of cells for single-nucleus RNA-seq (snRNA-seq) permits profiling of frozen tissues, where isolation of cells would be challenging [23, 24].

The ability to profile single cells within a heterogenous pool is particularly suited to applications of complex solid tissues. Single-cell studies in the brain have revealed many previously unknown subtypes within known major cell-types [25]. By leveraging splicing dynamics, it also is possible to infer a temporal ordering of cells along differentiation trajectories [26]. Furthermore, single-cell applications have been especially successful in the field of cancer genomics by dissecting tumor heterogeneity [5] and profiling T cell exhaustion [22].

Although powerful in its resolution, single-cell genomics is currently limited by its relatively small scale capabilities. The current cost and low-throughput nature of these approaches mean that the sample sizes are roughly 10-100 times smaller than with bulk-tissue methods. Methodologies have been adapted to circumvent this limitation. Samples with natural genetic variation, such a cohort of individuals, can be pooled and later de-multiplexed by leveraging genetic variation inherently profiled in sequencing reads [28]. Deconvolution and decomposition methods estimate cell-type proportions in bulk-tissue data with larger sample sizes [29, 30, 2]. Thus, the development of approaches that increase sample sizes and integrate across bulk-tissue experiments allow for population-based association studies at cell-type resolution.

1.3 Contributions of this thesis work to the current state of knowledge

The second chapter of this thesis details our work in developing a method to computationally process single-cell level RNA-seq data from frozen human tissues [1]. Proof-of-concept experiments for single-cell technologies have typically used fresh tissues, such as blood, in which isolation of

a cell suspension is relatively straightforward. However, many well-phenotyped biobanks store frozen tissue, for which obtaining a clean suspension of cells or nuclei is technically challenging. Cells are susceptible to lysis when thawed, and higher amounts of ambient RNA tend to encapsulate in droplets. This makes cell/nucleus identification more difficult. Our approach, called Debris Identification using Expectation Maximization (DIEM), removes empty and highly contaminated droplets from single-cell and single-nucleus RNA-seq experiments. We show that this results in a higher level of certainty in clustering and downstream results. Overall, our tool can help researchers take advantage of frozen tissues for single-cell level investigations.

In the third chapter of this thesis, I describe our computational approach that helps integrate single-cell and bulk-tissue RNA-seq [2]. This integration leverages the advantages of each method for association studies where larger sample sizes are required. Our tool, called Bisque, estimates cell-type proportions in bulk-tissue RNA-seq using single-cell level data. Either overlapping samples or cell-type marker information can be used to decompose bulk samples. This approach is useful for researchers to estimate cell-type proportions for use in association studies. Cell-type abundance can be studied directly to infer novel biological connections or included as a covariate to remove this variation if treated as a confounder.

The fourth chapter of this thesis presents an application of the above two methods to single-cell level RNA-seq data of HCC liver tumor and non-tumor samples [3]. Previous single-cell studies in HCC have been limited by their sample sizes, making population-level inferences of HCC cell-types challenging. We first integrated our own HCC snRNA-seq data with two previously published data sets to maximize our power to detect cell-types. Then, we used this integrated reference to estimate proportions in bulk RNA-seq cohorts. This integration elucidated the importance of a proliferating cell-type in HCC survival outcomes and its association with *TP53* mutations.

In the fifth chapter of this thesis, I describe results from our adipose tissue snRNA-seq analysis (Alvarez et al manuscript in preparation). Here, we multiplexed samples using natural genetic variation to sequence subcutaneous adipose tissue nuclei from over 100 individuals. In addition, we combined adipose bulk-tissue RNA-seq data, consisting of over 300 individuals, to validate our

results. The main contributions of this work are our subcell-type analysis. We carefully identified cell states within the major known cell-types in the adipose tissue. Then, we discovered associations between subcell-type abundance and cardiometabolic traits. Furthermore, these correlations were present in our integrative analysis that applied canonical correlation analysis (CCA) between the single-nucleus and bulk-tissue RNA-seq data sets. Our results show that cell-type specific signals were present in bulk-tissue data and associated with cardiometabolic traits.

REFERENCES

[1] Ogden, C. L., Carroll, M. D., Kit, B. K. & Flegal, K. M. Prevalence of childhood and adult obesity in the united states, 2011-2012. *JAMA* **311**, 806–814 (2014). URL https://doi.org/10.1001/jama.2014.732.

[2] Ward, Z. J. *et al.* Projected u.s. state-level prevalence of adult obesity and severe obesity. *New England Journal of Medicine* **381**, 2440–2450 (2019). URL https://doi.org/10.1056/NEJMsa1909301.

[3] Cawley, J. & Meyerhoefer, C. The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics* **31**, 219–230 (2012). URL https://doi.org/10.1016/j.jhealeco.2011.10.003.

[4] Spiegelman, B. M. & Flier, J. S. Obesity and the regulation of energy balance. *Cell* **104**, 531–543 (2001). URL https://doi.org/10.1016/S0092-8674(01)00240-9.

[5] Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. *Nature Reviews Genetics* **23**, 120–133 (2022). URL https://doi.org/10.1038/s41576-021-00414-z.

[6] Claussnitzer, M. *et al.* Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine* **373**, 895–907 (2015). URL https://doi.org/10.1056/NEJMoa1502214.

[7] Longo, M. *et al.* Adipose tissue dysfunction as determinant of obesity-associated metabolic complications. *International Journal of Molecular Sciences* **20** (2019). URL https://doi.org/10.3390/ijms20092358.

[8] Tchoukalova, Y. D. *et al.* Regional differences in cellular mechanisms of adipose tissue gain with overfeeding. *Proceedings of the National Academy of Sciences* **107**, 18226–18231 (2010). URL https://doi.org/10.1073/pnas.1005259107.

[9] Hotamisligil, G. S., Shargill, N. S. & Spiegelman, B. M. Adipose expression of tumor necrosis factor-α: Direct role in obesity-linked insulin resistance. *Science* **259**, 87–91 (1993). URL https://doi.org/10.1126/science.7678183.

[10] Weisberg, S. P. *et al.* Obesity is associated with macrophage accumulation in adipose tissue. *The Journal of Clinical Investigation* **112**, 1796–1808 (2003). URL https://doi.org/10.1172/JCI19246.

[11] Nishimura, S. *et al.* Cd8+ effector t cells contribute to macrophage recruitment and adipose tissue inflammation in obesity. *Nature Medicine* **15**, 914–920 (2009). URL https://doi.org/10.1038/nm.1964.

[12] Li, L. *et al.* Obesity is an independent risk factor for non-alcoholic fatty liver disease: evidence from a meta-analysis of 21 cohort studies. *Obesity Reviews* **17**, 510–519 (2016). URL https://doi.org/10.1111/obr.12407.

[13] Marchesini, G. *et al.* Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology* **37**, 917–923 (2003). URL https://doi.org/10.1053/jhep.2003.50161.

[14] Anstee, Q. M., Reeves, H. L., Kotsiliti, E., Govaere, O. & Heikenwalder, M. From nash to hcc: current concepts and future challenges. *Nature Reviews Gastroenterology & Hepatology* **16**, 411–428 (2019). URL https://doi.org/10.1038/s41575-019-0145-7.

[15] Stenkula, K. G. & Erlanson-Albertsson, C. Adipose cell size: importance in health and disease. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **315**, R284–R295 (2018). URL https://doi.org/10.1152/ajpregu.00257.2017.

[16] Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101–113 (2004). URL https://doi.org/10.1038/nrg1272.

[17] Wang, Z., Gerstein, M. & Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63 (2009). URL https://doi.org/10.1038/nrg2484.

[18] Poitou, C. *et al.* Bariatric surgery induces disruption in inflammatory signaling pathways mediated by immune cells in adipose tissue: A rna-seq study. *PLOS ONE* **10**, e0125718 (2015). URL https://doi.org/10.1371/journal.pone.0125718.

[19] Glastonbury, C. A., Couto Alves, A., El-Sayed Moustafa, J. S. & Small, K. S. Cell-type heterogeneity in adipose tissue is associated with complex traits and reveals disease-relevant cell-specific eqtls. *The American Journal of Human Genetics* **104**, 1013–1024 (2019). URL https://doi.org/10.1016/j.ajhg.2019.03.025.

[20] Picelli, S. *et al.* Full-length rna-seq from single cells using smart-seq2. *Nature Protocols* **9**, 171–181 (2014). URL https://doi.org/10.1038/nprot.2014.006.

[21] Macosko, E. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015). URL https://doi.org/10.1016/j.cell.2015.05.002.

[22] Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017). URL https://doi.org/10.1038/ncomms14049.

[23] Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015). URL https://doi.org/10.1038/nature14590.

[24] Habib, N. *et al.* Massively parallel single-nucleus rna-seq with dronc-seq. *Nature Methods* **14**, 955–958 (2017). URL https://doi.org/10.1038/nmeth.4407.

[25] Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* **347**, 1138–1142 (2015). URL https://doi.org/10.1126/science.aaa1934.

[26] La Manno, G. *et al.* Rna velocity of single cells. *Nature* **560**, 494–498 (2018). URL https://doi.org/10.1038/s41586-018-0414-6.

[27] Losic, B. *et al.* Intratumoral heterogeneity and clonal evolution in liver cancer. *Nature Communications* **11**, 291 (2020). URL https://doi.org/10.1038/s41467-019-14050-z.

[28] Kang, H. M. *et al.* Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature Biotechnology* **36**, 89–94 (2018). URL https://doi.org/10.1038/nbt.4042.

[29] Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453–457 (2015). URL https://doi.org/10.1038/nmeth.3337.

[30] Rahmani, E. *et al.* Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature Communications* **10**, 3417 (2019). URL https://doi.org/10.1038/s41467-019-11052-9.

[31] Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications* **11**, 1971 (2020). URL https://doi.org/10.1038/s41467-020-15816-6.

[32] Alvarez, M. *et al.* Enhancing droplet-based single-nucleus rna-seq resolution using the semi-supervised machine learning classifier diem. *Scientific Reports* **10**, 11019 (2020). URL https://doi.org/10.1038/s41598-020-67513-5.

[33] Alvarez, M. *et al.* Human liver single nucleus and single?cell rna sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival. *Genome Medicine* **14**, 50 (2022). URL https://doi.org/10.1186/s13073-022-01055-5.

CHAPTER 2

Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM

SCIENTIFIC REPORTS
nature research

OPEN

Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM

Marcus Alvarez[1,12], Elior Rahmani[2,12], Brandon Jew[3], Kristina M. Garske[1], Zong Miao[1,3], Jihane N. Benhammou[1,5], Chun Jimmie Ye[4], Joseph R. Pisegna[1,5], Kirsi H. Pietiläinen[6,7], Eran Halperin[1,2,9,10,11] & Päivi Pajukanta[1,3,8]✉

Single-nucleus RNA sequencing (snRNA-seq) measures gene expression in individual nuclei instead of cells, allowing for unbiased cell type characterization in solid tissues. We observe that snRNA-seq is commonly subject to contamination by high amounts of ambient RNA, which can lead to biased downstream analyses, such as identification of spurious cell types if overlooked. We present a novel approach to quantify contamination and filter droplets in snRNA-seq experiments, called Debris Identification using Expectation Maximization (DIEM). Our likelihood-based approach models the gene expression distribution of debris and cell types, which are estimated using EM. We evaluated DIEM using three snRNA-seq data sets: (1) human differentiating preadipocytes in vitro, (2) fresh mouse brain tissue, and (3) human frozen adipose tissue (AT) from six individuals. All three data sets showed evidence of extranuclear RNA contamination, and we observed that existing methods fail to account for contaminated droplets and led to spurious cell types. When compared to filtering using these state of the art methods, DIEM better removed droplets containing high levels of extranuclear RNA and led to higher quality clusters. Although DIEM was designed for snRNA-seq, our clustering strategy also successfully filtered single-cell RNA-seq data. To conclude, our novel method DIEM removes debris-contaminated droplets from single-cell-based data fast and effectively, leading to cleaner downstream analysis. Our code is freely available for use at https://github.com/marcalva/diem.

Single-cell RNA sequencing (scRNA-seq) has grown considerably in use over the previous decade and permitted a transcriptomic view into the composition of heterogeneous mixtures of cells[1,2]. Recent advances in droplet-based microfluidics have created a high-throughput opportunity to assay single cells by scaling up previous well-based technologies to tens to hundreds of thousands of cells[3]. Single-nucleus RNA sequencing (snRNA-seq), where nuclei are used instead of cells, has allowed the critical extension of single-cell based technologies to solid tissues where isolation and suspension of individual cells is difficult or impossible[4]. For example, snRNA-seq has been

[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA. [2]Department of Computer Science, School of Engineering, UCLA, Los Angeles, CA 90095, USA. [3]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA. [4]Department of Epidemiology and Biostatistics, Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, UCSF, San Francisco, USA. [5]Vache and Tamar Manoukian Division of Digestive Diseases, UCLA, Los Angeles, CA, USA. [6]Obesity Research Unit, Research Programs Unit, Diabetes and Obesity, University of Helsinki, Biomedicum Helsinki, Helsinki, Finland. [7]Obesity Center, Endocrinology, Abdominal Center, Helsinki University Central Hospital and University of Helsinki, Helsinki, Finland. [8]Department of Human Genetics, Institute for Precision Health, David Geffen School of Medicine at UCLA, Gonda Center, Room 6335B, 695 Charles E. Young Drive South, Los Angeles, CA 90095-7088, USA. [9]Department of Anesthesiology, UCLA Health, Los Angeles, CA 90095, USA. [10]Department of Computational Medicine, School of Medicine, UCLA, Los Angeles, CA 90095, USA. [11]Institute for Precision Health, School of Medicine, UCLA, Los Angeles, CA 90095, USA. [12]These authors contributed equally: Marcus Alvarez and Elior Rahmani. ✉email: ppajukanta@mednet.ucla.edu

used successfully to identify cell types in the brain[5]. Another practically important application of sequencing nuclei is identifying cell types in frozen tissue, from which it is often not feasible to isolate intact cells, whereas nuclei can still be successfully isolated[6].

Droplet-based snRNA-seq encapsulates individual nuclei into a water-in-oil emulsion that contains reagents for generating cDNA and ligating droplet-specific oligonucleotide barcodes. After library construction and sequencing, the mapped reads can be assigned to droplets of origin. The input nuclei suspension is prepared so that all reads associated with one barcode originate from one nucleus. However, RNA originating from lysed cellular components (such as the cytoplasm) or from outside the cell can become encapsulated in droplets as well. Since these reads have the same barcode, contaminated RNA cannot be readily distinguished from nuclear RNA. To apply snRNA-seq to tissues, homogenization of the tissue is usually required to break apart the extracellular matrix and release nuclei from cells[4]. This can release higher amounts of debris and lead to more background RNA contamination[7]. This contamination of droplets with extranuclear RNA can lead to a biased increase in expression of these genes. Using mitochondrial RNA, we show that this results in clusters driven by background RNA, as well as contamination of clusters representing true cell types. As droplet-based snRNA-seq is increasingly applied to various solid tissues, there is an urgent need to accurately filter contaminated droplets.

A common practice to distinguish cell/nuclei- vs. background-containing barcodes relies on removing droplets below a hard cutoff of the number of reads, unique molecular indexes (UMI), or genes detected in a droplet[3,8–11]. This ad hoc cutoff is typically set by ranking barcodes by their total UMI counts and visually selecting a knee point, where a steep dropoff in counts occurs[3,12]. Droplets with higher counts are expected to contain cells or nuclei, whereas droplets with lower counts are expected to contain ambient RNA. However, a clear separation between the two may not occur, especially if the amount of debris is high and the droplet RNA content is low, as we show is the case with frozen solid tissues. Additionally, an ad hoc cutoff of the percent of reads originating from the mitochondria (a measure of extranuclear contamination) can help to filter droplets[12]. Again, the choice of a cutoff may be arbitrary or unclear. The recent method EmptyDrops[12] addresses this filtering issue for scRNA-seq by estimating a Dirichlet-Multinomial distribution of the ambient RNA. It then removes droplets by testing if their expression profile deviates significantly from the ambient profile using a Monte Carlo approach[12]. However, while this works for single-cell, we show that these methods underperform when applied to snRNA-seq.

Here we show that, in snRNA-seq, using a hard cutoff to remove droplets can result in a substantial loss of nuclear droplets and inclusion of debris droplets. Importantly, we demonstrate that including these contaminated droplets can lead to spurious clustering and false positive cell types. To overcome this, we built a fast filtering pipeline that uses a likelihood-based approach to model debris and cell type RNA distributions with a multinomial distribution. The parameters of the model are inferred using semi-supervised EM[13,14], where all droplets below a hard count threshold are fixed as debris. Then, the droplets are scored based on their expression of genes enriched in the debris set. This multinomial-based clustering approach has been successfully applied to the information retrieval and text mining fields[15]. Similar to reads, word occurrences in a document can be modeled with a multinomial distribution, and documents can belong to separate topics, leading to a mixture model.

We developed this pipeline into an approach, termed Debris Identification using EM (DIEM), which robustly removes background droplets from both scRNA-seq and snRNA-seq data. In contrast to hard count and EmptyDrops filtering, DIEM accurately models debris and cell types and can quantify the amount of contamination in individual droplets. This resulted in more accurate filtering and higher quality clustering of snRNA-seq data, particularly when applied to frozen tissue. We also found that DIEM can effectively filter scRNA-seq data.

Results

snRNA-seq produces clusters driven by high amounts of background RNA contamination. Isolation of nuclei for snRNA-seq relies on lysis of the cell membrane, releasing cytoplasmic RNA, in addition to cell-free RNA, into the solution. This extranuclear RNA can become encapsulated into droplets, with or without nuclei, and lead to biases in downstream analysis; particularly, it may lead to spurious or contaminated cell-types in downstream clustering. We evaluated the extent of contamination and its effect on clustering in three distinct snRNA-seq data sets: 1. in vitro differentiating human preadipocytes (DiffPA) (n = 1), 2. freshly dissected mouse brain tissue (n = 1), and 3. frozen human subcutaneous adipose tissue (AT) (n = 6). We initially ran a clustering analysis in the three data sets by filtering out droplets with a hard-count threshold[3,8–11]. This threshold can be selected manually, as the knee point[3], or by dividing the total count of the 99% quantile of expected cells by 10[16]. Since we observed that the knee point could not be reliably estimated or was not evident in the AT samples (Fig. 1a), we used the quantile-based threshold for further analyses.

To assess levels of extranuclear RNA contamination, we primarily used the percentage of reads that are spliced in a droplet. The poly-T capture probes used in drop-seq [3] and the 10X platform can also hybridize to adenosine tracts present in introns, allowing for quantification of unspliced pre-mRNAs[17]. We expected that a higher percent of cytoplasmic ambient RNA would be spliced in comparison to nuclear RNA, and thus contaminated droplets would have a higher proportion of spliced reads. We found that in all three data sets, the percent of spliced reads correlated negatively with total UMI counts (Fig. S1a). Furthermore, we found that the percentage of reads spliced generally showed a bimodal distribution, with nuclear and background droplets centered below and above roughly 50%, respectively (Fig. S1b). For each of the 8 experiments, we calculated a midpoint to separate the nuclear and background distributions (see "Methods"). This was performed independently for each experiment as they exhibited distinct distributions (Fig. S1b). To evaluate clusters, we specified those with a mean percent of reads spliced of at least 50% as debris and classified those with less than 50% as cell types consistent with expressed marker genes, as we observed this was the average value across the experiments and that the 6 adipose tissue samples were combined. In addition to the percentage of reads spliced, we evaluated extranuclear contamination using the percentage of UMIs aligning to the mitochondria (MT%) and to the

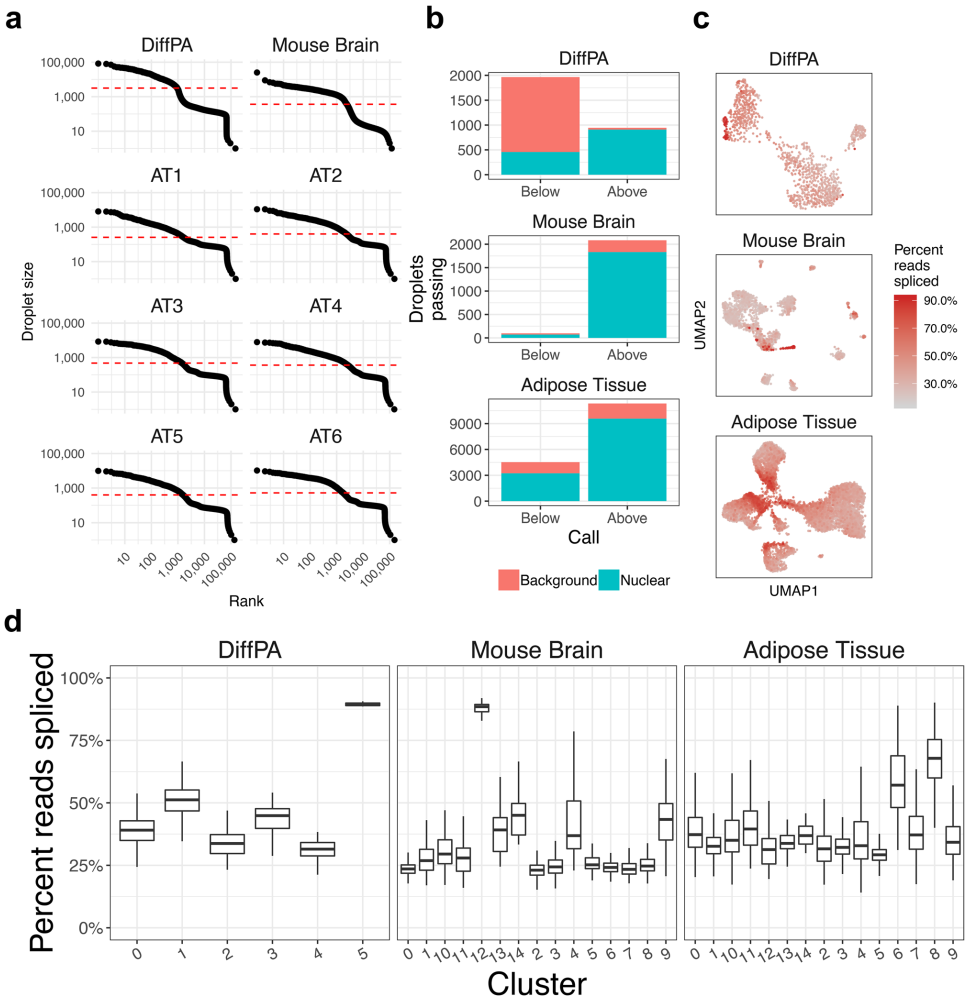


Figure 1. Applying a hard count threshold fails to remove droplets contaminated with background RNA in snRNA-seq. (**a**) Barcode-rank plots showing the droplet size (the total number of UMI read counts) of each droplet in descending order for the differentiating preadipocytes (DiffPA), mouse brain, and six human frozen adipose tissue (AT) snRNA-seq samples. The dotted red line indicates the quantile-based threshold. (**b**) The number of droplets above and below the quantile-based hard-count threshold is shown. The height of the red bar indicates the number of background droplets in the category indicated in the x-axis, while the height of the blue bar indicates the number of nuclear droplets. Background and nuclear droplets are defined using the percent spliced reads. Ideally, all nuclear droplets would occur above the threshold and all background droplets would occur below. (**c**) UMAP[33] visualization of droplets in each of the three data sets with droplets colored by the percent of reads spliced. (**d**) The droplets above the quantile threshold were clustered using Seurat[20]. The x-axis shows the clusters, and the y-axis shows the distribution of the percent of reads spliced for each cluster. Background droplets with a high percent of reads spliced tend to cluster together.

nuclear-localized lincRNA *MALAT1*[18] (MALAT1%). We chose to incorporate mitochondrial RNA as a measure extranuclear RNA contamination because it is one of the only true sources of background RNA and is present

in all snRNA-seq data sets. However, we note that other sources of extranuclear RNA can exist. Hemoglobin mRNA, which is predominantly expressed in erythrocytes, can also serve as another negative control for tissues where blood is present[19]. We also found that the percentage of reads spliced correlated positively with MT% and negatively with MALAT1% (Fig. S1c, d). As the percentage of spliced reads is more independent from gene expression, we primarily used this metric as an estimate of contamination within droplets.

To test whether a hard count threshold could effectively remove debris-contaminated nuclei, we investigated the relationship between total counts and the percent of reads spliced. We selected the hard count threshold for each of the 8 independent samples based on a quantile[16] (Fig. 1a). We found that this threshold failed to remove all background droplets and incorrectly removed nuclear droplets (Fig. 1b, c). For example, in the DiffPA dataset, the quantile threshold correctly kept a large proportion of nuclear droplets (908 of 944) but incorrectly removed 457 droplets (Fig. 1b). Of the 11,331 passing droplets in the 6 adipose tissue samples, only 9,578 (84.5%) droplets were nuclear (Fig. 1b). We found that no single count threshold could effectively discriminate the nuclear and background droplets (Fig. S1a). We further investigated the downstream effect on clustering to see if there was any evidence of background RNA driving spurious clusters. In the DiffPA, mouse brain, and adipose tissue data sets, there were 2, 1, and 2 clusters that had a mean percent reads spliced greater than 50%, respectively (Fig. 1d). Additionally, we observed droplets with a high MT% and clusters that were enriched for mitochondrial RNA (Fig. S2). Overall, a hard count threshold failed to discriminate nucleus-containing droplets from debris droplets when using percent reads spliced and MT% to quantify contamination.

Nuclear and debris droplets demonstrate distinct RNA profiles. Since the total UMI count in a droplet does not always distinguish nuclei from debris, we postulated that the expression profile of a droplet could be used to differentiate them if there were sufficient differences in RNA abundance between cell types and debris. Specifically, we hypothesized that there would be genes with sub-cellular localized RNA products that show differential abundance between droplets containing nuclear vs. ambient RNA. Thus, we evaluated the extent of differences between the debris and nuclear RNA profiles. We separated droplets into debris- and nuclear-enriched groups using a threshold of 100 total UMI counts. Although a large number of droplets above 100 UMI counts consist of debris and would lead to a loss of power, we use this threshold to ensure that no droplets below it contain nuclei. We evaluated the difference between the debris and nuclear RNA profiles by running a paired differential expression (DE) analysis in the six human AT samples. Of 19,934 genes detected, 3,417 (17.1%) were DE between the nuclear- and debris-enriched groups at a Bonferroni-adjusted p-value threshold of 0.05 (Fig. 2a). To see if these differences were preserved across the DiffPA, mouse brain, and six AT data sets, we correlated the nuclear vs. debris log fold changes of the genes in common. Among the 8,924 genes expressed in all three data sets, we found that all log fold changes were significantly correlated ($p < 2.2 \times 10^{-16}$) across all pairs (mean $R = 0.56$), with the human data sets showing the highest correlations (Fig .S3).

Since the nuclear-enriched group is not homogeneous, but rather originates from distinct cell types with different RNA distributions, we also looked at differences between the debris group and cell types. In addition, we compared the cell type-debris differences with the cell type-cell type differences. Using the six AT samples, we ran a paired DE analysis between the cell types and debris droplets (total UMI counts < 100). Among 14 debris-cell type pairs, the average percent of genes that are DE was 5.8% (Fig. 2b). We then compared this to the DE between a cell type and all other cell types. Among these 14 pairs, the average percent of genes DE between cell types was slightly lower at 4.5% (t-test $p = 0.23$; Fig. 2b, c). Overall, we found significant differences between debris and nuclei RNA profiles, and that the differences between debris and cell types were within the same order of magnitude as the cell type-cell type differences.

Overview of a novel EM-based approach to cluster and remove debris droplets from snRNA-seq data. Since we observed differences in RNA abundance between cell types and debris, we developed an approach to remove debris-containing droplets based on the distribution of read counts. Our approach assigns individual debris scores to filter out droplets. We first cluster droplets using a multinomial mixture model. To estimate the parameters of the mixture model, we run semi-supervised expectation maximization[13,14] by fixing droplets that fall below a threshold of 100 counts as debris. The majority of these droplets are assumed to contain ambient RNA, and thus we leverage this feature by fixing the labels throughout EM. After fitting the model, we assign droplets to clusters based on their posterior probability. Then, droplets are scored based on their expression of genes enriched in the debris set. DIEM then filters out droplets based on their individual scores. Figure 3a shows an overview of this model. We termed this method Debris Identification using Expectation Maximization (DIEM). We ran DIEM on the DiffPA, mouse brain, and six AT sample and compared our approach with the quantile-based method and the EmptyDrops method in the DropletUtils package[12].

Although debris scores are used to filter out individual droplets, we run clustering to better initialize the debris and cell type groups. Droplets are clustered using a multinomial mixture model and the parameters are fit using semi-supervised EM. To initialize the EM, we run k-means with a pre-specified number of cell types k. After the initialization, semi-supervised EM estimates the parameters of the multinomial mixture model while fixing the labels of the low-count droplets to the debris cluster. The mixture model consists of $k + 1$ clusters corresponding to the debris cluster along with the cell type clusters initialized by k-means (Fig. 3a). Here, we set k to 20 for all experiments, although we noticed robust results across a range of k greater than 1 to 50 (Fig. S4). While it is possible to remove droplets that have high posterior probability of belonging to the debris cluster, we noticed that some of the cell type clusters produced by the mixture model contained contaminated droplets with high a percent of reads spliced in the snRNA-seq data sets (Fig. S5). Since only removing the debris cluster would fail to account for this, we developed an approach to estimate contamination in individual droplets instead (see "Methods"). Briefly, DIEM runs differential expression between the droplets in the debris and cell type clusters

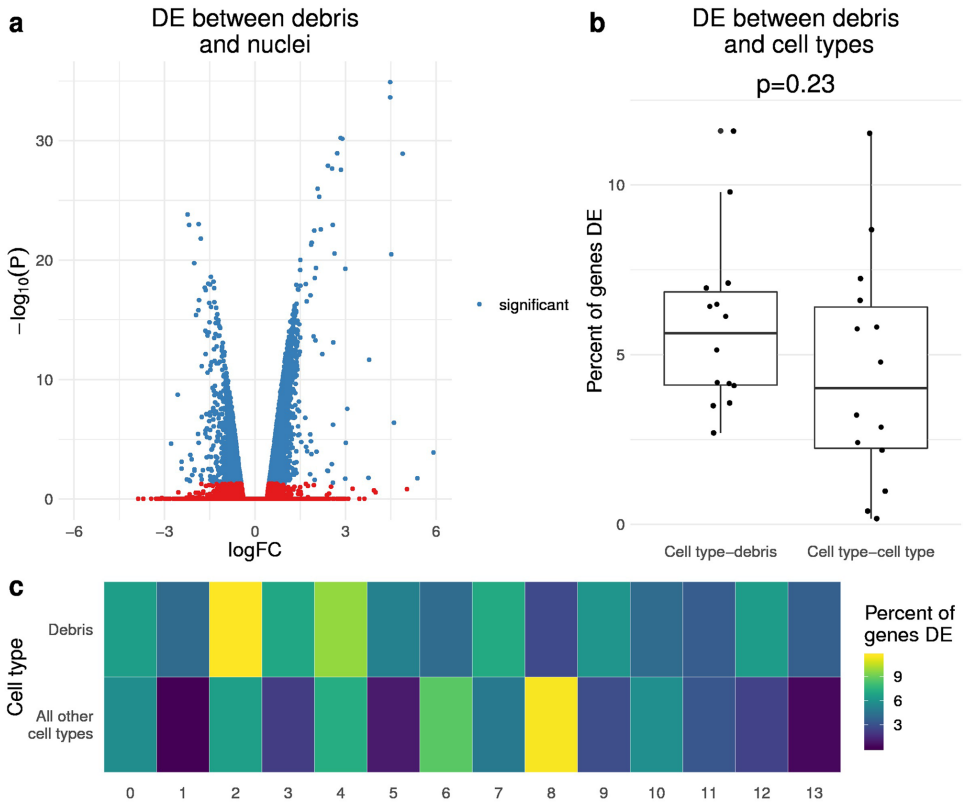


Figure 2. Debris-containing and nuclei-containing droplets show distinct gene expression profiles. (**a**) Differential expression (DE) between droplets with less than 100 UMI counts (debris) and greater than or equal to 100 UMI counts (nuclei) in the 6 human adipose tissue (AT) samples. The volcano plot shows the log fold change on the x-axis and negative log transformed p-value on the y-axis. The genes colored in blue are DE with a Bonferroni-corrected p-value < 0.05. A positive log fold change indicates over-expression in the debris group. (**b**, **c**) For each of the 14 cell types identified after clustering the quantile filtered droplets, we ran differential expression between the cell type and the debris group, or between the cell type and all other cell types in the combined adipose tissue data set. Cell types are estimated from clustering droplets that pass quantile-based filtering. A (**b**) box plot shows the percent of expressed genes that are DE (Bonferroni p < 0.05) between a cell type-debris pair, and a cell type-cell type pair. The p-value was calculated from a student's t-test between cell type-debris percent and cell type-cell type percent. The (**c**) heatmap shows the percent of total genes expressed in the cell type (x-axis column) that are significantly differentially expressed between the debris droplets (first row) or droplets in all other cell types (second row). This shows that the DE between a cell type and the debris group is similar to the DE between different cell types.

and calculates a debris score based on the debris-enriched genes. We found that this debris score correlated highly with the percent of spliced reads in all 8 independent snRNA-seq experiments (mean Pearson $R = 0.89$; Fig. 3b). To filter our droplets, we use a threshold t where those with a score above this value are removed. We investigated the effect of varying t from 0 to 1. As expected, the number of passing droplets increased with t (Fig. S6). However, the proportion of background droplets and contamination in the kept droplets also increased as t was increased (Fig. S6). We therefore set t to 0.5 for all experiments in the manuscript. Intuitively, this value represents the threshold that lies between the least contaminated cluster and the debris cluster.

The incorporation of clusters should result in a more realistic model of the snRNA-seq data. DIEM directly models debris and cell type clusters to more accurately specify the debris and cell type droplets for calculating the debris score. We asked whether the clusters identified by DIEM corresponded to valid biological cell types. DIEM identified 3 major cell types in the DiffPA, consisting of preadipocyte-like, fibroblast-like, and adipocyte

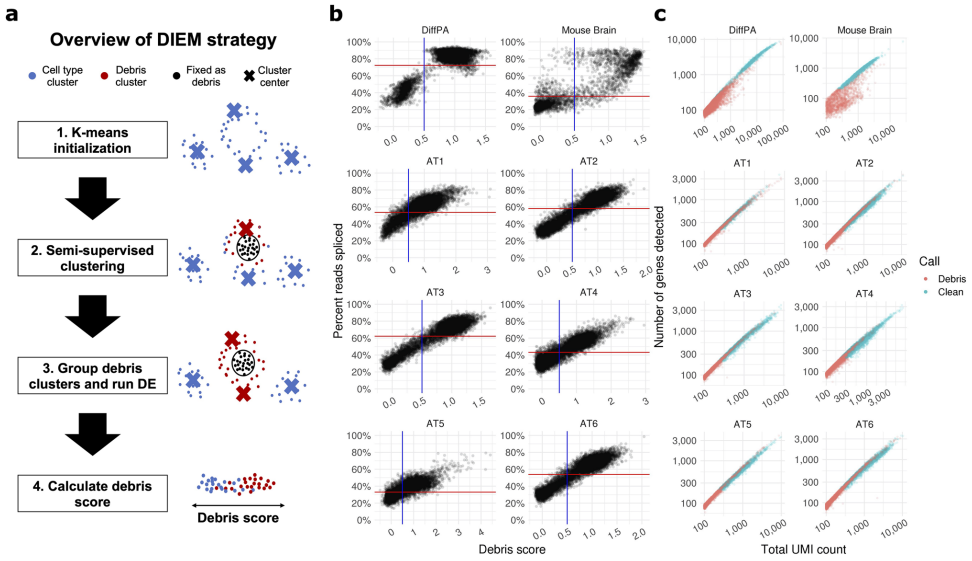


Figure 3. Debris scoring predicts background RNA contamination in snRNA-seq droplets. (**a**) Overview of DIEM approach to remove debris-contaminated droplets. Expectation Maximization (EM) is used to estimate the parameters of a multinomial mixture model consisting of debris and cell type groups. The label assignments of droplets below a pre-specified threshold (100 total counts) are fixed to the debris group, while the test set droplets above this rank are allowed to change group membership. The mixture model is initialized by running k-means. After parameter estimation, droplets are grouped into the debris cluster(s) or cell type clusters based on their posterior probabilities. Debris scores are calculated for each droplet by summing the normalized expression of debris-enriched genes, which are specified by differential expression between the debris and cell type clusters. Droplets can be filtered based on their cluster assignment or on their debris score. (**b**) The debris score of a droplet and the percent of reads spliced exhibit a significant correlation in the differentiating preadipocytes (DiffPA), mouse brain, and human frozen adipose tissue (AT) data sets (mean $R = 0.89$). The horizontal red line indicates the sample-specific midpoint that separates nuclear and background droplets. The vertical blue line indicates the threshold cutoff of 0.5 we used, where droplets with a debris score less than 0.5 are classified as clean. **c,** Scatterplots of droplets from snRNA-seq of the DiffPA, mouse brain, and AT data sets, with total unique molecular index (UMI) counts on the x-axis and total number of genes detected on the y-axis. Droplets are colored by the DIEM classification. Those in red are removed as debris while the blue droplets are kept as nuclei.

cells (Fig. S7a). In the adipose tissue data sets, we found that the up-regulated cluster markers corresponded to the known major cell types in adipose, including immune, endothelial, fibroblast, and adipocyte cell-types (Fig. S7b). We then compared the DIEM clusters to those identified by the established method Seurat[20] We found that the Seurat clusters generally overlapped with the DIEM clusters (mean percent overlap 73.0% across the 8 independent samples; Fig. S7c). Together, these results suggest that DIEM accurately identifies cell types and can leverage this cell-type heterogeneity to filter debris droplets.

DIEM filtering results in a higher proportion of nuclear droplets and less contaminated clusters in snRNA-seq. We ran DIEM on the adipocyte, mouse brain, and six adipose tissue samples. We observed that DIEM removed droplets across a range of total UMI counts (Fig. 3c). We then evaluated the extent of extra-nuclear contamination, as well as its effect on clustering, across the quantile, EmptyDrops, and DIEM methods. We first quantified the number of nuclear and background droplets that passed filtering in each of the 8 experiments. In the DiffPA data set, the DIEM and quantile methods kept a larger proportion of nuclear droplets. Among the passing droplets, 1,337 of 1,339 (99.9%), 1,360 of 1,579 (86.1%), and 908 of 944 (96.2%) were nuclear in the DIEM, EmptyDrops, and quantile droplets, respectively (Fig. 4a). In the mouse brain data set, all three methods produced similar results. We found that 1,850 of 2,010 (92.0%), 1,868 of 2,080 (89.8%), and 1,832 of 2,083 (87.6%) passing droplets were nuclear in the DIEM, EmptyDrops, and quantile droplets, respectively (Fig. 4a). Across all 6 adipose tissue samples, 12,117 of 12,715 (95.7%), 10,110 of 11,502 (87.9%), and 9,578 of 11,331 (84.5%) passing droplets were nuclear in the DIEM, EmptyDrops, and quantile droplets, respectively (Fig. 4a). We further investigated these filtering methods in each of the adipose tissue samples. We found that the percent of DIEM passing droplets that were nuclear was significantly higher when compared to EmptyDrops

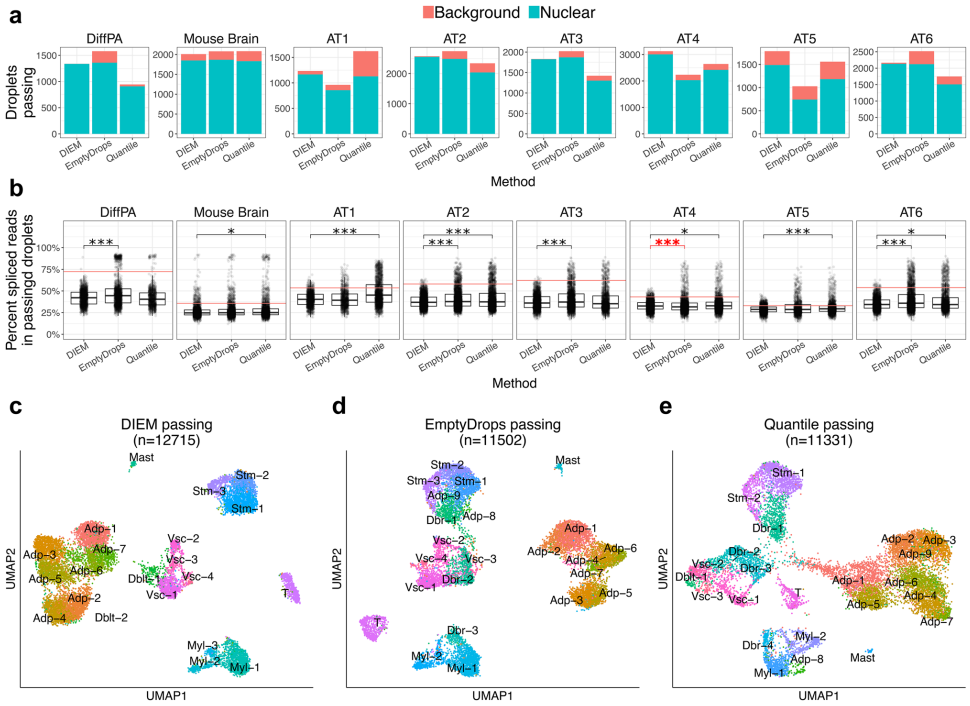


Figure 4. DIEM filtering keeps an increased number and proportion of nuclear droplets in snRNA-seq. (**a**) The bar plots show the number and type of droplets that pass the indicated filtering method in the differentiating preadipocytes (DiffPA), mouse brain, and six human frozen adipose tissue (AT) snRNA-seq samples. The height of the blue bar indicates the number of nuclear droplets that pass filtering, while the height of the red bar indicates the number of background droplets. DIEM filtering tends to result in a higher number and proportion of nuclear droplets. Background and nuclear droplets are defined using the percent spliced reads. (**b**) The percent of reads spliced is shown in a boxplot for droplets that pass the indicated filtering method in the DiffPA, mouse brain, and six AT snRNA-seq samples. The horizontal red line indicates the sample-specific midpoint, where droplets above and below are background and nuclear, respectively. A Mann-Whitney U test was performed between DIEM and EmptyDrops[12], and DIEM and quantile-filtered droplets. DIEM shows a decrease in percent spliced reads for all comparisons (black bar and asterisks) except for AT4 with EmptyDrops (red bar and asterisks). P-values were corrected for multiple testing using Bonferroni and are shown in the upper portion of the plot (*$p<0.05$; **$p<0.005$; ***$p<0.0005$). (**c**) UMAP[33] visualization of clusters after filtering with the indicated method in the combined adipose tissue snRNA-seq data set. Clusters were identified with Seurat[20] and classified as adipocyte (Adp), doublet (Dblt), myeloid (Myl), T cell, mast, and stromal (Stm) cell types according to their up-regulated genes. A cluster was classified as debris (Dbr) if it had a mean percent of spliced reads above 50%.

and the quantile approach (paired Wilcoxon $p=0.03$). Overall, the DIEM method tended to keep a higher number and proportion of nuclear droplets in the 8 snRNA-seq experiments.

Next, we compared the fraction of spliced reads in individual droplets for the independent experiments using a Mann–Whitney U test (Fig. 4b). When compared to EmptyDrops, DIEM had a significantly lower spliced reads fraction for the DiffPA, AT2, AT3, and AT6 samples. Although DIEM droplets had a lower mean of percent spliced reads for the AT4 sample, the Mann–Whitney U test yielded a significantly higher rank for the spliced reads fraction when compared to EmptyDrops (Bonferroni-corrected $p<0.05$; Fig. 4b). When compared to the quantile-passing droplets, DIEM had a significantly lower spliced reads fraction for the mouse brain and 5 of the 6 adipose tissue samples (Bonferroni-corrected $p<0.05$; Fig. 4b). None of the quantile-filtered droplets produced samples with a lower percent of reads spliced than DIEM, suggesting the quantile droplets contain more ambient RNA. Taken together, these results suggest that the DIEM-passing droplets comprise more nuclear droplets when using the percent of reads spliced as a measure of contamination.

We then looked at the effect of filtering on clustering results. We clustered passing droplets using Seurat[20] to unbiasedly evaluate the clustering results based on each of the three methods. We considered clusters with a

mean percent spliced reads of at least 50% as debris clusters and classified those with less than 50% as cell types consistent with expressed marker genes (see "Methods"). In the DiffPA dataset, DIEM removed a debris cluster that was present after filtering with both the quantile and EmptyDrops methods (Fig. S8a, b). Additionally, both EmptyDrops and DIEM identified a low count cluster that showed evidence of containing nuclei (Fig. S9), highlighting how a hard count threshold can result in removing cell types with lower counts. In the mouse brain data set, both EmptyDrops and DIEM removed a background cluster (median spliced reads 88.5%) that was present in the quantile method (Fig. S8c, d). For the adipose tissue data set, we combined the 6 individual filtered data sets and ran Seurat clustering with a resolution value of 2 to accommodate the larger number of droplets. DIEM resulted in 21 clusters, while EmptyDrops and the quantile method yielded 23 clusters. We then classified clusters based on their marker genes (Figs. 4c–e and S10). None of the DIEM clusters had a mean percent of reads spliced above 50% (Fig. S11a). However, EmptyDrops filtering resulted in 3 debris clusters while the quantile approach yielded 4 (Fig. S11a). There was a high overlap of over 50% in the major cell types between all 3 methods, but clusters consisting of smaller numbers of droplets tended to be spread across related cell types (Fig. S11b). Overall, these results suggest that DIEM preserves cell types while removing clusters characterized by high extranuclear contamination when compared to the EmptyDrops and quantile approaches.

DIEM filtering removes a higher proportion of contaminated droplets and clusters in snRNA-seq. We next investigated whether filtering incorrectly removed nucleus-containing droplets and possibly true cell types. Similar to the above analysis, we quantified the number of nuclear and debris droplets that were removed by each of the three methods. In the DiffPA dataset, 1,542 of 1,570 (98.2%), 1,325 of 1,330 (99.6%), and 1,508 of 1,965 (76.7%) removed droplets were background droplets after DIEM, EmptyDrops, and quantile filtering, respectively (Fig. 5a). The DIEM and EmptyDrops methods performed similarly in the mouse brain data set, and both outperformed the quantile approach. We found that 117 of 171 (68.4%), 65 of 101 (64.4%), and 26 of 98 (26.5%) removed droplets were classified as background in the DIEM, EmptyDrops, and quantile filtering, respectively (Fig. 5a). Across all adipose tissue samples, we found that the DIEM-removed droplets consisted of a higher proportion of background-derived droplets. We found that 2,499 of 3,140 (79.6%), 1,651 of 4,353 (37.9%), 1,290 of 4,524 (28.5%) removed droplets were background droplets in the DIEM, EmptyDrops, and quantile filtering, respectively (Fig. 5a). We investigated the percent of background droplets in those removed in each of the 6 adipose tissue samples as well. The percent of the removed droplets that were background was significantly higher with DIEM when compared to EmptyDrops and the quantile approach (paired Wilcoxon $p = 0.03$). We found that EmptyDrops incorrectly removed a much higher number of nuclear droplets in the AT4 and AT5 samples (Fig. 5a). EmptyDrops filtered out 1,038 in the AT4 and 782 in the AT5 samples, whereas DIEM removed 63 and 36 nuclear droplets, respectively (Fig. 5a). Overall, DIEM tended to remove a higher number and proportion of background droplets than EmptyDrops or the quantile approach.

We next investigated the amount of extranuclear contamination in the individual filtered-out droplets using a Mann–Whitney U test. We found that DIEM removed more background droplets with a significantly higher percent of reads spliced in all 8 experiments when compared to the quantile approach (Bonferroni-corrected $p < 0.05$; Fig. 5b). When compared to EmptyDrops, DIEM-removed droplets had a significantly higher percent of spliced reads for 5 of the 6 adipose tissue samples (Bonferroni-corrected $p < 0.05$; Fig. 5b). Neither the EmptyDrops nor the quantile method resulted in significantly more contamination in the removed debris droplets than DIEM. These results suggest that the DIEM-removed droplets contained fewer nuclei when using the percent of reads spliced as a measure of contamination.

Among the droplets removed by the three filtering methods, we sought further evidence that they originated from cell types. We clustered the removed droplets in the adipose tissue and looked to see if they consisted of biological cell types. We again considered clusters with a mean percent of reads spliced of at least 50% as debris clusters and classified those with less than 50% as cell types consistent with expressed marker genes (Figs. S1 and S12). Among the 8 clusters present in the DIEM-removed droplets, all had an average percent of reads spliced above 50%, suggesting that these consist of largely contaminated droplets (Figs. 5c and S12a). The EmptyDrops-removed droplets formed 11 clusters, 6 of which were debris. The other 5 clusters consisted of adipocyte, vascular, and stromal cell types (Figs. 5d and S12a). The quantile-removed droplets formed 7 cell type and 2 debris clusters. The cell type clusters consisted of adipocyte, stromal, T cell, myeloid, and mast cell types (Figs. 5e and S12a). Taken together, we found that the clusters formed by DIEM-removed droplets had more extranuclear contamination than those from EmptyDrops and the quantile method.

Interestingly, we found that the debris clusters formed by all filtering methods exhibited cell type-specific expression (Fig. S12b–e). This suggests that nucleus-containing droplets exhibit a range of extranuclear contamination in snRNA-seq experiments from frozen tissue. Furthermore, we found that droplets that had high read counts of the macrophage marker *CD14*[21] tended to have higher extranuclear contamination and were more often filtered out by DIEM (Figs. S10f and S12d). This suggests that *CD14+* macrophages are more susceptible to damage or contamination and may imply that nuclei isolation or the snRNA-seq assay may introduce a bias in cell type capture.

DIEM filtering removes debris from single-cell RNA-seq. In addition to filtering snRNA-seq, we also investigated whether our approach could be applied to single-cell RNA-seq data. We found that the debris scoring approach of individual droplets did not effectively distinguish empty vs. cell droplets in the 68,000 PBMC single-cell RNA-seq experiment[16] (Fig. S13). DIEM gave a high debris score to a cell type cluster with high read counts, suggesting the debris-enriched genes and thus the debris score were less specific in discriminating debris droplets from all cell types in this PBMC single-cell RNA-seq data set. Although the threshold could be increased to accommodate this cell type, we found that simply removing droplets belonging to the fixed debris

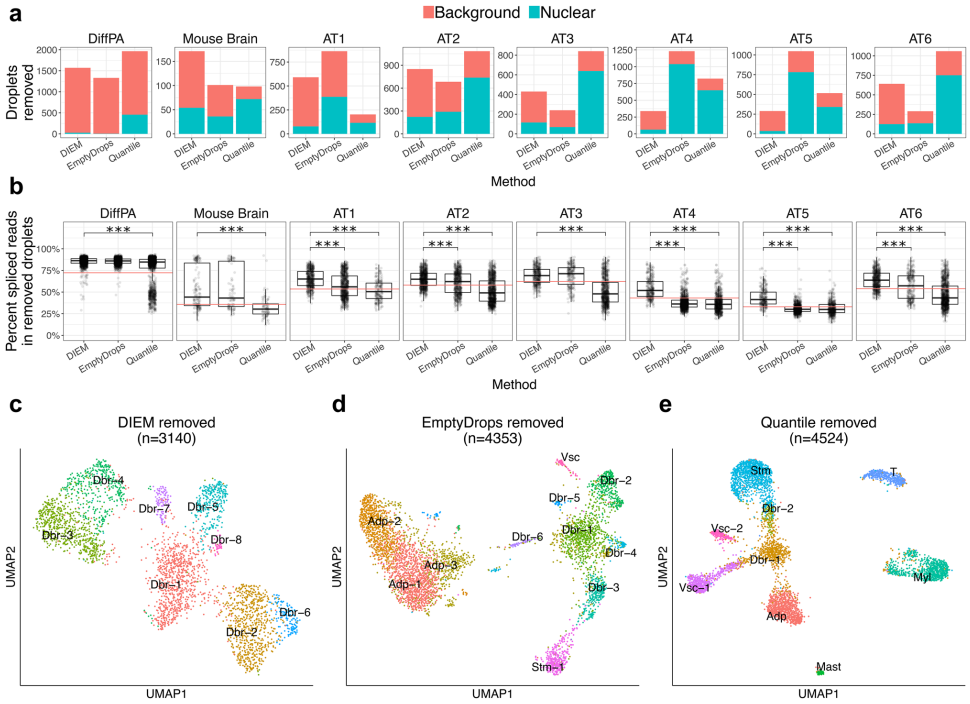


Figure 5. DIEM filtering removes fewer numbers of nuclei in snRNA-seq. (**a**) The bar plots show the number and type of droplets that are removed by the indicated filtering method in the differentiating preadipocytes (DiffPA), mouse brain, and six human frozen adipose tissue (AT) snRNA-seq samples. The height of the blue bar indicates the number of nuclear droplets that are removed while the height of the red bar indicates the number of background droplets. Background and nuclear droplets are defined using the percent spliced reads. DIEM filtering tends to result in a higher number and proportion of nuclear droplets. Removal of large numbers of nuclear droplets and low numbers of background droplets indicates poor performance. (**b**) The percent of reads spliced is shown in a boxplot for droplets removed by the filtering method in the DiffPA, mouse brain, and six AT snRNA-seq samples. The horizontal red line indicates the sample-specific midpoint, where droplets above and below are background and nuclear, respectively. A Mann-Whitney U test was performed between DIEM and EmptyDrops[12], and DIEM and quantile removed droplets. DIEM shows an increase in percent of reads spliced for all comparisons. P-values were corrected for multiple testing using Bonferroni and are shown in the upper portion of the plot (*$p < 0.05$; **$p < 0.005$; ***$p < 0.0005$). (**c**) UMAP[33] visualization of clustering of removed droplets with the indicated method in the combined adipose tissue snRNA-seq data set. Clusters were classified as adipocyte (Adp), doublet (Dblt), myeloid (Myl), T cell, mast, and stromal (Stm) cell types according to their up-regulated genes. A cluster was classified as debris (Dbr) if it had a mean percent of spliced reads above 50%.

cluster was effective in removing empty droplets in the single cell RNA-seq data. Both DIEM and EmptyDrops kept all 69,981 droplets that had at least 200 genes detected. To evaluate the effect of filtering, we removed this threshold to more completely characterize the two methods. Among 78,024 comparable droplets with at least 100 UMI counts, EmptyDrops kept 77,585 and DIEM kept 75,847. The 1,927 droplets unique to EmptyDrops showed a high percent of reads aligning to the mitochondria and to *MALAT1* (Fig. 6a). The 189 unique droplets to DIEM showed MT% and MALAT1% levels similar to the shared droplets that passed both filtering methods (Fig. 6a). Although these metrics no longer serve as negative and positive controls as they do in snRNA-seq, they are consistent with a ruptured cell membrane. This suggests that EmptyDrops retains droplets with dying cells whereas DIEM removes them. We next evaluated the clusters formed by these droplets. DIEM-filtered droplets formed 18 clusters while EmptyDrops resulted in 19 clusters. As expected, there was a general one-to-one correspondence between the clusters. However, the droplets with high MT% and MALAT1% formed a cluster that was absent in the DIEM results (Fig. 6b, c). Overall, we found that EmptyDrops and DIEM provided similar results in the PBMC single-cell RNA-seq data.

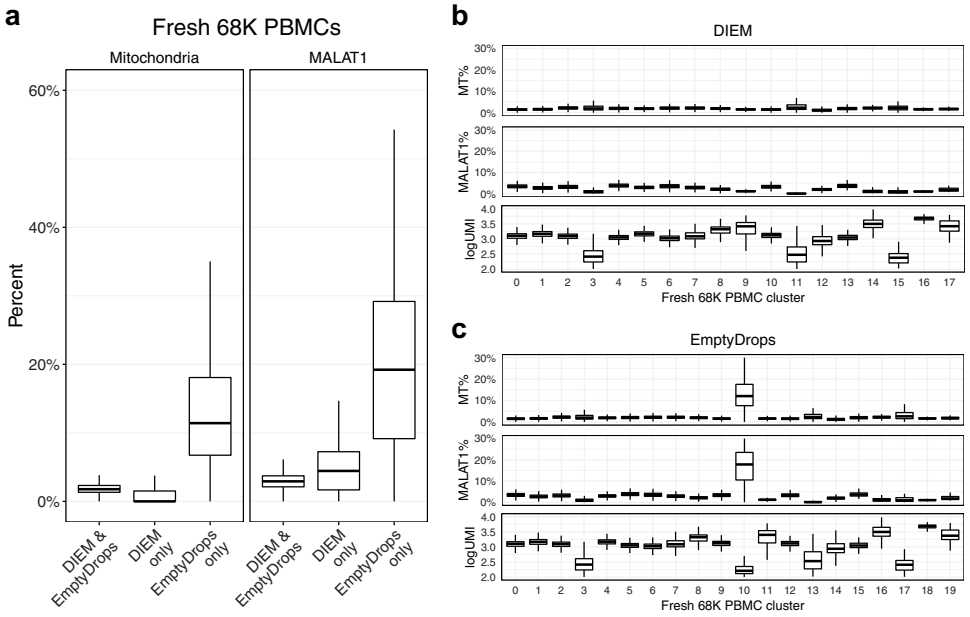


Figure 6. DIEM filtering in single-cell RNA-seq of fresh PBMCs results in robust cell type identification. (**a**) Boxplots showing the percent of unique molecular indices (UMIs) mapping to the mitochondria (left) and the percent of MALAT1 UMIs (right) in the fresh 68 K peripheral blood mononuclear cells (PBMC) data set[16]. The DIEM and EmptyDrops[12] set includes the droplets identified by both DIEM and EmptyDrops (n = 75,658), while the EmptyDrops only set (n = 1,927) and the DIEM only set (n = 189) include droplets uniquely kept by each method. The droplets uniquely kept by EmptyDrops have a higher percent of reads aligned to the mitochondrial and MALAT1 genes, consistent with a ruptured cell membrane. (**b**, **c**) Boxplots show the percent of UMIs aligning to the mitochondrial genome (MT%), to the nuclear-localized MALAT1[18] (MALAT1%), and the log total number of UMIs in a droplet for clusters in the PBMC single-cell RNA-seq data set. These measures are plotted for the (**b**) clusters from the DIEM-kept droplets and the (**c**) clusters from the EmptyDrops-kept droplets. Clusters were identified with Seurat[20]. The droplets uniquely kept by EmptyDrops form a distinct cluster with high MT% and MALAT1%.

Discussion

The snRNA-seq approach is an adaptation of scRNA-seq that allows for cell-type identification when isolation of a cell suspension is not possible, such as in frozen tissues. We have shown here that snRNA-seq is subject to background contamination from extranuclear RNA and that it can drive spurious clusters and false positive cell-types if not properly accounted for. We also show that current methods, such as the commonly applied hard count threshold, do not effectively address this problem in snRNA-seq. To this end, we searched for and found differences in the gene expression profiles from the debris and cell types. This motivated us to develop DIEM in order to use the RNA profile of a droplet for filtering contaminated snRNA-seq experiments. We found that DIEM efficiently removed debris-contaminated droplets while preserving cell types in snRNA-seq data from fresh cells, fresh tissue, and frozen tissue inputs.

DIEM first clusters all droplets generated from the experiment to identify droplets belonging either to the debris cluster or putative cell type clusters. This allows for a general separation of droplets into debris and cell type groups, and thus more accurate differential expression analysis between background and cell type droplets. Although it is possible to simply remove droplets that cluster as debris, we found that clusters with high amounts of contamination still existed in snRNA-seq. Therefore, scoring and filtering individual droplets allow for finer classification as well as quantification of the amount of contamination. The debris score threshold that removes droplets can be adjusted according to the desired tolerance for contamination. In addition, this estimate can be used as a covariate in downstream analyses, such as clustering and differential expression analyses. The debris score, however, cannot be assumed to exist on the same scale in independent experiments. The scores are normalized relative to the best and worst cluster means within the sample. Thus, when integrating multiple samples, the debris scores may not be comparable, particularly if the distribution of extranuclear RNA is different across the samples. Finally, the scoring approach relies on accurate estimation of debris-enriched genes. If there

are little or no genes that are increased in the background distribution, the resulting debris score will likely be inaccurate. Although we found that the debris scores in all 8 snRNA-seq experiments were correlated with the level of extranuclear contamination as measured by the percent of reads spliced, this was less successful in the PBMC single-cell RNA-seq data. This may be due to smaller differences between the background and cell type profiles in scRNA-seq.

Since we found that extranuclear RNA contamination exists across a wide range of UMI counts, using a hard count threshold underperformed in comparison to EmptyDrops and DIEM. Both DIEM and EmptyDrops[12] remove droplets based on their expression distribution. When compared to the EmptyDrops method[12], however, we found that DIEM had a higher accuracy in filtering heterogeneous snRNA-seq data sets as assessed by the percent of reads that are spliced as a metric of extranuclear RNA. EmptyDrops, however, was originally developed and tested on single cell data and thus, the assumptions behind the model are different than that of DIEM. EmptyDrops only models the background RNA distribution and uses Monte Carlo sampling to determine how significant the deviation of a droplet is from it. It also safeguards from removing cell-types that are similar to the background by assuming that all droplets above a calculated knee point are true cell-containing droplets. DIEM directly models cell types by clustering and this may allow for more accurate grouping of debris droplets. We have shown that the difference between the cell types and the debris are within the same order of magnitude as the differences between the cell types, highlighting the need to account for heterogeneity. We found that both EmptyDrops and the quantile approach removed more nuclear droplets and kept a higher proportion of contaminated droplets, but the major adipose cell types were still identified. However, since the DIEM-filtered droplets contained less extranuclear contamination, the resulting clusters were also characterized by less debris on average. This is beneficial for both accurate cell type clustering and identification.

Even though snRNA-seq recovers less RNA than scRNA-seq and thus retrieves less information about cell types, there are advantages to using nuclei over cells. For example, snRNA-seq has been shown to reduce dissociation biases present in scRNA-seq, leading to more accurate profiling of cell types in tissue[22]. Another important reason to use snRNA-seq is that scRNA-seq may be practically impossible. This can occur with frozen tissues, since thawing cells is known to lyse the outer membranes and preclude a suspension of single cells required for droplet-based technologies[3]. This prevents the application of scRNA-seq to biobanked snap-frozen human tissues. In order to leverage existing, phenotyped human datasets with biobanked tissues, snRNA-seq may be the only viable option to profile cell types. We have shown that snRNA-seq of frozen tissue results in contamination of droplets across a large range of UMI counts, making it difficult to remove background debris while maintaining an accurate cell type composition of the tissue. Even from fresh tissue and cells, we still observed downstream clusters affected by the extranuclear RNA. Therefore, we expect DIEM to help produce cleaner snRNA-seq data sets from a variety of input sources, but especially from frozen tissues.

We focused the application of our approach on snRNA-seq data because there is a pressing need for debris filtering in data sets with lower RNA content. In single-cell RNA-seq, the higher RNA content of cells typically allows the total UMI count of a droplet to serve as a sufficient discriminator between debris and cells[3], although this may not always be the case[12]. However, running scRNA-seq on fresh human tissue at a large scale may be prohibitively difficult considering the requirement to immediately process a fresh biopsy for scRNA-seq. Therefore, snRNA-seq of frozen tissues offers a viable alternative to process samples at a higher throughput. Our method was designed to computationally remove background debris contamination from snRNA-seq data of frozen tissues. We expect that DIEM will enable the analysis of a larger number of samples from frozen tissue snRNA-seq data, thereby removing the need to coordinate the acquisition of fresh tissue samples and processing of single cell libraries.

Methods

Single-nucleus RNA-seq of human subcutaneous adipose tissue, differentiating preadipocytes, and mouse brain. Frozen subcutaneous adipose tissue was processed separately for each of the 6 samples. Tissue was minced over dry ice and transferred into ice-cold lysis buffer consisting of 0.1% IGEPAL, 10 mM Tris–Hcl, 10 mM NaCl, and 3 mM $MgCl_2$. After a 10 min incubation period, the lysate was gently homogenized using a dounce homogenizer and filtered through a 70 µm MACS smart strainer (Miltenyi Biotec #130-098-462) to remove debris. Nuclei were centrifuged at 500×*g* for 5 min at 4 °C and washed in 1 ml of resuspension buffer (RSB) consisting of 1X PBS, 1.0% BSA, and 0.2 U/µl RNase inhibitor. We further filtered nuclei using a 40 µm Flowmi cell strainer (Sigma Aldrich # BAH136800040) and centrifuged at 500×*g* for 5 min at 4 °C. Pelleted nuclei were re-suspended in wash buffer and immediately processed with the 10X Chromium platform following the Single Cell 3′ v2 protocol. After library generation with the 10X platform, libraries were sequenced on an Illumina NovaSeq S2 at a sequencing depth of 50,000 reads per cell. Reads were aligned to the GRCh38 human genome reference with Gencode v26 gene annotations[23] using the 10X CellRanger 2.1.1 pipeline. A custom pre-mRNA reference was generated to account for unspliced mRNA by merging all introns and exons of a gene into a single meta-exon.

We obtained and cultured the primary human white preadipocyte cells as recommended by PromoCell (PromoCell C-12731, lot 395Z024) for preadipocyte growth and differentiation into adipocytes. Cell media (PromoCell) was supplemented with 1% penicillin–streptomycin. We maintained the cells at 37 °C in a humidified atmosphere at 5% CO2. On day 6 of differentiation, we rinsed the cells with 1 × PBS and added ice-cold lysis buffer (3 mM $MgCl_2$, 10 mM Tris–HCl, 0.5% Igepal CA-630, 10 mM NaCl). The cells were gently scraped from the plate and centrifuged at 500×*g* for 5 min at 4 °C. Nuclei were washed with 1 ml of resuspension buffer (RSB; 1% BSA, 100 µl RNase inhibitor in 1 × PBS) and centrifuged again to remove cellular debris. After the second centrifugation, nuclei were washed with 1 ml RSB and filtered through a 40 µm filter. Cells were counted, then

centrifuged again and resuspended in the proper volume of RSB to obtain 2000 nuclei/µl. The 10X library preparation, sequencing, and data processing were done using the same protocol as for the adipose tissue.

For the mouse brain data, we downloaded the raw UMI count data matrix from the 10X website. The data set titled "2K Brain Nuclei from an Adult Mouse (> 8 weeks)" was downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/nuclei_2k. The 10X human 68K PBMC data were downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/fresh_68k_pbmc_donor_a.

Filtering droplets using a quantile threshold, EmptyDrops, and DIEM. Common methods for removing debris from snRNA-seq data rely on using a hard count threshold[3,8–11]. In the three data sets, we applied a quantile-based cutoff, similar to that implemented by the 10X CellRanger software. Droplets are ranked in decreasing order of total counts. The 99th percent quantile of the top C barcodes of total counts is divided by 10 to obtain the threshold T, where C is 3,000 for our analyses[16]. The 99th percentile is used to exclude any doublets from the derivation. Droplets with greater than or equal to T counts were included as nuclei. For comparison with EmptyDrops[12], we ran the method using default parameters. EmptyDrops calculates a Monte Carlo p-value that gives the probability that a droplet's expression profile is the same as that of the ambient RNA. We removed droplets with a false discovery rate (FDR) q value greater than 0.05. For the six adipose tissue samples, we applied filtering to each sample independently, as these are the result of individual experiments. We also tested DIEM filtering after combining the counts in the 6 adipose tissue samples and observed similar results (Fig. S14).

Estimating extranuclear RNA contamination in droplets from snRNA-seq. We used three metrics to estimate contamination of background RNA in the snRNA-seq data sets. We quantified the fraction of spliced reads using velocyto[17]. The BAM files from CellRanger were sorted by barcode ID using samtools[24], and spliced, unspliced, and ambiguous read counts were quantified for each gene. We then removed mitochondria (MT) reads to avoid confounding of the estimates, as the MT genes do not have introns. For each droplet, the unspliced and spliced UMI counts were added, and the percent of reads spliced was calculated as the fraction of all spliced reads over the sum of spliced and unspliced reads. We calculated the percent of UMIs aligned to the mitochondria (MT%) as the sum of reads aligned to the mitochondrial genome over the droplet's total UMI counts. The percent of UMIs aligned to *MALAT1* (MALAT1%) was calculated similarly.

We classified droplets as background or nuclear according to their percent of reads spliced. Since this metric showed a bimodal distribution that was distinct in each sample, we calculated the midpoint between the two distributions. To do so, we modeled the percent of reads spliced as a mixture of two gaussians. We fit the parameters using EM with the R package mixtools[25]. Then, the midpoint was calculated as the value in which the probability density was equal in the two distributions. Droplets with a percent of reads spliced above and below the midpoint were classified as background and nuclear, respectively. For clusters, we specified those with an average percent of reads spliced of at least 50% as debris and classified those with less than 50% as nuclear, as we observed that 50% was the average value of the midpoint across the experiments and that the 6 adipose tissue samples were combined.

Differential expression between nuclear-enriched and debris-enriched droplets. To identify genes differentially expressed (DE) between the background-enriched and nuclear-enriched groups, we set a hard count threshold to naively assign droplets to either group. Droplets with total UMI counts below 100 and greater than or equal to 100 were assigned to the background-enriched and nuclear-enriched groups, respectively. This ensures that the majority of droplets containing nuclei are found in the nuclear-enriched group. For each gene, reads were summed across all droplets in each of the two groups to estimate the RNA profiles. Read counts were normalized using trimmed mean of M-values (TMM) as implemented in edgeR[26,27]. For identifying differentially expressed genes, we used a paired design with the six adipose tissue samples by treating the background-enriched and nuclear-enriched counts of an individual as a paired sample (total n = 12). We then used the edgeR package[26,27] to run differential expression. We only kept genes with a counts per million (CPM) of greater than 0 in at least 6 of the 12 groups. Next, we used the estimateDisp function to estimate the dispersion with the paired design matrix. The quasi-likelihood fit and F test functions glmQLFit and glmQLFTest were used to calculate statistical significance. We adjusted for multiple testing using a Bonferroni-corrected p-value threshold of 0.05.

To identify DE genes between the debris and cell types, we used the clusters identified after quantile-based filtering to approximate the cell types. For each of the six samples, we subsampled the debris droplets (with total UMI counts less than 100) to 9,000 droplets to obtain a similar read depth as contained in the cell type groups. For the debris and cell type groups, reads were summed across the corresponding droplets to obtain the RNA profile used as input. Differential expression was performed by comparing debris vs. cell type or cell type vs. all other cell types using a paired design. The filtering and analysis was performed in the same manner as the debris vs. nuclear DE analysis above.

DIEM algorithm. DIEM first assigns droplets as originating from debris or cell types, and then calculates the level of contamination within droplets using the debris-enriched genes. To assign droplets to either debris or cell types, our filtering approach models droplet-based single-cell or single-nucleus data with a mixture of multinomial distributions. Particularly, droplet read counts are assumed to follow a multinomial with parameters conditional on the cluster. However, the parameters and droplet assignments are unknown for the droplets of interest. In addition, it is assumed that the majority of low count droplets contain ambient RNA. Therefore, we estimate the parameters of the model using semi-supervised expectation maximization (EM)[13,14]. This allows us to calculate the probability of the latent group variable given the data, and thus group debris and cell type drop-

lets. To initialize the parameters for EM, we cluster the droplets with k-means, where the number of cell types k is specified by the user. We include only droplets with at least 200 genes detected in this initialization step to avoid fitting clusters driven by empty debris droplets. After fitting the mixture model with EM, we calculate a debris score for each droplet based on the expression of genes enriched in the debris set.

In more detail, let X denote a g x N matrix containing the read/UMI counts from a single-cell or single-nucleus data set with g genes and N droplets. We include droplets with at least 1 read/UMI count. Our goal is to assign the N droplets into one of $K+1$ groups (K cell types and debris). We define x_i as the ith column of X giving the counts of droplet i and assume that it follows a multinomial distribution with the gene probabilities $\boldsymbol{\alpha_k} = p_{1,k}, \ldots, p_{G,k}$ conditional on group $k \in \{1, ..., K, K+1\}$. We model droplet expression using a multinomial distribution, and further model cell types and debris using a mixture of multinomials. The log-likelihood of the data is therefore:

$$logP(\boldsymbol{X}) = \sum_{i=1}^{N} log\left(\sum_{k=1}^{K+1} \pi_k Mult(\boldsymbol{x_i}|\boldsymbol{\alpha_k}, u_i)\right)$$

Here, u_i is the total number of read/UMI counts in droplet i, $\boldsymbol{\alpha_k}$ contains the multinomial parameters for group k, π_k is the mixing coefficient for group k, and *Mult* denotes the probability mass function of the multinomial distribution. Since an analytical solution cannot be derived based on the likelihood of the model, we define $z_i \in \{1, ..., K\}$ for each i as a latent indicator variable that describes the cell type or debris origin of the droplet. The complete log-likelihood of the data and the latent variables $Z = \{z_i\}_{i=1}^{N}$ now becomes:

$$logP(\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{N} \sum_{k=1}^{K+1} \mathbb{I}\{\mathrm{z_i} = \mathrm{k}\}[\log\pi_{\mathrm{k}} + \mathrm{logMult}(\boldsymbol{x_i}|\boldsymbol{\alpha_k}, \mathrm{u_i})]$$

where $\mathbb{I}\{\mathrm{z_i} = \mathrm{k}\}$ is an indicator variable for the assignment of droplet i. This formulation allows us to employ an EM algorithm and estimate the parameters $\alpha_1, \ldots, \alpha_k$ and $\pi_1, \ldots, \pi_k$ by maximizing the expected complete data log-likelihood. The latent indicator variables for the debris droplets with UMI counts below 100 remain fixed, thus effectively resulting in a semi-supervised EM.

Although it is possible to remove droplets that contain a high posterior probability of belonging to the fixed debris cluster, we employ a scoring strategy to quantify the level of contamination within individual droplets. This provides both a finer resolution in debris filtering and a direct estimate that can be used as a covariate in downstream analysis.

Droplets are divided into a test set and a debris set. The test set consists of the droplets we would like to classify, while the debris set consists of droplets that we assume to contain debris with high probability. The labels of droplets in the debris set are fixed throughout the EM iterations, while those of the test set are allowed to change. We define the test set as those droplets with at least T total counts, where we set T to a default value of 100. Only expressed genes with a counts per million (CPM) > 0 are included in the analysis.

Initialization of parameters for EM. The EM algorithm requires starting values for the parameters of the model. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ are initialized from the PCs of the cluster set of droplets using k-means. A proper initialization is important because mixture models can be sensitive to local optima[28]. Therefore, we run k-means, which has been shown to provide reasonable initial values for EM[29,30]. As the test set may contain many more empty droplets than the droplets of interest, we further define a cluster set for k-means as those droplets with at least 200 genes detected [20]. K-means is run on the on the first 30 PCs of the data using the kmeans function in R. Before running PCA, we first select the top $V = 2{,}000$ variable genes[20,31]. To do so, we first account for the relationship between the mean and variance[31,32]. The mean and variance of the raw gene counts are calculated and log transformed. To learn the relationship, we fit a locally weighted smoothing (LOESS) regression line between the normalized mean and variance using the loess function in R with a span = 0.3. We correct the variance for the expression level of a gene by subtracting the fitted variance from the observed variance. Finally, we rank the genes by their standardized variance and take the top $V = 2{,}000$ genes. PCA is run on the normalized counts, where the total droplet read counts are scaled to sum to the median read depth and then log transformed. PCA is performed on the adjusted counts on the cluster set and the variable genes, and the top 30 principal components are returned. Finally, k-means is run on these PCs, with the number of cell types k specified by the user. We use k = 20 for all experiments in the manuscript, unless otherwise specified. The initial parameters are estimated from the droplets assigned to these resulting clusters.

Estimation. The EM algorithm iteratively estimates the parameters and the posterior probabilities. Given $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\pi}}$, estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$, we calculate the posterior probability that droplet X_i belongs to cluster k

$$p(z_i = k|\boldsymbol{x_i}, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{k}}, \widehat{\boldsymbol{\pi}}) = \frac{p(\widehat{\boldsymbol{\pi}}_k)p(\boldsymbol{x_i}|z_i = k, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{k}})}{\sum_{j=1}^{K} p(\widehat{\boldsymbol{\pi}}_j)p(\boldsymbol{x_i}|z_i = j, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{j}})}$$

where $p(X_i|Z_i, \alpha_k)$ follows the multinomial given the parameters $\boldsymbol{\alpha_k}$ for cluster k and $p(\pi_k)$ follows a categorical distribution. The debris droplets with total counts below $T = 100$ have their z_i values kept fixed to the debris group. The maximum likelihood estimate of $\boldsymbol{\alpha_k}$ is calculated as the mean of the droplet counts weighted by their posterior probability of belonging to cluster k. We add a pseudocount of 10^{-10} to avoid collapsing the likelihood to 0. For π_k, the maximum likelihood estimate is calculated as the sum of $p(\boldsymbol{Z} = k|\boldsymbol{X})$ divided by the total number of droplets, so that $\pi_1, \ldots, \pi_K$ sum to one. These two steps iterate during EM, and the algorithm converges when the change in parameters is below ε, which we set to 10^{-4}. Droplets are assigned to the cluster that gives the maximum posterior probability.

Debris scoring and filtering of individual droplets. We assign a debris score to individual droplets to obtain a finer estimate of the amount of contamination. After clustering, we specify the set of debris clusters as the fixed cluster as well as any clusters that have an average number of genes detected less than d, where we set d to 200. The cell type group then consists of all other clusters. We estimate the debris score by summing a droplet's expression values of genes enriched in this debris set. To identify debris-enriched genes, we first run a Welch's t-test between the test set droplets in the debris and cell type clusters. Read counts are normalized by scaling the counts to sum to 1 and then log normalizing after adding a constant of 1. Then, genes with a log fold change greater than 0 and an FDR-corrected p-value less than 0.05 are specified as debris-enriched genes.

The debris score is estimated by summing the normalized expression values of the debris-enriched genes. Since the magnitude of this score is dependent on the number and expression of the debris-enriched genes, we scale the scores. We calculate the mean of all clusters, subtract the scores by the lowest cluster average, and divide them by average of the droplets in the debris cluster(s). This has the effect of setting the average of the lowest cluster to 0 and the debris cluster(s) to 1, so that scores are scaled relative to these clusters. Droplets in the snRNA-seq experiments are filtered using a threshold for the debris score. We keep droplets with a normalized debris score below t, where we set t to 0.5, although we note that these can be adjusted by the user accordingly.

Identifying cell types after filtering droplets. For all experiments, we ran a standardized clustering pipeline using Seurat v3.1.2[20]. After applying filtering, we only considered droplets with at least 200 genes detected [4] to ensure that each droplet had enough information for clustering. The count data were log-normalized using the NormalizeData function in Seurat, using a scaling factor equal to the median of total counts across droplets. For the six adipose tissue samples, we used a scaling factor equal to 1,000 to ensure that all samples were normalized equally. Additionally, we merged the normalized data of the six adipose tissue samples without batch correction, as we saw high overlap of clusters among the six samples (data not shown). The top 2,000 variable genes were then calculated using the FindVariableFeatures function.

Normalized read counts for each gene were scaled to mean 0 and variance 1. We calculated the first 30 PCs to use as input for clustering. We then ran the Seurat functions FindNeighbors and FindClusters with 30 PCs. In the FindClusters function, we used the default parameters with standard Louvain clustering and a default clustering resolution of 0.8, unless otherwise stated. For visualization, we ran UMAP[33] on the 30 PCs with default values. To identify marker genes for each cluster, we ran a Wilcoxon rank-sum test using the function FindAllMarkers with default parameters and only.pos = TRUE. We corrected for multiple testing using a false discovery rate (FDR) threshold of 0.05. Clusters were classified as doublets if the top marker genes consisted of an identifiable mixture of top markers between two cell types.

Ethics approval and consent to participate. All research was performed in accordance with the relevant institutional guidelines and regulations. Each of the 6 participants gave a written informed consent. The study protocol was approved by the Ethics Committee at the Helsinki University Hospital, Helsinki, Finland.

Data availability

The human single nucleus RNA-seq datasets generated and analyzed during the current study are available upon request from the corresponding author. The DIEM program is freely available for use at https://github.com/marcalva/diem. The code for the analysis is available at https://github.com/marcalva/DIEM2019.

Received: 28 October 2019; Accepted: 4 June 2020
Published online: 03 July 2020

References

1. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
2. 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**, e4.346-e4.360 (2016).
3. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
5. Habib, N. *et al.* Div-Seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
6. Krishnaswami, S. R. *et al.* Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499 (2016).
7. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental considerations for single-cell RNA sequencing approaches. *Front. Cell Dev. Biol.* **6**, 108 (2018).
8. Hu, P. *et al.* Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-Seq. *Mol. Cell* **68**, 1006-1015.e7 (2017).
9. Lacar, B. *et al.* Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 1–13 (2016).
10. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
11. Zeng, W. *et al.* Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Res.* **44**, e158 (2016).
12. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
13. Dempster, A. P. P., Laird, N. M., Rubin, D. B. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–22 (1977).
14. Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm?. *Nat. Biotechnol.* **26**, 897–899 (2008).
15. Nigam, K., Mccallum, A. K., Thrun, S. & Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* **39**, 103–134 (2000).

16. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).
17. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
18. Miyagawa, R. *et al.* Identification of cis- and trans-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. *RNA* **18**, 738–741 (2012).
19. Hardison, R. C. Evolution of hemoglobin and its genes. *Cold Spring Harbor Perspect. Med.* **2**, a011627–a011627 (2012).
20. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902.e21 (2019).
21. Ziegler-Heitbrock, H. W. L. & Ulevitch, R. J. CD14: cell surface receptor and differentiation marker. *Immunol. Today* **14**, 121–125 (1993).
22. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).
23. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
24. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. Mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
26. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
27. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
28. Biernacki, C., Celeux, G. & Govaert, G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.* **41**, 561–575 (2003).
29. Steinley, D. & Brusco, M. J. Evaluating mixture modeling for clustering: recommendations and cautions. *Psychol. Methods* **16**, 63–79 (2011).
30. McLachlan, G. J., Lee, S. X. & Rathnayake, S. I. Finite mixture models. *Annu. Rev. Stat. Appl.* **6**, 355–378 (2019).
31. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
32. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 1–5 (2019).
33. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38 (2019).

Acknowledgements

We thank the Finnish individuals who participated in the adipose tissue study, as well as Jaakko Kaprio and Aila Rissanen for their contributions. We also thank the Finnish Institute for Molecular Medicine Single Cell Analytics Core for performing library preparation and sequencing for the adipose tissue samples. This study was funded by the National Institutes of Health (NIH) grants HL-095056, HL-28481, and U01 DK105561. M.A. was supported by the HHMI Gilliam Fellowship and the NIH T32HG002536. JRP, JNB and PP were supported by an NIH DK41301 grant. E.H., B.J., and E.R. were partially supported by the National Science Foundation (Grant No. 1705197). E.H., E.R., and B.J. were partially supported by NIH/NHGRI HG010505-02. E.H. was also partially funded by NIH 1R56MD013312, NIH 1R01MH115979, NIH 5R25GM112625, and NIH 5UL1TR001881. B.J. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650604. K.M.G. was supported by the NIH F31HL142180. Z.M was supported by the AHA grant 19PRE34430112. K.H.P. was supported by the Academy of Finland (272376, 266286, 314383, 315035), Finnish Medical Foundation, Finnish Diabetes Research Foundation, Novo Nordisk Foundation, Gyllenberg Foundation, Sigrid Juselius Foundation, Helsinki University Hospital Research Funds, Government Research Funds and University of Helsinki. J.N.B. was supported by the DDRC Pilot and Feasibility of the National Institutes of Health under award number DKP3041301 and the National Center for Advancing Translational Sciences at UCLA, CTSI Grant ULTR001881.

Author contributions

M.A., E.R., E.H., and P.P. conceived the study and designed the analysis. M.A., K.M.G., and J.B. designed, performed, and interpreted nuclei isolation experiments. M.A. and K.M.G. designed and performed snRNA-seq experiments and collected data. M.A., E.R. B.J., Z.M., K.M.G., and J.B. analyzed and interpreted snRNA-seq data. M.A., E.R., E.H. and P.P designed the DIEM method. J.R.P., C.J.Y., K.H.P. and P.P. designed and supervised nuclei isolation and snRNA-seq experiments. M.A., E.R., B.J., Z.M., E.H., and P.P. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-67513-5.

Correspondence and requests for materials should be addressed to P.P.

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

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

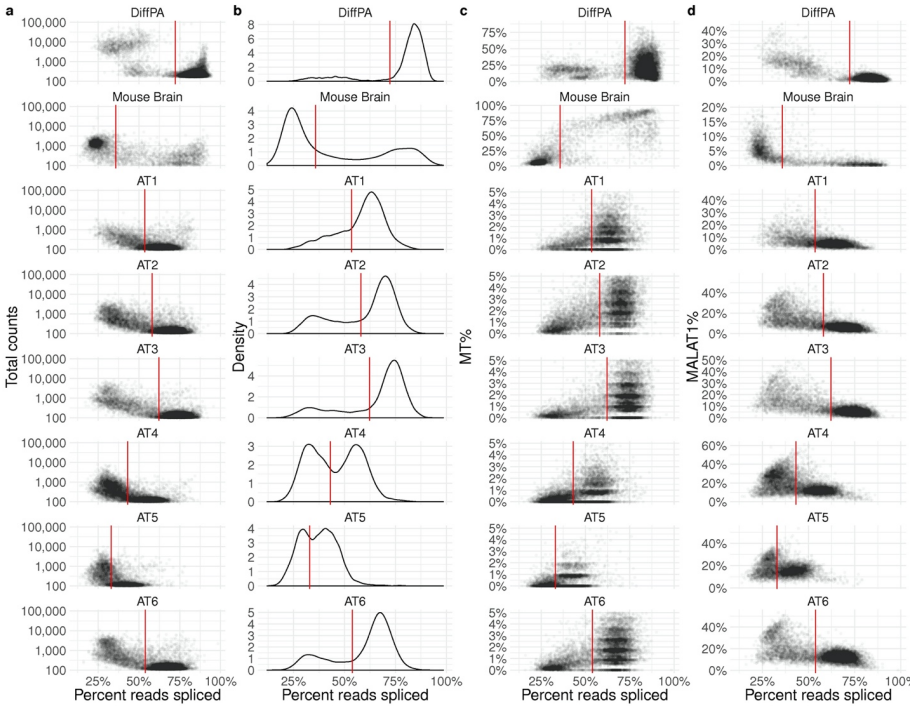


Figure S1. The percent of reads spliced separates nuclear from background RNA in snRNA-seq droplets.

The percent of reads spliced per droplet was calculated for each of the 8 independent samples in order to quantify extranuclear RNA contamination. To assess the effectiveness of this metric, we plotted the percent reads spliced against **(a)** total counts, **(b)** the density, **(c)** the percent of reads aligning to the mitochondria (MT%), and **(d)** the percent of reads aligning to *MALAT1* (MALAT1%). The human adipose tissue (AT) dataset was performed over 6 independent experiments. The spliced reads percent was calculated using Velocyto[17] after removing mitochondrial reads. As each sample demonstrated a distinct distribution of spliced reads, we estimated a cutoff (see methods) for each sample (vertical red line). Droplets with a percent of

reads spliced below the cutoff were classified as nuclear, and those greater than or equal to the cutoff as background.

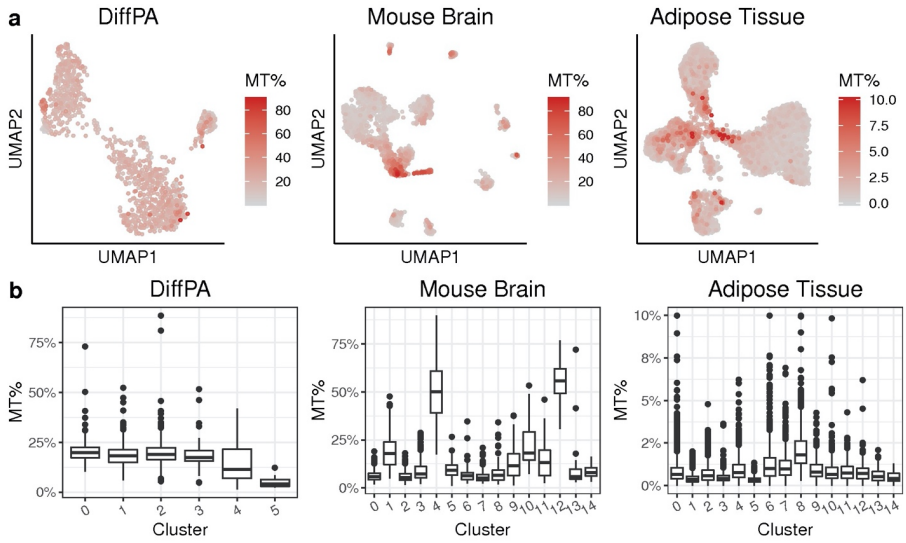


Figure S2. A hard count threshold fails to remove contaminated droplets and results in spurious clusters when assessed using MT%.

The estimation of background RNA when estimated using mitochondrial percent (MT%) in a droplet shows how a hard count threshold fails to remove contaminated droplets **a,** UMAP[33] visualizations for the differentiating preadipocytes (DiffPA), mouse brain, and human frozen adipose tissue (AT) data sets show clustering of contaminated droplets. **b,** boxplots of MT% in clusters after processing the filtered droplets with Seurat[20]. The quantile-based approach was used to select the hard count threshold.

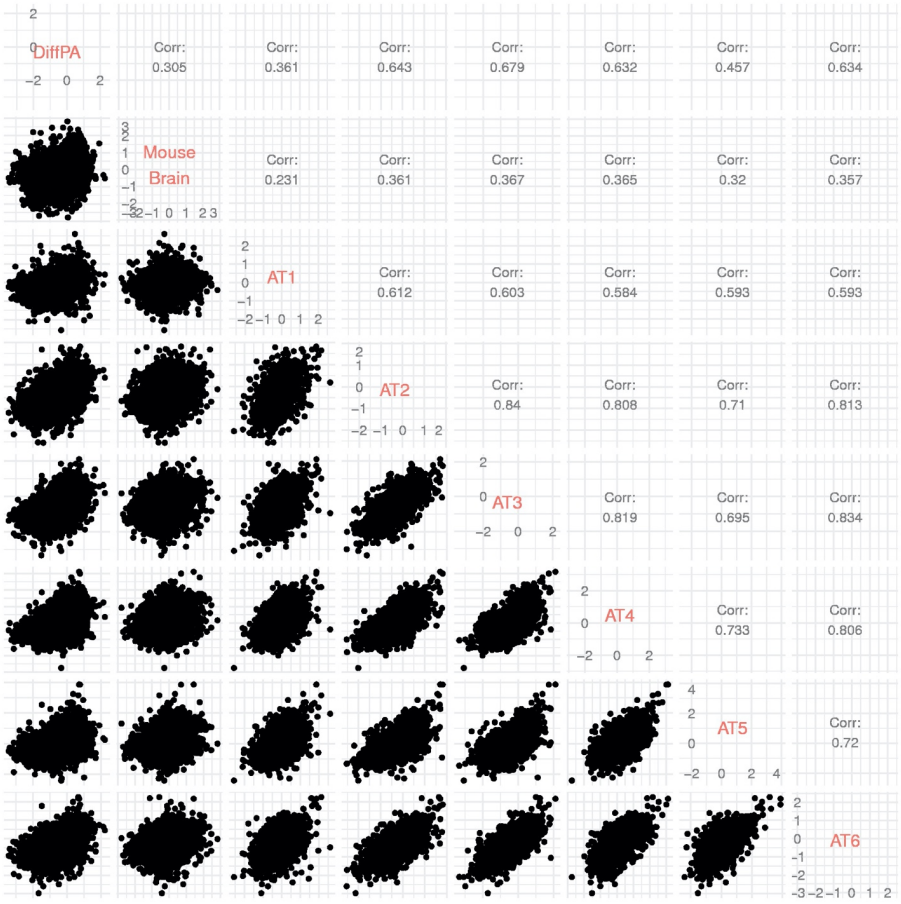


Figure S3. Preservation of differential RNA profiles between nuclear-enriched and background-enriched droplets.

Correlation plots of log fold changes across different snRNA-seq experiments. For each of the 8 experiments (differentiating preadipocytes (DiffPA), mouse brain, and six human frozen adipose tissue (AT) snRNA-seq samples), the log2 fold change of the counts per million (CPM) for each gene is calculated between the nuclear-enriched and background-enriched droplets. Nuclear-

enriched and background-enriched droplets are those with UMI counts greater than or equal to, and less than 100 UMI counts, respectively.

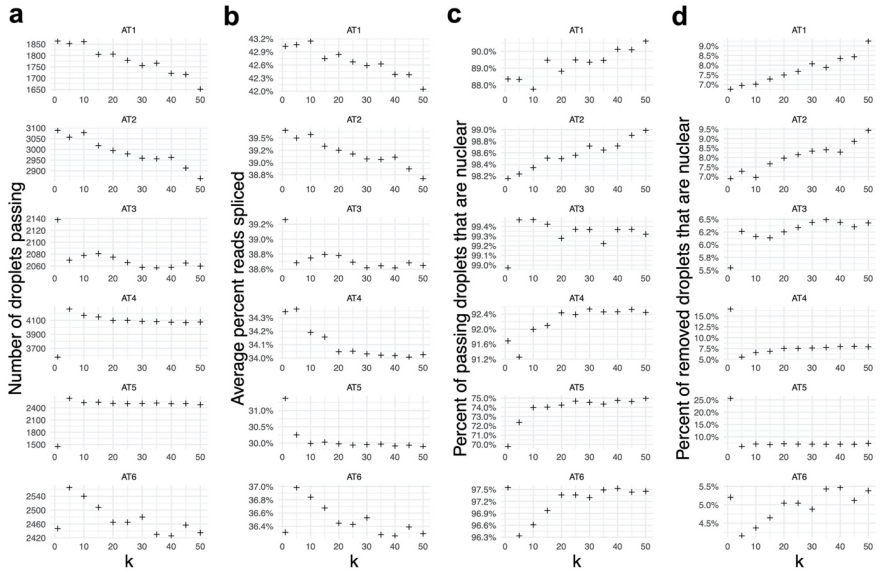


Figure S4. Effect of the number of clusters on classification accuracy in a mixture model. The number of clusters k was varied across values of 1, 5, …, 50 in the six adipose tissue samples. DIEM was run for each indicated k using a threshold value *t* of 0.5. The **(a)** number of droplets passing filtering and with a number of genes detected of at least 200, **(b)** the average percent of reads spliced, **(c)** the percent of passing droplets that are nuclear, and **(d)** the percent of removed droplets that are nuclear are shown. Nuclear droplets are defined as those with a percent of spliced reads below the sample-specific midpoint. Background and nuclear droplets are defined using the percent spliced reads.

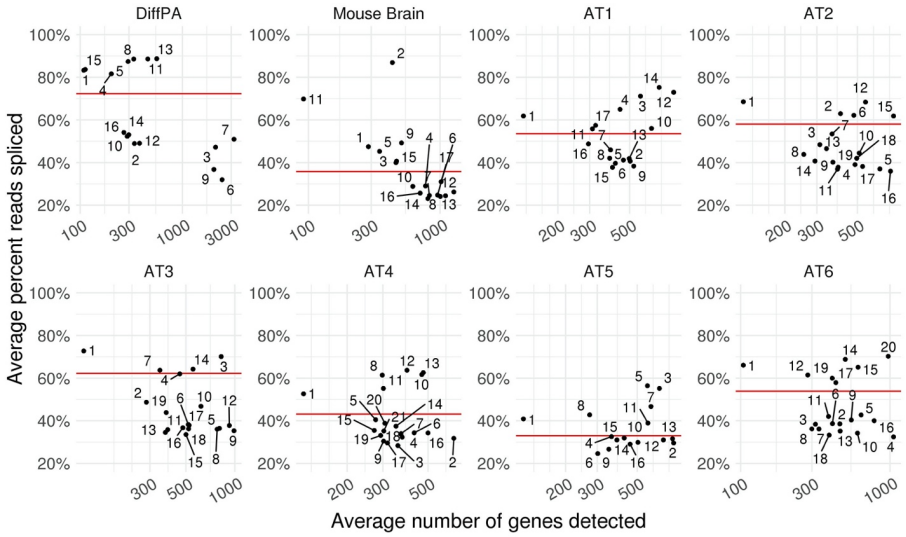


Figure S5. Single-nucleus RNA-seq produces clusters with high levels of background contamination.

The clusters produced by DIEM for the differentiating preadipocytes (DiffPA), mouse brain, and human frozen adipose tissue (AT) data sets are shown. Cluster 1 corresponds to the fixed debris cluster. The average number of genes detected in a cluster is plotted against the average percent of reads spliced. As each sample demonstrated a distinct distribution of spliced reads, we estimated a cutoff that separates nuclear and background droplets (see methods) for each sample (horizontal line). Clusters above and below the line indicate the background and nuclear clusters, respectively. This shows that clusters with high numbers of genes are susceptible to contamination.

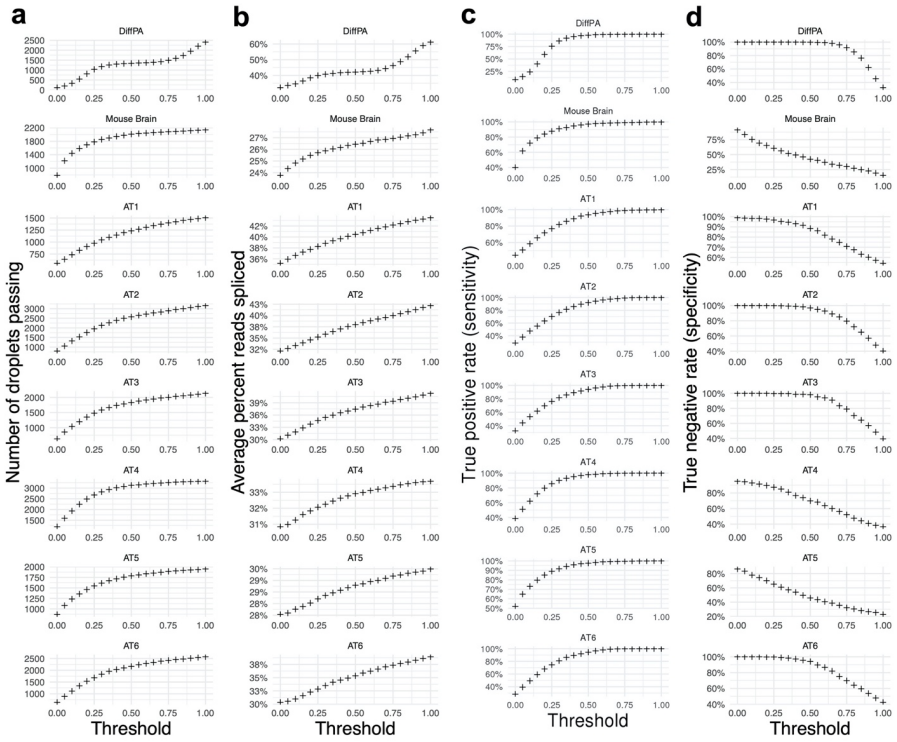


Figure S6. Increasing the threshold parameter *t* increases sensitivity and decreases specificity.

The figure shows the effect of varying the threshold parameter *t* from 0 to 1. DIEM was run on the differentiating preadipocytes (DiffPA), mouse brain, and human frozen adipose tissue (AT) data sets using k=20 clusters. **a,** The number of droplets that pass filtering and with a number of genes detected of at least 200. **b,** The average percent of reads spliced in droplets that pass DIEM filtering. **c,d,** The true positive and true negative rates are calculated for droplets with at least 200 genes detected. Nuclear and background droplets are defined as those with a percent of spliced reads below and above the sample-specific midpoint. **c,** The true positive rate (sensitivity), calculated as the percent of all nuclear droplets that correctly pass filtering, is

plotted against the threshold value. **d,** The true negative rate (specificity), calculated as the percent of all background droplets that are correctly removed, is plotted against the threshold value.

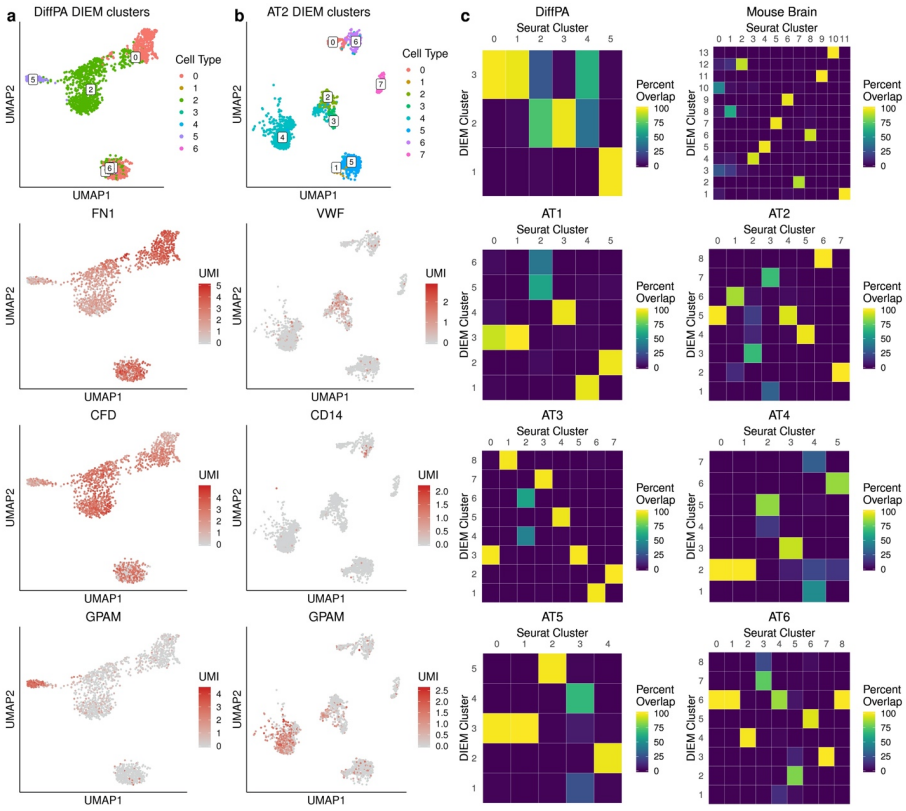


Figure S7. Accurate modeling of major cell types by the multinomial mixture model in DIEM.

a,b, UMAP[33] visualizations of clustering results after applying DIEM filtering for the **(a)** differentiating preadipocytes (DiffPA), and **(b)** adipose tissue sample 2 (AT2). The top panel shows the clusters identified by Seurat[20], while the bottom clusters show cell type marker expression in these clusters. The DiffPA data set consists of preadipocytes (expressing *CFD*), fibroblasts (expressing *FN1*), and adipocytes (expressing *GPAM*), while the AT consists of adipocyte (expressing *GPAM*), immune (expressing *CD14*), endothelial (expressing *VWF*), and

stromal cell types. **c,** Overlap of clusters identified by the DIEM mixture model with those from Seurat. Each panel shows the results from one of the eight independent data sets. The rows of the heatmap correspond to clusters identified by DIEM, while the columns correspond to Seurat clusters. Brighter values indicate a higher overlap. The percent overlap is defined as the number of shared droplets divided by the minimum size of the clusters in the pair. The average overlap was 73.0% across corresponding clusters for the DIEM clusters.

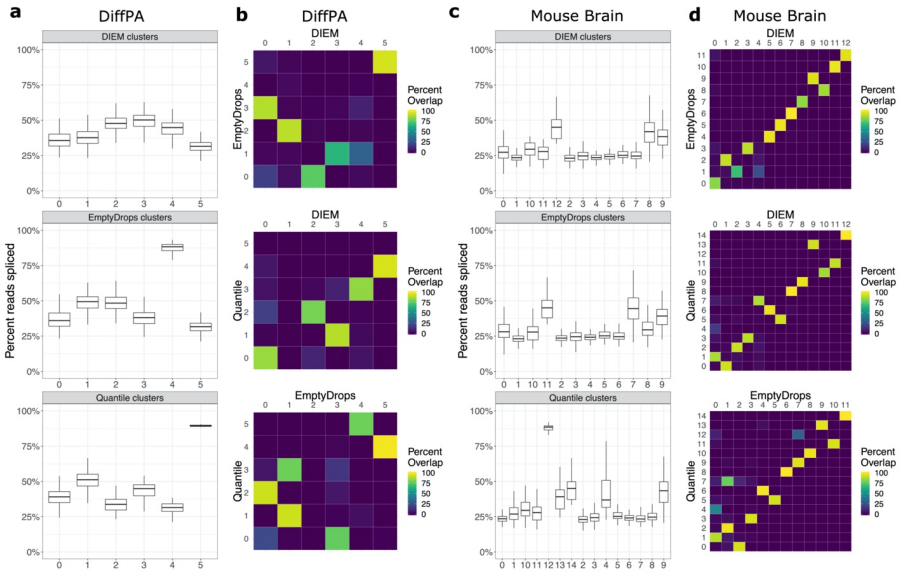


Figure S8. DIEM filtering reduces contamination in clusters in the differentiating pre-adipocyte and mouse brain single-nucleus RNA-seq experiments.

a,b, The **(a)** distribution of the percent of reads spliced for droplets in clusters identified by Seurat[20] after filtering with each of the three methods in the differentiating preadipocytes (DiffPA) is shown in a box plot. The **(b)** overlap of the resulting DiffPA clusters between the three filtering methods is shown in a heatmap. **c,d,** The **(c)** distribution of the percent of reads spliced for droplets in Seurat clustering after filtering with each of the three methods in the mouse brain is shown in a box plot. The **(d)** overlap of the resulting mouse brain clusters between the three filtering methods is shown in a heatmap. Brighter values in the heatmap indicate a higher percent overlap between the methods. Major cell types are preserved across the filtering methods.

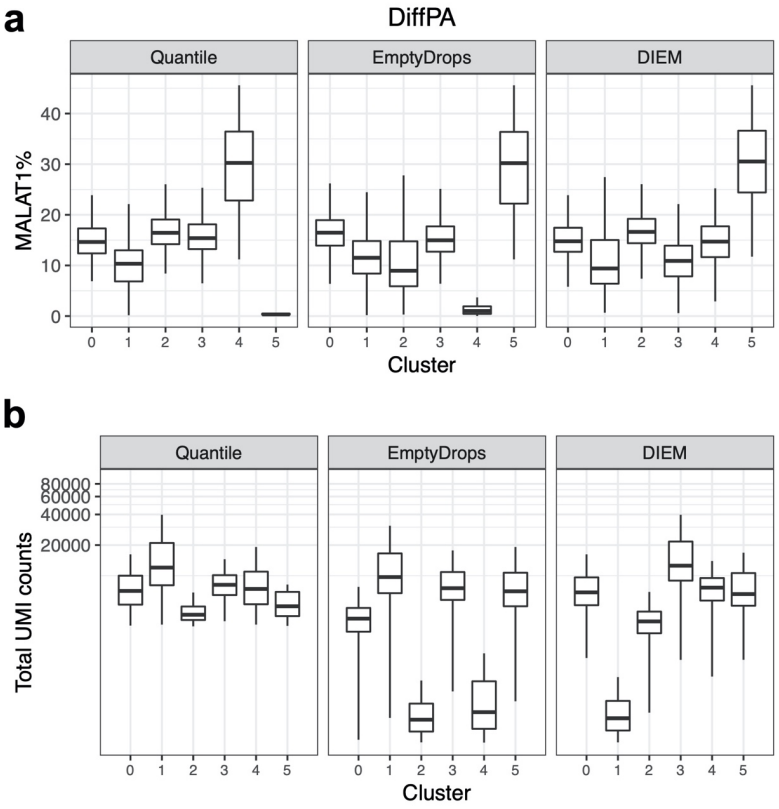


Figure S9. DIEM removes clusters with high MALAT1 expression and is able to keep clusters with low read counts in the DiffPA snRNA-seq data set.

a, Boxplots showing the percent of UMIs mapping to *MALAT1* (MALAT1%) per droplet in the differentiating preadipocytes (DiffPA). MALAT1% of clusters are compared across the quantile-based, EmptyDrops[12], and DIEM filtering methods. *MALAT1* is a nuclear-localized lincRNA[18], which suggests that the RNA is of nuclear origin. **b,** Boxplots showing the total number of UMIs

per droplet in the differentiating preadipocytes (DiffPA). Clusters are compared across the three filtering methods.

CHAPTER 3

Accurate estimation of cell composition in bulk expression through robust integration of single-cell information

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-15816-6 OPEN

Accurate estimation of cell composition in bulk expression through robust integration of single-cell information

Brandon Jew[1,10], Marcus Alvarez[2,10], Elior Rahmani[3], Zong Miao[1,2], Arthur Ko[2], Kristina M. Garske[2], Jae Hoon Sul[1,4], Kirsi H. Pietiläinen[5,6], Päivi Pajukanta[1,2,7✉] & Eran Halperin[2,3,7,8,9✉]

We present Bisque, a tool for estimating cell type proportions in bulk expression. Bisque implements a regression-based approach that utilizes single-cell RNA-seq (scRNA-seq) or single-nucleus RNA-seq (snRNA-seq) data to generate a reference expression profile and learn gene-specific bulk expression transformations to robustly decompose RNA-seq data. These transformations significantly improve decomposition performance compared to existing methods when there is significant technical variation in the generation of the reference profile and observed bulk expression. Importantly, compared to existing methods, our approach is extremely efficient, making it suitable for the analysis of large genomic datasets that are becoming ubiquitous. When applied to subcutaneous adipose and dorsolateral prefrontal cortex expression datasets with both bulk RNA-seq and snRNA-seq data, Bisque replicates previously reported associations between cell type proportions and measured phenotypes across abundant and rare cell types. We further propose an additional mode of operation that merely requires a set of known marker genes.

[1] Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA 90095, USA. [2] Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA. [3] Department of Computer Science, School of Engineering, UCLA, Los Angeles, CA 90095, USA. [4] Department of Psychiatry and Biobehavioral Sciences, UCLA, Los Angeles, CA 90095, USA. [5] Obesity Research Unit, Research Program for Clinical and Molecular Metabolism, University of Helsinki, Helsinki 00014, Finland. [6] Obesity Center, Endocrinology Abdominal Center, Helsinki University Central Hospital and University of Helsinki, Helsinki 00260, Finland. [7] Institute for Precision Health, School of Medicine, UCLA, Los Angeles, CA 90095, USA. [8] Department of Anesthesiology, UCLA Health, Los Angeles, CA 90095, USA. [9] Department of Computational Medicine, School of Medicine, UCLA, Los Angeles, CA 90095, USA. [10] These authors contributed equally: Brandon Jew, Marcus Alvarez. ✉email: ppajukanta@mednet.ucla.edu; ehalperin@cs.ucla.edu

Bulk RNA-seq experiments typically measure total gene expression from heterogeneous tissues, such as tumor and blood samples[1,2]. Variability in cell-type composition can significantly confound analyses of these data, such as in identification of expression quantitative trait loci (eQTLs) or differentially expressed genes[3]. Cell-type heterogeneity may also be of interest in profiling changes in tissue composition associated with disease, such as cancer[4] or diabetes[5]. In addition, measures of cell composition can be leveraged to identify cell-specific eQTLs[6,7] or differential expression[6] from bulk data.

Traditional methods for determining cell-type composition, such as immunohistochemistry or flow cytometry, rely on a limited set of molecular markers and lack in scalability relative to the current rate of data generation[8]. Single-cell technologies provide a high-resolution view into cellular heterogeneity and cell-type-specific expression[9–11]. However, these experiments remain costly and noisy compared to bulk RNA-seq[12]. Collection of bulk expression data remains an attractive approach for identifying population-level associations, such as differential expression regardless of cell-type specificity. Moreover, many bulk RNA-seq studies that have been performed in recent years resulted in a large body of data that is available public databases such as dbGAP and GEO. Given the wide availability of these bulk data, the estimation of cell-type proportions, often termed decomposition, can be used to extract large-scale cell-type-specific information.

There exist a number of methods for decomposing bulk expression, many of which are regression-based and leverage cell-type-specific expression data as a reference profile[13]. CIBERSORT[14] is a SVM-regression-based approach, originally designed for microarray data that utilizes a reference generated from purified cell populations. A major limitation of this approach is the reliance on sorting cells to estimate a reference gene expression panel. BSEQ-sc[15] instead generates a reference profile from single-cell expression data that is used in the CIBERSORT model. MuSiC[16] also leverages single-cell expression as a reference, instead using a weighted non-negative least-squares regression (NNLS) model for decomposition, with improved performance over BSEQ-sc in several datasets.

The distinct nature of the technologies used to generate bulk and single-cell sequencing data may present an issue for decomposition models that assume a direct proportional relationship between the single-cell-based reference and observed bulk mixture. For example, the capture of mRNA and chemistry of library preparation can differ significantly between bulk tissue and single-cell RNA-seq methods, as well as between different single-cell technologies[17,18]. Moreover, some technologies may be measuring different parts of the transcriptome, such as nuclear pre-mRNA in single-nucleus RNA-seq (snRNA-seq) experiments as opposed to cellular and extra-cellular mRNA observed in traditional bulk RNA-seq experiments. As we show later, these differences may introduce gene-specific biases that break down the correlation between cell-type-specific and bulk tissue measurements. Thus, while single-cell RNA-seq technologies have provided unprecedented resolution in identifying expression profiles of cell types in heterogeneous tissues, these profiles generally may not follow the direct proportionality assumptions of regression-based methods, as we demonstrate here.

We present Bisque, a highly efficient tool to measure cellular heterogeneity in bulk expression through robust integration of single-cell information, accounting for biases introduced in the single-cell sequencing protocols. The goal of Bisque is to integrate the different chemistries/technologies of single-cell and bulk tissue RNA-seq to estimate cell-type proportions from tissue-level gene expression measurements across a larger set of samples. Our reference-based model decomposes bulk samples using a single-cell-based reference profile and, while not required, can leverage single-cell and bulk measurements for the same samples for further improved decomposition accuracy. This approach employs gene-specific transformations of bulk expression to account for biases in sequencing technologies as described above. When a reference profile is not available, we propose BisqueMarker, a semi-supervised model that extracts trends in cellular composition from normalized bulk expression samples using only cell-specific marker genes that could be obtained using single-cell data. We demonstrate using simulated and real datasets from brain and adipose tissue that our method is significantly more accurate than existing methods. Furthermore, it is extremely efficient, requiring seconds in cases where other methods require hours; thus, it is scalable to large genomic datasets that are now becoming available.

Results

Method overview (Bisque). A graphical overview of Bisque is presented in Fig. 1. Our reference-based decomposition model requires bulk RNA-seq counts data and a reference dataset with read counts from single-cell RNA-seq. In addition, the single-cell data should be labeled with cell types to be quantified. A reference profile is generated by averaging read count abundances within each cell type in the single-cell data. Given the reference profile and cell proportions observed in the single-cell data, our method learns gene-specific transformations of the bulk data to account for technical biases between the sequencing technologies. Bisque can then estimate cell proportions from the bulk RNA-seq data using the reference and the transformed bulk expression data using non-negative least-squares (NNLS) regression.

Evaluation of decomposition performance in adipose tissue. We applied our method to 106 bulk RNA-seq subcutaneous adipose tissue samples collected from both lean and obese individuals, where 6 samples have both bulk RNA-seq and snRNA-seq data available (Table 1). Each of the participants gave a written informed consent. The study protocol was approved by the Ethics Committee at the Helsinki University Hospital, Helsinki, Finland. Adipose tissue consists of several cell types, including adipocytes that are expected to be the most abundant population. Adipose tissue also contains structural cell types (i.e. fibroblasts and endothelial cells) and immune cells (i.e. macrophages and T cells)[19]. These 5 cell-type populations were identified from the snRNA-seq data (Supplementary Fig. 1a).

We observed significant biases between the snRNA-seq and bulk RNA-seq data in samples that had both data available. We found that the linear relationship between the pseudo-bulk (summed snRNA-seq reads across cells) and the true bulk expression varied significantly by each gene (Fig. 2a). Specifically, we observed best fit lines relating these expression levels between technologies with a mean slope of roughly 0.30 and a variance in slope of 5.67. In our model, a slope of 1 would indicate no bias between technologies. We further investigated whether gene expression differences between the bulk and snRNA-seq were the same across individuals and experiments. Comparing log-ratios of RNA-seq to snRNA-seq expression levels, we found that the majority of gene biases were preserved across individuals, tissues, and experiments ($R = 0.75$ across experiments) (Supplementary Fig. 3), providing evidence that technological differences drive consistent gene expression differences across bulk and snRNA-seq methods.

We performed simulations based on the adipose snRNA-seq data to demonstrate the effect of technology-based biases between the reference profile and bulk expression on decomposition performance. In these analyses, we benchmarked Bisque and

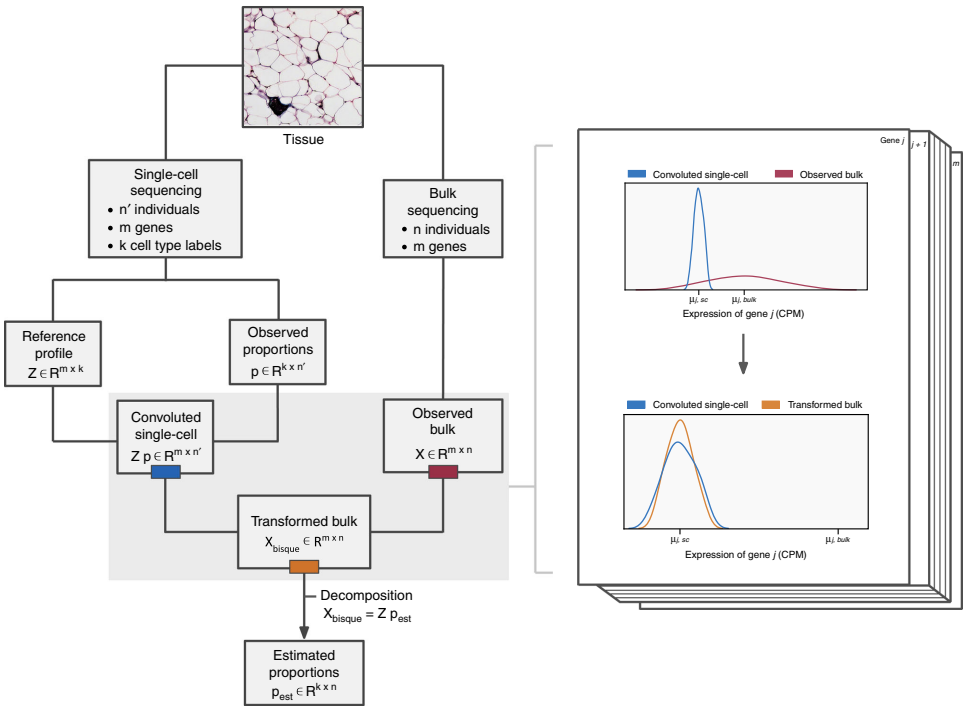


Fig. 1 Graphical overview of the Bisque decomposition method. We integrate single-cell and bulk expression by learning gene-specific bulk transformations (pictured on *right*) that align the two datasets for accurate decomposition.

Table 1 Summary of snRNA-seq and bulk expression datasets used for benchmarking Bisque and existing methods.

Tissue	Number of samples	Bulk RNA-seq platform	snRNA-seq platform	snRNA-seq samples	Total nuclei	Average nuclei per individual	Number of cell types
Subcutaneous adipose	106	Illumina NovaSeq	10x Genomics Chromium	6	10,947	1824	5
Dorsolateral prefrontal cortex	636	Illumina HiSeq	10x Genomics Chromium	8	68,028	8503	11

three existing methods (MuSiC, BSEQ-sc, and CIBERSORT). Briefly, we simulated bulk expression for 6 individuals by summing the observed snRNA-seq read counts. To model discordance between the reference and bulk, we applied gene-specific linear transformations of the simulated bulk expression. For each gene, the coefficient and intercept of the linear transformation were sampled from half-normal distributions with increasing variance. In this model, a higher variance corresponds to a larger bias between sequencing experiments. Although these transformations closely mirrored the Bisque decomposition model, they utilized the true snRNA-seq counts for each individual whereas Bisque learned these transformations using the reference profile generated from averaging these counts across all cells. Hence, this simulation framework introduced additional noise that Bisque does not entirely model. We evaluated decomposition performance by comparing proportion estimates to the proportions observed in the snRNA-seq data in terms of global Pearson correlation (R) and root-mean squared deviation (RMSD). Owing to the small number of samples, we applied leave-one-out cross-validation to predict the cell composition of each individual using the remaining snRNA-seq samples as training data for each method. In these simulations, Bisque remained robust ($R \approx 0.85$, RMSD ≈ 0.07) at higher levels of simulated bias between the bulk and snRNA-seq-based reference (Fig. 2b).

Next, we performed this cross-validation benchmark on the observed bulk RNA-seq data for these 6 individuals and found that Bisque ($R = 0.923$, RMSD $= 0.074$) provided significantly improved global accuracy in detecting each cell type over existing methods (Table 2, Supplementary Fig. 1b). MuSiC ($R = -0.111$, RMSD $= 0.427$), BSEQ-sc ($R = -0.113$, RMSD $= 0.432$), and CIBERSORT ($R = -0.131$, RMSD $= 0.416$) severely underestimated the proportion of adipocytes (the most abundant population in adipose tissue) while overestimating the endothelial cell fraction. We also benchmarked CIBERSORTx[20], which employs a batch correction mode to account for biases in

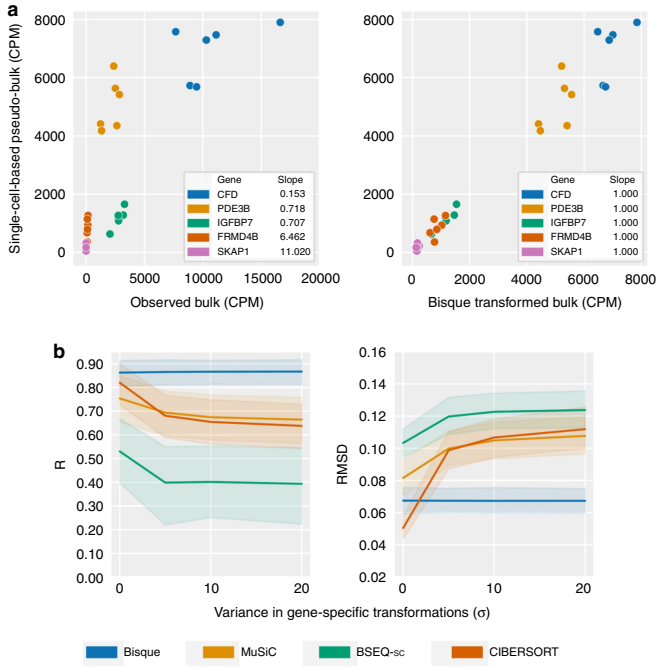


Fig. 2 The effect of discrepancies between a single-cell-based reference and bulk expression on decomposition. a Observed discrepancies in real data between single-nucleus and bulk expression for selected marker genes (*left*) for six individuals. Each color corresponds to a gene. On the left, observed bulk expression on the *x* axis is plotted against the pseudo-bulk expression on the *y* axis, where pseudo-bulk expression is calculated by summing the single-cell-based reference with cell proportions as weights. On the right, the Bisque transformation of bulk expression is on the *x* axis. Bisque recovers a one-to-one relationship by transforming the bulk expression for improved decomposition accuracy (*right*). **b** Simulation of bulk expression for six individuals based on true adipose snRNA-seq data with increasing gene-specific differences. These differences are modeled as a linear transformation of the summed snRNA-seq counts with coefficient and intercept sampled from half-normal distributions with parameter as indicated on the *x* axis. At $\sigma = 0$, the simulated bulk is simply the sum of the observed single-cell read counts. Performance on *y* axis measured in global Pearson correlation (*R*) (*left*) and root-mean squared deviation (RMSD) (*right*). *Shaded regions* indicate 95% confidence intervals based on bootstrapping with *central lines* indicating the mean observed value. Bisque remains robust to increasing gene-specific variation between single-cell and bulk expression levels. Source data are provided as a Source Data file.

Table 2 Leave-one-out cross-validation in subcutaneous adipose using 6 samples with snRNA-seq and bulk RNA-seq data available.

Method	*R*	RMSD
Bisque	**0.923 ± 0.064**	**0.074 ± 0.034**
CIBERSORTx	0.687 ± 0.450	0.099 ± 0.046
MuSiC	−0.111 ± 0.182	0.427 ± 0.058
BSEQ-sc	−0.113 ± 0.180	0.432 ± 0.058
CIBERSORT	−0.131 ± 0.176	0.416 ± 0.059

Proportions based on snRNA-seq were used as a proxy for the true proportions. Performance measured in Pearson correlation (*R*) and root-mean-square deviation (RMSD) across all 5 identified cell types in each sample. Reported values were averaged across the 6 samples with standard deviation indicated. Bold values indicate the highest performing method with respect to each metric. Source data are provided as a Source Data file.

sequencing technologies. Although CIBERSORTx ($R = 0.687$, RMSD = 0.099) outperformed existing methods, Bisque provided improved accuracy. It should be noted that cell-specific accuracy is more informative than global *R* and RMSD; however, these small sample sizes did not provide robust measures of within-cell-type performance in this cross-validation framework (Supplementary Fig. 1c). We were able to slightly improve the number of detected cell populations by MuSiC, BSEQ-sc, and CIBERSORT when we considered only snRNA-seq reads aligning to exonic regions of the transcriptome, indicating that intronic reads introduced increasing discrepancy between snRNA-seq and bulk RNA-seq in the context of decomposition. However, given that a significant portion of the nuclear transcriptome consists of pre-mRNA, this filtering process removed over 40% of cells detected in the snRNA-seq data. Moreover, Bisque provided improved accuracy over existing methods using this exonic subset of the snRNA-seq data (Supplementary Fig. 1d).

We then applied these decomposition methods to the remaining 100 bulk samples and found that the distribution of cell-proportion estimates produced by Bisque were most concordant with the expected distribution inferred from the limited number of snRNA-seq samples and previously reported proportions[21,22] (Fig. 3a). Although these benchmarks provided a measure of calibration (i.e. the ability to detect cell populations in expected ranges), they did not provide measurements of cell-specific proportion accuracy across individuals. In order to

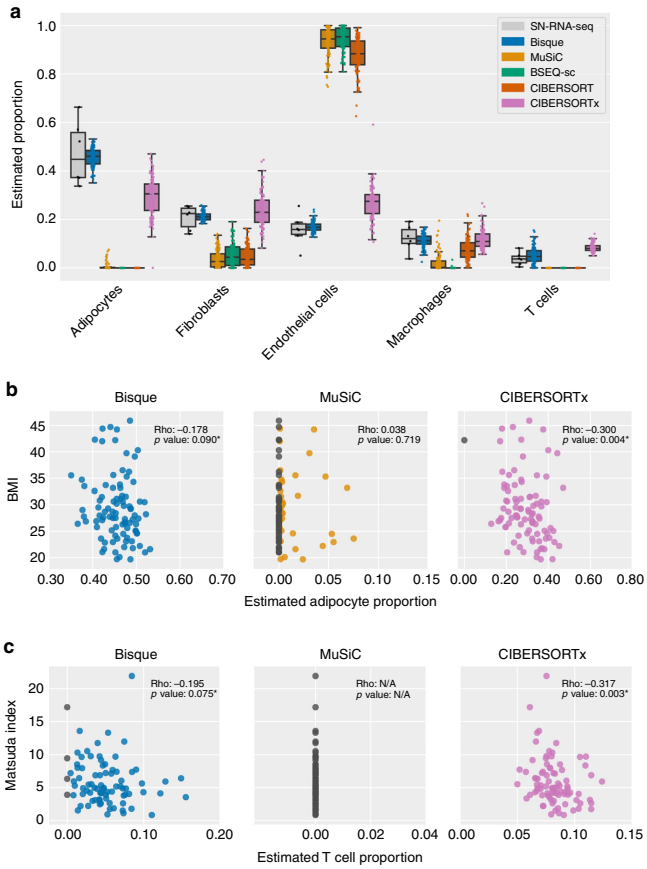


Fig. 3 Decomposition benchmark in human subcutaneous adipose tissue. a Comparison of decomposition estimates from 100 individuals with estimates from 6 individuals with snRNA-seq data available. Each color represents a benchmarked method. *Boxes* indicate the quartiles of the estimated proportions with *whiskers* extending 1.5 times the interquartile range. *Points* are individual samples that are represented by the boxplot. **b**, **c** Scatterplots comparing decomposition estimates with measured phenotypes in 100 individuals. Reported 'rho' corresponds to Spearman correlation and *p*-values indicate the significance of these correlations, with an *asterisk* denoting significance after correction for covariates in a linear-mixed model. CIBERSORT and BSEQ-sc are not shown as they did not detect these cell populations. These examples include the most abundant (adipocytes) and rarest (T cells) cell types identified in the snRNA-seq data. Significance of associations reported in Supplementary Table 1. **b** Adipocyte proportion has been observed to negatively correlate with BMI so we expected a negative correlation. Bisque ($p = 0.030$) and CIBERSORTx ($p = 0.001$) produced significant negative associations after correcting for sex, age, age-squared, and relatedness in a linear-mixed model. **c** T-cell proportion has previously been reported to positively correlate with insulin resistance. Matsuda index decreases with higher insulin resistance so we expected a negative correlation. Bisque ($p = 0.002$) and CIBERSORTx ($p = 0.046$) produced significant negative associations after correcting for diabetes status, sex, age, age-squared, and relatedness in a linear-mixed model. Source data are provided as a Source Data file.

evaluate cell-specific accuracy, we replicated previously reported associations between cell proportions and measured phenotypes. Specifically, we compared cell-proportion estimates from each method to body mass index (BMI) and Matsuda index, a measure of insulin resistance. We measured the significance of these association based on *t*-values estimated in a linear-mixed model accounting for age, age-squared, and sex as fixed effects and relatedness as a random effect.

Obesity is associated with adipocyte hypertrophy, the expansion of the volume of fat cells[23]; thus, we expected a negative association between adipocyte proportion and BMI. Bisque, MuSiC, and CIBERSORTx produced adipocyte proportion estimates that replicate this behavior, while BSEQ-sc and CIBERSORT were unable to detect this cell population (Fig. 3b). The adipocyte proportion estimates produced by Bisque ($p = 0.030$) and CIBERSORTx ($p = 0.001$) had a significant negative association with BMI (Supplementary Table 1a). In addition, macrophage abundance has been shown to increase in adipose tissue with higher levels of obesity, concomitant with a state of low grade inflammation[24]. Each method detected macrophage

populations that positively associated with BMI; however, only Bisque ($p < 0.001$), BSEQ-sc ($p = 0.004$), and CIBERSORTx ($p = 0.034$) reached significance (Supplementary Table 1b).

T cells were the least abundant cell-type population identified from the snRNA-seq data, constituting around 4% of all sequenced nuclei. The abundance of T cells has been observed to positively correlate with insulin resistance[25]. Thus, we compared decomposition estimates for T-cell proportions to Matsuda index. As a lower Matsuda index indicates higher insulin resistance, we expect a negative association between T-cell proportion and Matsuda index. Proportion estimates produced by Bisque and CIBERSORTx followed this trend while the remaining existing methods did not identify T cells in the bulk samples (Fig. 3c). We found this association significant for Bisque ($p = 0.002$) and CIBERSORTx ($p = 0.046$) (Supplementary Table 1c) after correcting for diabetes status, as Matsuda index may not be informative in these individuals[26].

Evaluation of decomposition performance in cortex tissue. We also benchmarked these decomposition methods using expression data collected from the dorsolateral prefrontal cortex (DLPFC). This dataset was generated by the Rush Alzheimer's Disease (AD) Center[27] and includes 636 postmortem bulk RNA-seq samples. The Religious Orders Study and Rush Memory and Aging Project were approved by an IRB of Rush University Medical Center. Both bulk RNA-seq and snRNA-seq data were collected from 8 of the individuals (Table 1). Using the same pipeline we used to process the adipose dataset, we identified 11 clusters: 3 neuronal subtypes, 2 interneuronal subtypes, 2 astrocyte subtypes, oligodendrocytes, oligodendrocyte progenitor cells, and microglia (Supplementary Fig. 2a). We observed a higher overlap in marker genes for these clusters than in those identified in the adipose dataset (average of 10% of marker genes shared between clusters in DLPFC compared to 3% in adipose) (Supplementary Fig. 4a, b).

We again applied leave-one-out cross-validation on the 8 individuals with both RNA-seq and snRNA-seq data available. In this example, we randomly sampled 25% of the nuclei in the snRNA-seq data to accommodate CIBERSORTx (which is currently web-based and restricts the size of files that can be processed). Bisque was able to detect each cell population identified from the snRNA-seq data with high global accuracy ($R = 0.924$, RMSD = 0.029) while MuSiC ($R = -0.192$, RMSD = 0.173), BSEQ-sc ($R = 0.098$, RMSD = 0.120), and CIBERSORT ($R = -0.281$, RMSD = 0.197) did not detect a number of cell populations (Table 3, Supplementary Fig. 2b, c). Bisque also provided higher accuracy than CIBERSORTx ($R = 0.671$, RMSD = 0.070). However, we found that the performance of the existing methods improved when estimates with subtypes were summed together (Supplementary Fig. 2d). Although each method was able to quantify major cell populations after merging subtypes, Bisque was able to distinguish between these closely related cell populations. Interestingly, we found that in both adipose and DLPFC, endothelial cell proportions were overestimated by each of the existing methods.

Table 3 Leave-one-out cross-validation in dorsolateral prefrontal cortex using 8 samples with snRNA-seq and bulk RNA-seq data available.

Method	*R*	RMSD
Bisque	**0.924 ± 0.062**	**0.029 ± 0.010**
CIBERSORTx	0.671 ± 0.153	0.070 ± 0.019
MuSiC	−0.192 ± 0.107	0.173 ± 0.013
BSEQ-sc	0.098 ± 0.216	0.120 ± 0.023
CIBERSORT	−0.281 ± 0.049	0.197 ± 0.012

Proportions based on snRNA-seq were used as a proxy for the true proportions. Performance measured in Pearson correlation (*R*) and root-mean-square deviation across all 11 identified cell types in each sample. Reported values were averaged across the 8 samples with standard deviation indicated. We performed these experiments with 25% of the snRNA-seq data in order to accommodate the file size limit of the current web-based implementation of CIBERSORTx. Bold values indicate the highest performing method with respect to each metric. Source data are provided as a Source Data file.

We applied these decomposition methods to the remaining 628 individuals and compared the distribution of estimates to the proportions observed in the 8 snRNA-seq samples. We found that Bisque was able to detect each cell population and produced estimates that were closest in mean to the snRNA-seq observations (Fig. 4a). The increased accuracy of Bisque over existing methods persisted when we merged closely related subtypes (Supplementary Fig. 2e). Moreover, immunohistochemistry (IHC) analyses on a 70 of these samples found similar proportions of major cell populations[28], confirming the relative accuracy of snRNA-seq-based estimates of cell proportions.

Again, to determine cell-specific decomposition accuracy, we replicated known associations between cell-type proportions and measured phenotypes in the 628 individuals. For these analyses, we compared cell-proportion estimates to each individual's Braak stage and physician cognitive diagnostic category at time of death. Braak stage is a semiquantitative measure of neurofibrillary tangles, ranging in value from 0 to 5 with increasing severity. The cognitive diagnostic category provides a semiquantitative measure of dementia severity, where a code of 1 indicates no cognitive impairment and 5 indicates a confident diagnosis of AD by physicians. We determined the significance of these associations based on *t*-values estimated by a linear regression model that accounted for age, age-squared, and sex.

Neuronal death is a hallmark symptom of AD[29]. Therefore, we expected to find a negative association between cognitive diagnosis and neuron proportion. We found that each decomposition method provides estimates of total neuron proportion that tend to decrease with cognitive diagnostic category (Fig. 4b). Each method generates proportions with negative association with cognitive diagnosis. Each method, excluding BSEQ-sc, reached significance in this model ($p \leq 0.001$ for each method) (Supplementary Table 2a). As another example, we compared each individual's Braak stage to their estimated proportion of microglia, a relatively small cell population that constituted roughly 5% of the sequenced nuclei. Microglia activation has been observed to increase with AD severity[30]. We used Braak stage as a proxy for AD severity and expected a positive association between microglia proportion and Braak stage. Bisque and MuSiC provided estimates that follow this expected trend (Fig. 4c). Only Bisque produced estimates with a significant positive association ($p = 0.001$) (Supplementary Table 2b). Interestingly, we observed a decrease in microglia proportions estimated by Bisque in Braak stage 6 individuals, which has been previously observed in AD patients[31].

Runtime comparison of reference-based decomposition methods. Given the large amounts of transcriptomic data that are becoming available, we also benchmarked these decomposition methods in terms of runtime. In the subcutaneous adipose dataset, which included 100 bulk RNA-seq samples and 6 snRNA-seq samples with about 1800 nuclei sequenced per individual, Bisque was able to estimate cell proportions efficiently compared to existing methods. Bisque (1 s) and MuSiC (1 s) provided decomposition estimates faster than BSEQ-sc (26 s), CIBERSORT (27 s), and CIBERSORTx (389 s) (Fig. 5a). Bisque also provided improved efficiency in processing the reduced DLPFC dataset, which included 628 bulk RNA-seq samples and 8

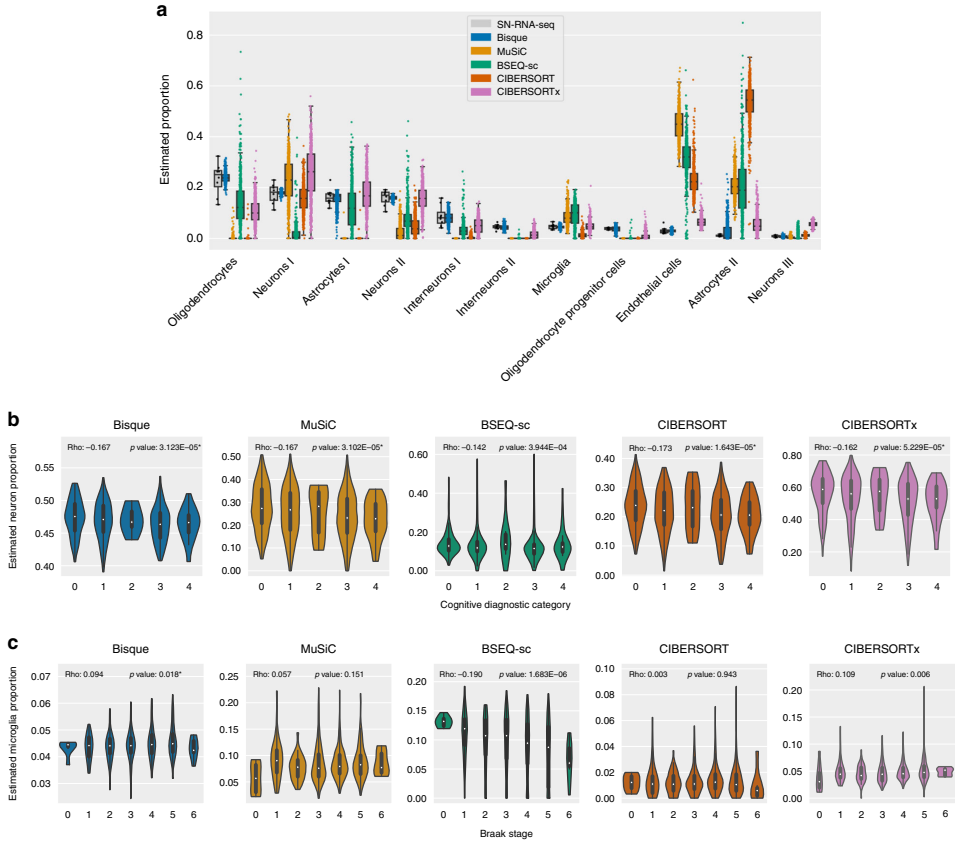


Fig. 4 Decomposition benchmark in human dorsolateral prefrontal cortex tissue. We randomly sampled 25% of the nuclei in the snRNA-seq data to accommodate the file size limit of the web-based implementation of CIBERSORTx at the time of writing. **a** Comparison of decomposition estimates from 628 individuals with estimates from 8 individuals with snRNA-seq data available. Each color represents a benchmarked method. *Boxes* indicate the quartiles of the estimated proportions with *whiskers* extending 1.5 times the interquartile range. *Points* are individual samples that are represented by the boxplot. **b**, **c** Violin plots depicting association of decomposition estimates aggregated into major cell types with measured phenotypes in 628 individuals. Reported 'rho' corresponds to Spearman correlation and *p*-values indicate the significance of these correlations, with an *asterisk* denoting both an expected effect direction and significance after correction for covariates. Examples shown are for the most abundant (neurons) and least abundant (microglia) populations detected in the snRNA-seq data. Significance of associations reported in Supplementary Table 2. **b** Neuronal degeneration has been observed in patients diagnosed with Alzheimer's disease (AD). Cognitive diagnostic category measures a physician's diagnosis of cognitive impairment (CI), with 0 indicating no CI and 4 indicating a confident AD diagnosis. We expected a negative correlation between neuron proportion and cognitive diagnostic category. **c** Microglia proportion has been observed to positively correlate with increased severity of AD symptoms, such as neurofibrillary tangles. Braak stage provides a semiquantitative measure of tangle severity, so we expected an overall positive correlation between microglia proportion and Braak stage. In addition, a decrease in microglia abundance has been previously reported at Braak stages 5 through 6 in AD patients. Only Bisque produced estimates with a significant positive association ($p = 0.001$) after correcting for sex, age, and age-squared in a linear regression model. Source data are provided as a Source Data file.

snRNA-seq samples with around 2125 nuclei per individual. Bisque (4 s) and MuSiC (10 s) estimated cell proportions relatively quickly compared to BSEQ-sc (273 s), CIBERSORT (298 s), and CIBERSORTx (6566 s) (Fig. 5b).

Robustness of the reference-based decomposition model. Our reference-based decomposition method is based on the assumption that cell populations are equally represented in single-cell and bulk RNA sequencing of the same tissue samples. As this assumption may be violated[32], we explored the performance of our model as we relaxed this assumption in simulations. First, we simulated snRNA-seq data where cell proportions were increasingly biased. Using the DLPFC snRNA-seq data, we downsampled or upsampled the cells identified as microglia at varying levels and performed decomposition. Indeed, the absolute estimates produced by Bisque propagated these shifts in snRNA-seq proportions. However, we found that our estimated microglia proportions, regardless of these shifts,

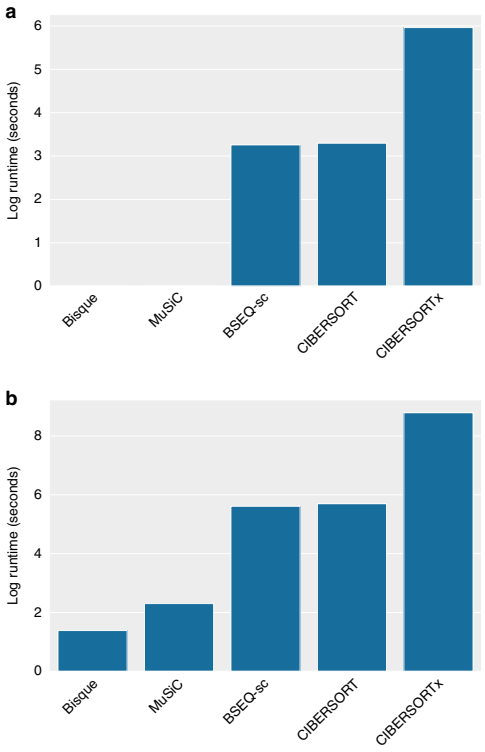


Fig. 5 Runtime comparisons in log-transformed seconds for benchmarked reference-based decomposition methods. a Runtime for subcutaneous adipose dataset, which included 100 RNA-seq samples and 6 snRNA-seq samples with around 1800 nuclei per individual. **b** Runtime for dorsolateral prefrontal cortex dataset, which included 628 RNA-seq samples and 8 snRNA-seq samples. We benchmarked each method using around 2125 nuclei per snRNA-seq sample. Source data are provided as a Source Data file.

maintained an expected positive association with Braak stage. This positive association served as evidence for the correlation between these estimates and the true microglia proportions (Supplementary Fig. 5a). Given these results, we suggest that users take note of this behavior if both the mean abundances are important for downstream analysis and the single-cell reference data is known to be significantly biased against specific cell populations of interest.

Next, we simulated a situation where an unknown cell population contributes to bulk expression but is not represented in the snRNA-seq reference data. For situations where this unknown contribution varies across the bulk dataset, we simulated bulk expression by mixing the observed bulk expression for the DLPFC dataset with increasing amounts of expression observed in the adipose dataset. To determine the effect of unknown cell populations on our model, we analyzed the distribution of residual norms produced by the method. These residual norms provide a measure of the difference between the vector of observed bulk and expression reference weighted by the estimated proportions across all genes for each individual. As we increased the contribution from unknown cell types, the residual norm values tend to increase (Supplementary Fig. 5b). In our simulation framework, this variability in unknown cell-type contribution could be qualitatively identified by the presence of a multimodal residual norm distribution.

Given that single-cell datasets still remain relatively small compared to bulk datasets, we also explored the impact of sample size in the reference single-cell data on the performance of Bisque. In the DLPFC dataset, we saw a drop in performance when using less than four randomly selected snRNA-seq samples (Supplementary Fig. 5c). This threshold is likely to differ between experiments, though we recommend at least three single-cell samples to generate reference data.

Finally, as marker gene selection can vary between studies, we were interested in the performance of Bisque as we varied the number of marker genes. Again, we measured cell-type proportion estimation performance for microglia in the DLPFC dataset by correlating the estimates with Braak stage, which is known to have a positive association. We recalculated this correlation as we removed marker genes for this cell type. We removed marker genes in order of both decreasing and increasing log-fold change, which provides a measure of the importance of marker genes for identifying this cell type. In both procedures, we observe that as we remove an increasing percentage of the 102 identified marker genes, performance remains stable until a shared drop off point around 75% (Supplementary Fig. 5d). As we observed this trend in both marker gene removal schemes, we assume that a relatively few number of marker genes, regardless of their log-fold change magnitude, can be used to accurately estimate cell-type proportions. These results suggest that as long as a core set of marker genes are present, variations in less important marker genes will have little effect on downstream analyses.

Marker-based decomposition using cell-type marker genes. Although a reference profile from snRNA-seq can help to decompose bulk-level gene expression, it may not be available for the same dataset. The majority of bulk RNA-seq datasets do not have corresponding snRNA-seq data in the same set of individuals. However, marker gene information from prior experiments can still be applied to distinct expression datasets of the same tissue. The basis of most decomposition methods relies on the logic that as the proportion of a cell type varies across individuals, the expression of its marker genes will tend to correlate in the same direction as its cell-type proportion. This linear co-variation can be captured in a principal components analysis (PCA). Under the same argument, the more cell-type-specific a marker gene is, the more its expression will reflect its cell-type proportion. These observations form the basis for BisqueMarker, a weighted PCA-based (wPCA) decomposition approach. Genes that are more specifically expressed within a cell type will provide more information than genes with shared expression across cell types. To estimate cell-type proportions without the use of cell-type-specific gene expression information, we applied wPCA to bulk-level adipose tissue expression.

For each cell type, we extracted the first PC from a wPCA of the expression matrix of its markers. The expression matrix was corrected for the first global expression PC as a covariate so that wPCA estimates would not reflect technical variation. We first confirmed that these genes were distinct across cell types. If 2 cell types share a high proportion of marker genes, the wPCA estimates from bulk RNA-seq will correlate highly. We then investigated whether the second or third PC could have represented cell-type proportions. The percent of variance explained by the first PC was typically 30–60% across adipose cell types, and additionally, over 90% of the markers correlated in the same direction as the first PC. In contrast, roughly 50–70% of

markers correlated in the same direction as the second or third PC. As performed for reference-based decomposition, we correlated phenotypes with cell-type proportions estimated by BisqueMarker. We identified the same associations as with reference-based decomposition, demonstrating its validity when a reference is not available (Supplementary Table 1). Similarly, we observed the same trends between estimated cell-type abundances and phenotypes as we did using our reference-based method in the DLPFC cohort (Supplementary Table 2).

Discussion

Bisque effectively leverages single-cell information to decompose bulk expression samples, outperforming existing methods in datasets with snRNA-seq data available. In simulations, we demonstrated that the decomposition accuracy of Bisque is robust to increasing variation between the generation of the reference profile and bulk expression, which is a significant issue when comparing snRNA-seq and bulk RNA-seq data. In observed bulk expression, our reference-based method accurately estimates cell proportions that are consistent with previously reported distributions and reliably detects rare cell types. We found that these estimates consistently follow expected trends with measured phenotypes, suggesting that cell-specific estimates of proportion are sufficiently accurate to extract relevant biological signals. In addition, differences in tissue structure can lead to significant differences in the quality of single-cell expression data[33]. We demonstrated the improved performance of our method in adipose and DLPFC, two distinct tissues, suggesting that Bisque is robust across different tissue types.

The cell-type proportion estimates determined by Bisque may be utilized to effectively identify cell-type-specific interactions, such as expression quantitative trait loci (eQTLs), and adjust for confounding effects from variability in cell populations. With this reference-based approach, single-cell sequencing of a subset of samples from large-scale bulk expression cohorts can provide high power to detect cell-specific associations in complex phenotypes and diseases.

However, we note that there are limitations to this reference-based method that users should consider. First, if the number of individuals with single-cell data available is small, the reference profile and gene-specific transformations may become unreliable. In addition, a key assumption of our transformation framework is that single-cell-based estimates of cell proportions accurately reflect the true proportions we wish to estimate. As a result of this assumption, Bisque provides estimates of cell proportions reported by the single-cell technology used to generate the reference data. Given that snRNA-seq can provide less bias in isolating specific cell types compared to scRNA-seq[34,35], we expect these estimates to be useful for downstream analyses such as those previously discussed. Nevertheless, the accuracy of Bisque may decrease if the proportion of cell types captured by single-cell experiments differs significantly from the true physiological distributions. Therefore, we advise users to take caution if there is a known significant bias in the single-cell measurements of a tissue, such as severe underrepresentation of a cell type of interest[32,35], that can affect downstream analysis. Our results demonstrate that even with these limitations, Bisque can be used to provide cell-type specific biological insight in relevant datasets.

In cases where these described issues may be significant, BisqueMarker provides cell-type abundance estimations using only known marker genes. Although this reference-free method may be less accurate than reference-based methods, it does not depend on single-cell based estimates of cell proportions or expression profiles, but rather on the fact that the expression in certain genes differs across different cell types; moreover, this method also does not model explicitly the expression level, and it is thus robust to biases in the single-cell sequencing protocol. We found that BisqueMarker estimates followed expected trends with measured phenotypes; however, it should be noted that this method estimates relative differences in abundances that cannot be compared across cell types. Also, given the semi-supervised nature of this method, these cell-type abundance estimates may include signals from technical or other biological variation in the data. Therefore, we highly suggest applying this method to data that is properly normalized with sources of undesired variation removed.

Methods

Processing bulk expression data. Paired-end reads were aligned with STAR v2.5.1 using default options. Gene counts were quantified using featureCounts v1.6.3. For featureCounts, fragments were counted at the gene-name level. Alignment and gene counts were generated against the GRCh38.p12 genome assembly. STAR v2.5.1 and GRCh38.p12 were included with CellRanger 3.0.2, which was used to process the single-nucleus data.

Processing single-nucleus expression data. Reads from single nuclei sequenced on the 10x Genomics Chromium platform were aligned and quantified using the CellRanger 3.0.2 count function against the GRCh38.p12 genome assembly. To account for reads aligning to both exonic and intronic regions, each gene transcript in this reference assembly was relabeled as an exon as CellRanger counts exonic reads only. We perform this additional step since snRNA-seq captures both mature mRNA and pre-mRNA, the latter of which includes intronic regions.

After aggregating each single-nucleus sample with the CellRanger aggr function, the full dataset was processed using Seurat v3.0.0[36]. The data were initially filtered for genes expressed in at least 3 cells and filtered for cells with reads quantified for between 200 and 2500 genes. We further filtered for cells that had percentage of counts coming from mitochondrial genes less than or equal to 5%. The data were normalized, scaled, and corrected for mitochondrial read percentages with sctransform v0.2.0[37] using default options.

To identify clusters, Seurat employs a shared nearest neighbor approach. We identified clusters using the top 10 principal components of the processed expression data with resolution set at 0.2. The resolution parameter controls the number of clusters that will be identified, and suggested values vary depending on the size and quality of the dataset. We chose a value that produced 6 clusters in the adipose dataset and 13 clusters in the DLPFC dataset and visualized the clustering results with UMAP[38].

Marker genes were identified by determining the average log-fold change of expression of each cluster compared to the rest of the cells. We identified marker genes as those with an average log-fold change above 0.25. The significance of the differential expression of these genes was determined using a Wilcoxon rank sum test. Only genes that were detected in at least 25% of cells were considered. Clusters with many mitochondrial genes as markers (nine genes detected in both datasets) were removed from both datasets. In addition, a cluster with only three marker genes was removed from the DLPFC datasets. Finally, we remove mitochondrial genes from the list of marker genes for decomposition as we assume reads aligning to the mitochondrial genome originate from extra-nuclear RNA in the snRNA-seq dataset (targeting nuclear RNA).

Clusters were labeled by considering cell types associated with the identified marker genes. Marker genes were downloaded from PanglaoDB[39] and filtered for entries validated in human cells. For each gene, we count the possible cell-type labels. Each cluster was labeled as the most frequent cell type across all of its marker genes, with each label associated with a gene weighted by the average log-fold change. If multiple clusters shared a cell-type label, we consider each cluster a subtype of this label.

Exon-aligned reads were processed in the same exact procedure but snRNA-seq data was aligned to just exonic regions. Cluster names were manually changed for both datasets when aligned to exons to match the clusters from intronic reads as well. Specifically, for clusters identified in the exonic data not found in the full data, we relabeled as the label with the highest score found in the full data. These relabeled clusters were similar in proportion to the corresponding cluster in the full dataset.

Bisque reference-based decomposition model. We assume that only a subset of genes are relevant for estimating cell-type composition. For the adipose and DLPFC datasets, we selected the marker genes identified by Seurat as described previously. Moreover, we filter out genes with zero variance in the single-cell data, unexpressed genes in the bulk expression, and mitochondrial genes. We convert the remaining gene counts to counts-per-million to account for variable sequencing depth. For m genes and k cell types, a reference profile $\mathbf{Z} \in \mathbf{R}^{m \times k}$ is generated by averaging relative abundances within each cell type across the entire single-cell dataset.

Although there is a strong positive correlation between bulk and single-cell-based pseudo-bulk (summed single-cell counts) expression data, we observe that

the relationship is not one-to-one and varies between genes. This behavior indicates that the distribution of observed bulk expression may significantly differ from the distribution of the single-cell profile weighted by cell proportions. We propose transforming the bulk data to maximize the global linear relationship across all genes for improved decomposition. Our goal is to recover a one-to-one relationship between the transformed bulk and expected convolutions of the reference profile based on single-cell based estimates of cell proportions. This transformed bulk expression better satisfies the assumptions of regression-based approaches under sum-to-one constraints.

Cell-type proportions $\mathbf{p} \in \mathbf{R}^{k \times n'}$ are determined by counting the cells with each label in the single-cell data for n' individuals. Given these proportions and the reference profile $\mathbf{Z}$, we calculate the pseudo-bulk for the single-cell samples as $\mathbf{Y} = \mathbf{Zp}$, where $\mathbf{Y} \in \mathbf{R}^{m \times n'}$. For each gene j, our goal is to transform the observed bulk expression across all n bulk samples $\mathbf{X}_j \in \mathbf{R}^n$ to match the mean and variance of $\mathbf{Y}_j \in \mathbf{R}^{n'}$; hence, the transformation of $\mathbf{X}_j$ will be a linear transformation.

If individuals with both single-cell and bulk expression are available, we fit a linear regression model to learn this transformation. Let $\mathbf{X}'_j \in \mathbf{R}^{n'}$ denote the expression values for these n' overlapping individuals. We fit the following model (with an intercept) and apply the model to the remaining bulk samples as our transformation:

$$Y_j = \beta_j X'_j + \epsilon_j \tag{1}$$

If there are no single-cell samples that have bulk expression available, we assume that the observed mean of $\mathbf{Y}_j$ is the true mean of our goal distribution for the transformed $\mathbf{X}_j$. We further assume that the sample variance observed in $\mathbf{Y}_j$ is larger than the true variance of the goal distribution, as the number of single-cell samples is typically small. We use a shrinkage estimator of the sample variance of $\mathbf{Y}_j$ that minimizes the mean squared error and results in a smaller variance than the unbiased estimator:

$$\tilde{\sigma}_j^2 = \frac{1}{n'+1} \sum_{i=1}^{n'} (Y_{i,j} - \bar{Y}_j)^2 \tag{2}$$

We transform the remaining bulk as follows:

$$X_{j,\text{transformed}} = \frac{X_j - \bar{X}_j}{\sigma_{X_j}} \tilde{\sigma}_j + \bar{Y}_j \tag{3}$$

where a bar indicates the mean value of the observed data and $\sigma_{\mathbf{X}j}$ is the unbiased sample variance of $\mathbf{X}_j$.

To estimate cell-type proportions, we apply non-negative least-squares regression with an additional sum-to-one constraint to the transformed bulk data. For individual i, we minimize the following with respect to the cell-proportion estimate $\mathbf{p}_i$:

$$||Zp_i - X_{i,\text{transformed}}||_2 s.t. p_i \geq 0, \sum p_i = 1 \tag{4}$$

Simulating bulk expression based on single-nucleus counts. We simulate the base bulk expression as the sum of all counts across cells/nuclei sequenced from an individual. To introduce gene-specific variation between the bulk and single-cell data, we sample a coefficient β_j and an intercept α_j from a half-normal (HN) distributions:

$$\beta_j \sim \text{HN}(\sigma) + 1 \tag{5}$$

$$\alpha_j \sim \text{HN}(\sigma) \tag{6}$$

At $\sigma = 0$, the base simulated bulk expression remains unchanged. We used a HN distribution to ensure coefficients and intercepts are positive. Although our method can handle negative coefficients, this simulation model assumes expression levels have a positive correlation across technologies. We performed 10 replicates of this data-generating process at each σ in 0, 5, 10, 20. Decomposition performance on these data were measured in terms of global R and RMSD and plotted with 95% confidence intervals based on bootstrapping.

Measuring significance of cell proportion-trait association. Reported associations were measured in terms of Spearman correlation. To determine the statistical significance of these associations while accounting for possible confounding factors, we applied two approaches. For the adipose dataset, which consisted entirely of twin pairs, we applied a linear-mixed-effects model (R nlme package) with random effects accounting for family. For the DLPFC dataset, we assumed individuals were unrelated and fit a simple linear model (R base package). In each model, we include cell-type proportion, age, age-squared, and sex as covariates. We introduced an additional covariate for diabetes status when regressing Matsuda index due to a known significant association between these two variables. We test whether the cell proportion-effect estimates deviate significantly from 0 using a t-test. Each R method implements the described model fitting and significance testing.

Bisque marker-based decomposition model. In order to estimate cell-type proportions across individuals without the use of a cell-type-specific gene expression panel as reference, we use a weighted PCA approach. BisqueMarker requires a set of marker genes for each cell type as well as the specificity of each marker determined by the fold-change from a differential expression analysis. Typical single-cell RNA-seq workflows calculate marker genes and provide both p-values and fold-changes, as in Seurat[36]. For each cell type, we take statistically significant marker genes (FDR < 0.05) ranked by p-value. A weighted PCA is calculated on the expression matrix using a subset of the marker genes by first scaling the expression matrix and multiplying each gene column by its weight (the log-fold change) $\mathbf{XW}$, where $\mathbf{X}$ is the sample by gene expression matrix and $\mathbf{W}$ is a diagonal matrix with entries equal to log-fold change of the corresponding gene. The bulk expression $\mathbf{X}$ should be corrected for global covariates so that the proportion estimates do not reflect this global variation. The first PC calculated from $\mathbf{XW}$ is used as the estimate of the cell-type proportion. This allows cell-type-specific genes to be prioritized over more broadly expressed genes. Alternatively, if weights are not available, PCA can be run on the matrix $\mathbf{X}$ and the first PC can be used.

In order to select marker genes, we iteratively run the above PCA procedure on a specified range of markers (from 25 to 200) and calculate the ratio of the first eigenvalue to the second. We then select the number of marker genes to use that maximizes this ratio. This procedure is similar to other methods which select the number of markers to use by maximizing the condition number of the reference matrix[13].

Software used. Single-nucleus RNA-seq data were aligned using CellRanger 3.0.2 against the GRCh38.p12 genome assembly. Bulk RNA-seq data were aligned with STAR 2.5.1 and quantified using featureCounts 1.6.3, both against the GRCh38.p12 genome assembly. R 3.5.1 was used for further processing and decomposition experiments. The Seurat v3.0.0 R package was used to filter, cluster, and identify cell-type marker genes from the single-nucleus data. The sctransform 0.2.0 R package was used to normalize and scale the single-nucleus data. Bisque 1.0, xbioc 0.1.7, Biobase 2.4.2, MuSiC 0.1.1, bseqsc 1.0, CIBERSORT v1.06, and CIBERTSORTx were all used for decomposition using the processed bulk and single-nucleus RNA-seq data. The R nlme 3.1-127 package was used for linear-mixed-model association. All visualizations and were generated with Python 3.7.2 using Seaborn 0.9.0, Matplotlib 3.0.3, Pandas 0.24.2, and Numpy 1.16.2, sklearn 0.20.3, and scipy 1.2.1.

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

Data availability

The adipose data used in these analyses are available from the corresponding authors upon reasonable request. The DLPFC data are available on Synapse (10.7303/syn3219045). Single-nucleus RNA-seq data (https://www.synapse.org/#!Synapse:syn16780177), bulk RNA-seq data (https://www.synapse.org/#!Synapse:syn3388564), and phenotypes (https://www.synapse.org/#!Synapse:syn3191087) are available under controlled use conditions set by human privacy regulations. A data use agreement is required to access these data. The source data underlying Tables 2 and 3, Figs. 2–5, Supplementary Tables 1 and 2, and Supplementary Figs. 1, 2, 3, 4, 5 are provided as a Source Data file.

Code availability

Bisque is available as an R package named "BisqueRNA" that is available on CRAN and Bioconda. The source code for this package is available at https://github.com/cozygene/bisque and is under the GPL-3 license.

Received: 3 June 2019; Accepted: 25 March 2020;
Published online: 24 April 2020

References

1. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68–A77 (2015).
2. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
3. Bruning, O. et al. Confounding factors in the transcriptome analysis of an in-vivo exposure experiment. *PLoS ONE* **11**, e0145252 (2016).
4. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* **12**, 298–306 (2012).
5. Rahier, J., Goebbels, R. M. & Henquin, J. C. Cellular composition of the human diabetic pancreas. *Diabetologia* **24**, 366–371 (1983).
6. Shen-Orr, S. S. et al. Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).

Supplementary Tables

a Association of adipocyte proportion estimates in adipose tissue with BMI

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	-0.178	0.090	-0.282	0.126	-2.240	**0.030**
MuSiC	0.038	0.719	-0.081	0.108	-0.754	0.455
BSEQ-sc	-	-	-	-	-	-
CIBERSORT	-	-	-	-	-	-
CIBERSORTx	-0.300	**0.004**	-0.361	0.100	-3.624	**0.001**
BisqueMarker	-0.227	**0.030**	-0.304	0.096	-3.154	**0.003**

b Association of macrophage proportion estimates in adipose tissue with BMI

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	0.389	**1.291e-04**	0.460	0.099	4.671	**3.078e-05**
MuSiC	0.065	0.540	0.034	0.110	0.308	0.760
BSEQ-sc	0.238	**0.022**	0.278	0.092	3.013	**0.004**
CIBERSORT	0.239	**0.022**	0.162	0.102	1.597	0.118
CIBERSORTx	0.273	**0.009**	0.224	0.102	2.192	**0.034**
BisqueMarker	0.296	**0.004**	0.253	0.103	2.465	**0.018**

c Association of T cell proportion estimates in adipose tissue with Matsuda index

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	-0.195	0.075	-0.387	0.116	-3.328	**0.002**
MuSiC	-	-	-	-	-	-
BSEQ-sc	-	-	-	-	-	-
CIBERSORT	-	-	-	-	-	-
CIBERSORTx	-0.317	**0.003**	-0.230	0.111	-2.068	**0.046**
BisqueMarker	-0.294	**0.007**	-0.188	0.100	-1.874	0.069

Supplementary Table 1: Significance of associations of estimated cell proportions and measured phenotypes in 100 subcutaneous adipose tissue samples. We fit a linear mixed-effects model (LMM) to account for the twin structure of the dataset as a random effect, with additional fixed effects to account for age, age-squared, and sex. Expected effect directions were based on previously reported findings. An entry of '-' indicates that the method did not detect the indicated cell population in any of the samples. Bold values were found to be significant at $\alpha = 0.05$ and in expected directions.
a Association of adipocyte proportion with BMI. A negative association was expected.
b Association of macrophage proportion with BMI. A positive association was expected.
c Association of T cell proportion with Matusda index, a measure of insulin resistance. A negative association was expected. An additional covariate accounting for diabetes status was added to the LMM due to previously reported significant associations with Matsuda index.
Source data are provided as a Source Data file.

a Association of neuron proportion estimates in DLPFC tissue with cognitive diagnosis

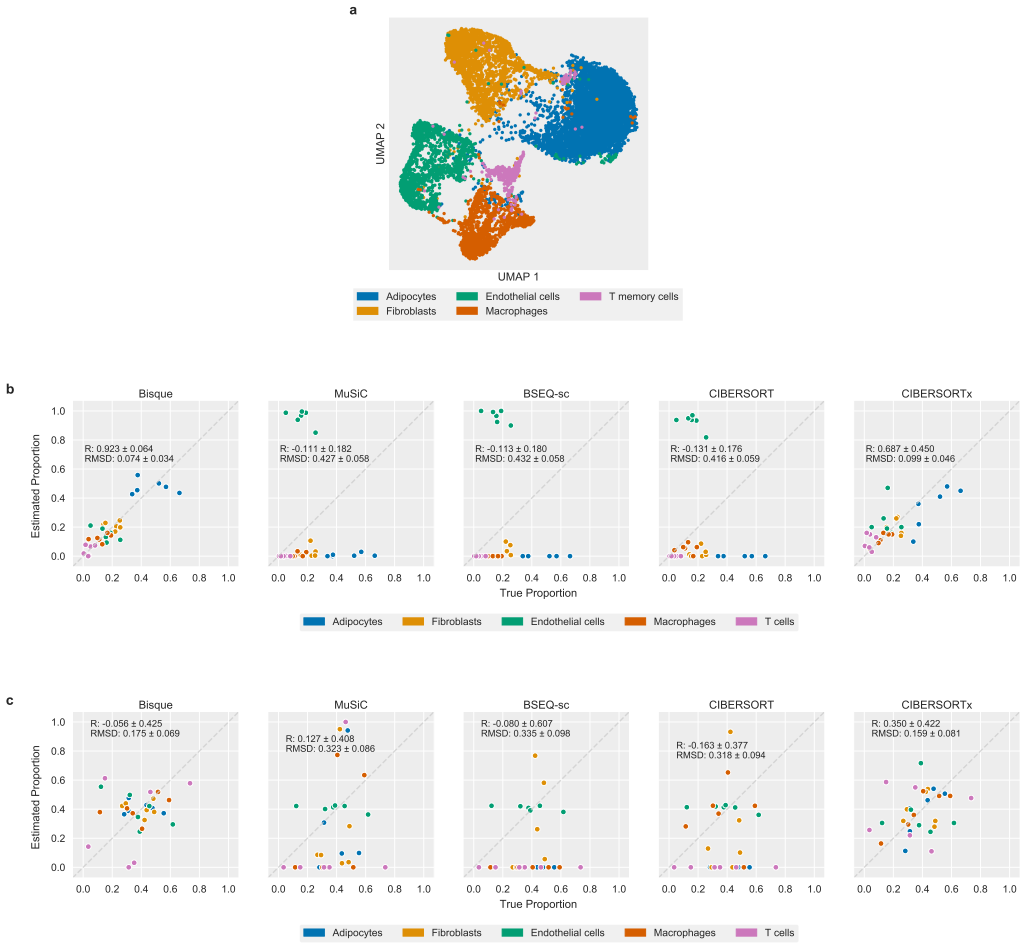
Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	-0.167	**3.123e-05**	-0.145	0.039	-3.705	**2.305e-04**
MuSiC	-0.167	**3.102e-05**	-0.147	0.039	-3.742	**1.995e-04**
BSEQ-sc	-0.142	**3.944e-04**	-0.053	0.039	-1.341	0.180
CIBERSORT	-0.173	**1.643e-05**	-0.155	0.039	-3.971	**7.998e-05**
CIBERSORTx	-0.162	**5.229e-05**	-0.127	0.039	-3.237	**0.001**
BisqueMarker	-0.141	**4.383e-04**	-0.142	0.039	-3.645	**2.897e-04**

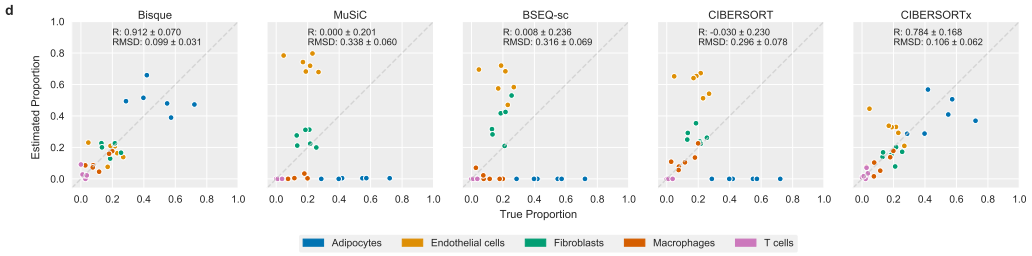
b Association of microglia proportion estimates in DLPFC tissue with Braak stage

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	0.094	**0.018**	0.118	0.037	3.220	**0.001**
MuSiC	0.057	0.151	0.019	0.037	0.509	0.611
BSEQ-sc	-0.190	1.683e-06	-0.166	0.037	-4.525	7.244e-06
CIBERSORT	0.003	0.943	-0.005	0.037	-0.137	0.891
CIBERSORTx	0.109	**0.006**	0.056	0.037	1.517	0.130
BisqueMarker	0.092	**0.021**	0.054	0.037	1.444	0.149

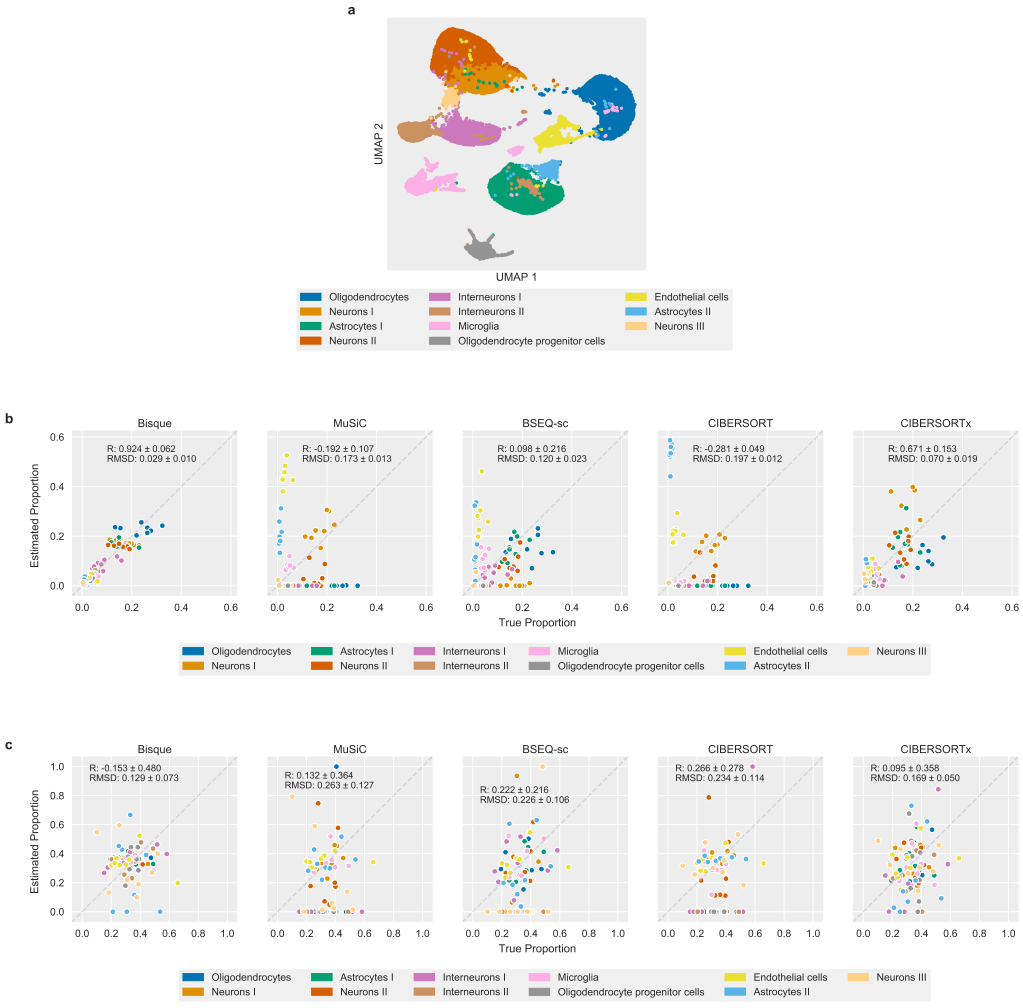
Supplementary Table 2: Significance of associations of estimated cell proportions and measured phenotypes in 628 DLPFC tissue samples. We fit a linear model with covariates to account for age, age-squared, and sex. Expected effect directions were based on previously reported findings. Bold values were found to be significant at $\alpha = 0.05$ and in expected directions.
a Association of neuron proportion with cognitive diagnosis category. A negative association was expected.
b Association of microglia proportion with Braak stage, a measure of neurofibrillary tangles. A positive association was expected.
Source data are provided as a Source Data file.

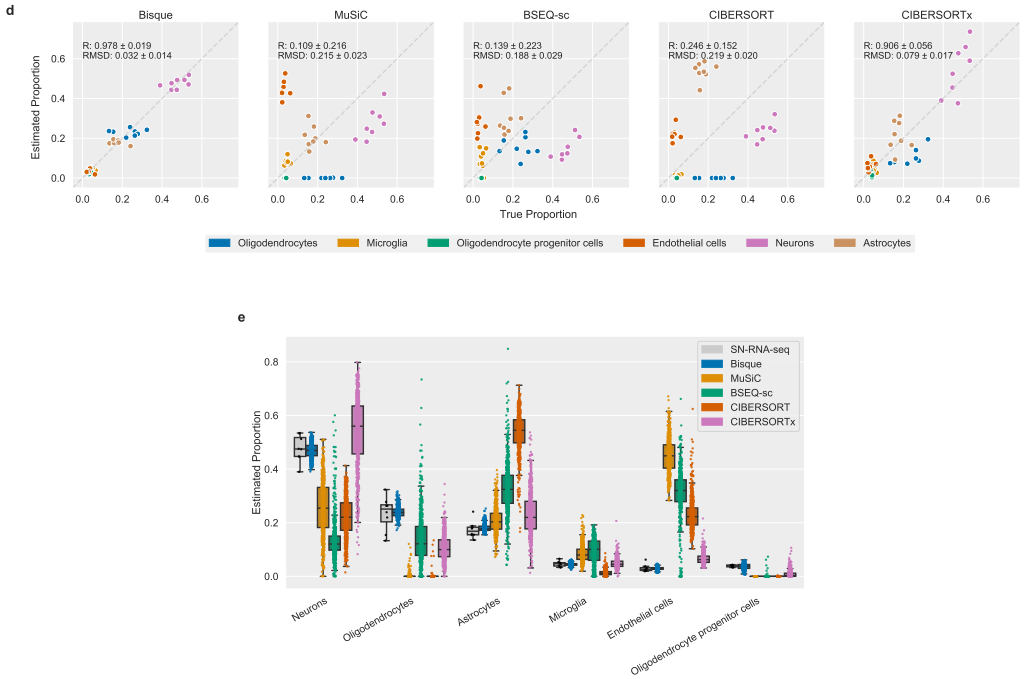
Supplementary Figures



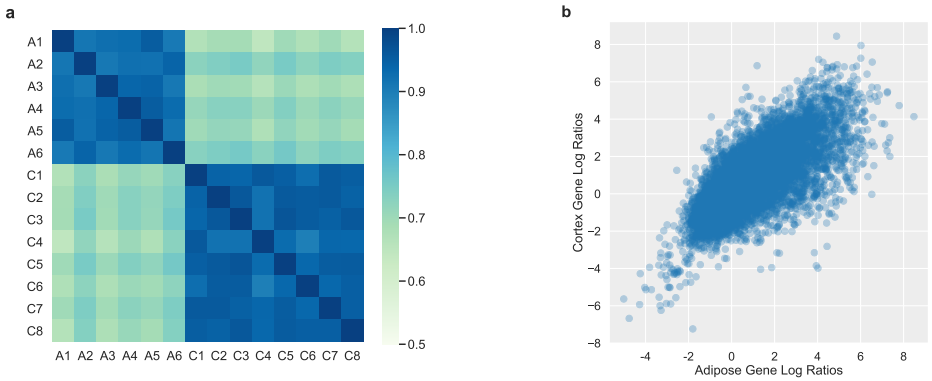


Supplementary Figure 1: Decomposition of human subcutaneous adipose tissue.
a UMAP projection of snRNA-seq data with 5 identified cell type clusters labeled.
b Leave-one-out cross-validation using 6 samples with snRNA-seq and bulk RNA-seq data available. Proportions based on snRNA-seq were used as a proxy for the true proportions on the x-axis. Estimated proportions for an individual were generated by each decomposition method using the remaining 5 individuals as training data. Each color represents one of the 5 identified cell populations.
c Leave-one-out cross-validation performance after normalization of estimates within each cell type to determine cell-specific accuracy. Normalized estimates are robust to inflation of global Pearson correlation by large cell populations; however, these metrics are noisy when considering only six individuals.
d Leave-one-out cross-validation performance on exon-aligned snRNA-seq data. Existing methods are able to detect additional cell populations using the exonic subset of the snRNA-seq data, though around 40% of the sequenced cells are filtered out.
Source data are provided as a Source Data file.

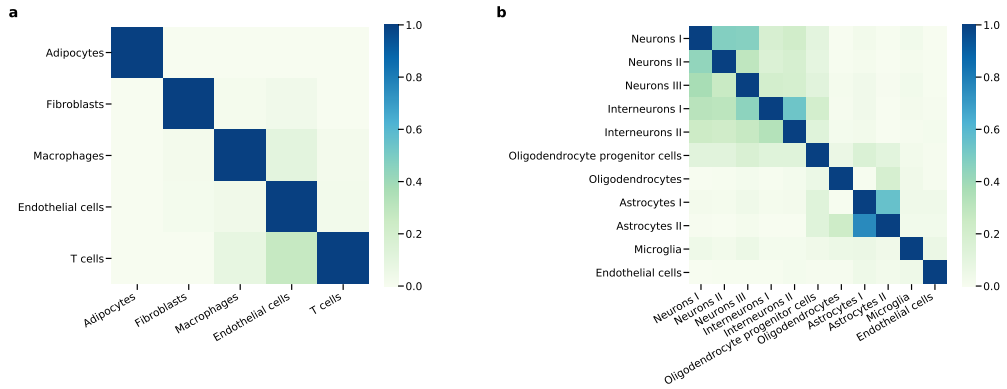

a
UMAP 2
UMAP 1
Oligodendrocytes
Neurons I
Astrocytes I
Neurons II
Interneurons I
Interneurons II
Microglia
Oligodendrocyte progenitor cells
Endothelial cells
Astrocytes II
Neurons III
b
Bisque
R: 0.924 ± 0.062
RMSD: 0.029 ± 0.010
MuSiC
R: -0.192 ± 0.107
RMSD: 0.173 ± 0.013
BSEQ-sc
R: 0.098 ± 0.216
RMSD: 0.120 ± 0.023
CIBERSORT
R: -0.281 ± 0.049
RMSD: 0.197 ± 0.012
CIBERSORTx
R: 0.671 ± 0.153
RMSD: 0.070 ± 0.019
Estimated Proportion
True Proportion
c
Bisque
R: -0.153 ± 0.480
RMSD: 0.129 ± 0.073
MuSiC
R: 0.132 ± 0.364
RMSD: 0.263 ± 0.127
BSEQ-sc
R: 0.222 ± 0.216
RMSD: 0.226 ± 0.106
CIBERSORT
R: 0.266 ± 0.278
RMSD: 0.234 ± 0.114
CIBERSORTx
R: 0.095 ± 0.358
RMSD: 0.169 ± 0.050
Estimated Proportion
True Proportion



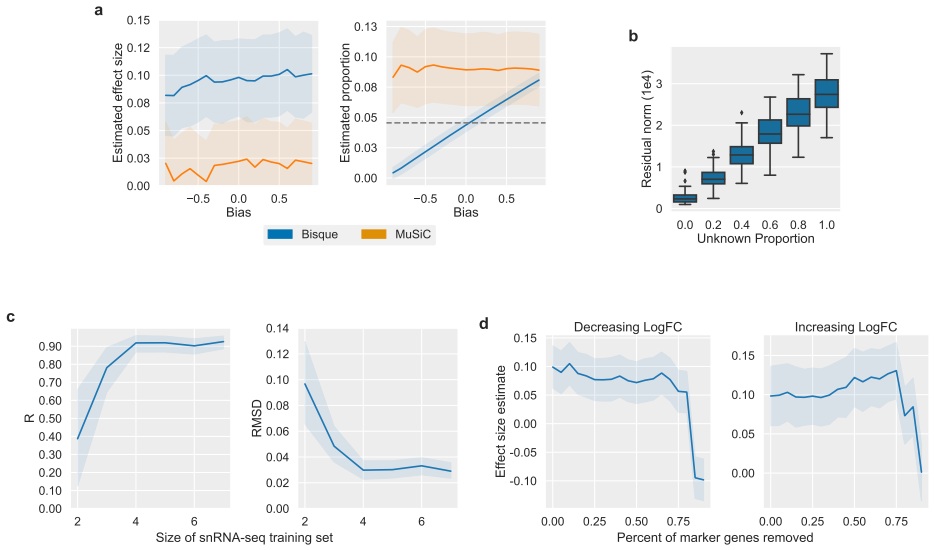
Supplementary Figure 2: Decomposition of human DLPFC tissue.
a UMAP projection of snRNA-seq data with 11 identified cell type clusters labeled.
b Leave-one-out cross validation using 8 samples with snRNA-seq and bulk RNA-seq data available. Proportions based on snRNA-seq were used as a proxy for the true proportions on the x-axis. Estimated proportions for an individual were generated by each decomposition method using the remaining 7 individuals as training data. Each color represents one of the 11 identified cell populations.
c Leave-one-out cross-validation performance after normalization of estimates within each cell type to determine cell-specific accuracy. As described previously, performance metrics on normalized data provide better measure of global accuracy but are noisy with small sample sizes.
d Leave-one-out cross-validation performance after merging closely related cell subtypes into 6 clusters. Performance of existing methods increases compared to decomposition into 11 clusters with related subtypes.
e Decomposition of remaining 628 individuals with cell subtype merging. The aggregated cell type proportions estimated from the 8 snRNA-seq samples are similar to IHC estimates for neurons and astrocytes from 70 individuals in the cohort (data not shown). Boxes indicate the quartiles of the estimated proportions with whiskers extending 1.5 times the interquartile range. Points are individual samples that are represented by the boxplot.
Source data are provided as a Source Data file.



Supplementary Figure 3: Consistency of snRNA-seq to bulk RNA-seq expression log-ratios across individuals, tissues, and experiments.
a Heatmap depicting Pearson correlation between pairs of individual's log-ratios of snRNA-seq expression to bulk RNA-seq gene expression measured in counts per million (CPM). A sample prefix of 'A' indicates an individual from the adipose dataset and 'C' indicates an individual from the cortex dataset. Correlation is high between individuals within experiments as well as between experiments/tissues, indicating the same genes are over/under-expressed in snRNA-seq when compared to bulk RNA-seq.
b Scatterplot of average snRNA-seq to bulk RNA-seq gene expression log-ratios across individuals in adipose dataset (x-axis) and cortex dataset (y-axis). Each point corresponds to a gene detected in both experiments, depicting the average ratio across all individuals for that tissue. The snRNA-seq to bulk RNA-seq ratios vary across genes and correlate (R=0.747) between these two experiments.
Source data are provided as a Source Data file.



Supplementary Figure 4: Shared marker genes between identified clusters in snRNA-seq data. Heatmaps of proportion of shared marker genes where an entry indicates the proportion of marker genes for the cluster on the x-axis that are found in the cluster on the y-axis.
a The 5 clusters identified in adipose tissue are relatively distinct in their marker genes.
b The 11 clusters identified in DLPFC tissue have several closely related subtypes, such as neurons and astrocytes.
Source data are provided as a Source Data file.



Supplementary Figure 5: Robustness of the reference-based decomposition model.
a Microglia cells in the 8 DLPFC snRNA-seq samples were upsampled or downsampled at various percentages, denoted as bias on the x-axis, to simulate reference data that may overrepresent or underrepresent a cell type of interest compared to bulk data. Decomposition performance, measured as the estimated effect size of microglia proportion on Braak stage (which is expected to be positive) on the y-axis was consistent for each method as the bias in the snRNA-seq reference varied (left). Effect sizes were fit in a linear regression model on 628 samples adjusting for age at death, age at death squared, and sex. Since Bisque utilizes proportions observed in the reference data under the assumption that they reflect physiological compositions, the simulated bias propagates to the estimated proportions (right). Shaded regions indicate standard error of estimates. On the left, the line indicates the estimated effect size of the linear model. On the right, the line indicates the mean cell proportion estimate.
b The DLPFC bulk RNA-seq data was mixed with the adipose bulk RNA-seq data at various proportions to simulate an unknown cell population in the bulk data that is not represented in the snRNA-seq reference data. In order to model the severity of the sample discordance, we compared the amount of adipose contamination, denoted as unknown proportion on the x-axis, to the residuals from the Bisque model (y-axis). As this simulated unknown proportion increases, the residuals of Bisque also increase. Each boxplot represents a distinct random subset of 100 samples from the DLPFC dataset that were mixed with 100 randomly selected samples from the adipose data. Boxes indicate the quartiles of the observed residual norms with whiskers extending 1.5 times the interquartile range. Points are outliers beyond this range.
c Leave-one-out cross-validation performance across the 8 samples in the DLPFC dataset after utilizing random subsamples of the snRNA-seq data as a reference. Performance, in terms of Pearson correlation (left) and RMSD (right), began to drop when using less than 4 individuals in the reference dataset. Shaded regions indicate 95% confidence interval with a line indicating the mean observed value.
d An increasing number of marker genes for the microglia cells in the DLPFC dataset were removed to determine the effect on decomposition performance. At each amount of genes removed (x-axis), performance was measured as the effect size of the estimated microglia proportion on Braak stage (y-axis). Effect sizes were fit in a linear regression model on 628 samples adjusting for age at death, age at death squared, and sex. Genes were removed in order of decreasing (left) or increasing (right) log-fold-change. In both settings, performance remained relatively consistent until around 75% of the 102 identified marker genes were removed. Shaded regions indicate standard error of estimated effects with a line indicating the actual estimated effect.
Source data are provided as a Source Data file.

CHAPTER 4

Human liver single nucleus and single cell RNA sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival

Alvarez *et al. Genome Medicine* (2022) 14:50
https://doi.org/10.1186/s13073-022-01055-5

Genome Medicine

RESEARCH Open Access

Human liver single nucleus and single cell RNA sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival

Marcus Alvarez[1†], Jihane N. Benhammou[2,3†], Nicholas Darci-Maher[1], Samuel W. French[4], Steven B. Han[5], Janet S. Sinsheimer[1,6,7], Vatche G. Agopian[8], Joseph R. Pisegna[1,3] and Päivi Pajukanta[1,7,9*]

Abstract

Background: Hepatocellular carcinoma (HCC) is a common primary liver cancer with poor overall survival. We hypothesized that there are HCC-associated cell-types that impact patient survival.

Methods: We combined liver single nucleus (snRNA-seq), single cell (scRNA-seq), and bulk RNA-sequencing (RNA-seq) data to search for cell-type differences in HCC. To first identify cell-types in HCC, adjacent non-tumor tissue, and normal liver, we integrated single-cell level data from a healthy liver cohort ($n = 9$ non-HCC samples) collected in the Strasbourg University Hospital; an HCC cohort ($n = 1$ non-HCC, $n = 14$ HCC-tumor, and $n = 14$ adjacent non-tumor samples) collected in the Singapore General Hospital and National University; and another HCC cohort ($n = 3$ HCC-tumor and $n = 3$ adjacent non-tumor samples) collected in the Dumont-UCLA Liver Cancer Center. We then leveraged these single cell level data to decompose the cell-types in liver bulk RNA-seq data from HCC patients' tumor ($n = 361$) and adjacent non-tumor tissue ($n = 49$) from the Cancer Genome Atlas (TCGA) multi-center cohort. For replication, we decomposed 221 HCC and 209 adjacent non-tumor liver microarray samples from the Liver Cancer Institute (LCI) cohort collected by the Liver Cancer Institute and Zhongshan Hospital of Fudan University.

Results: We discovered a tumor-associated proliferative cell-type, Prol (80.4% tumor cells), enriched for cell cycle and mitosis genes. In the liver bulk tissue from the TCGA cohort, the proportion of the Prol cell-type is significantly increased in HCC and associates with a worse overall survival. Independently from our decomposition analysis, we reciprocally show that Prol nuclei/cells significantly over-express both tumor-elevated and survival-decreasing genes obtained from the bulk tissue. Our replication analysis in the LCI cohort confirmed that an increased estimated proportion of the Prol cell-type in HCC is a significant marker for a shorter overall survival. Finally, we show that somatic mutations in the tumor suppressor genes *TP53* and *RB1* are linked to an increase of the Prol cell-type in HCC.

†Marcus Alvarez and Jihane N Benhammou are equally contributed to this work.

*Correspondence: ppajukanta@mednet.ucla.edu

[9] Institute for Precision Health, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
Full list of author information is available at the end of the article

Conclusions: By integrating liver single cell, single nucleus, and bulk expression data from multiple cohorts we identified a proliferating cell-type (Prol) enriched in HCC tumors, associated with a decreased overall survival, and linked to *TP53* and *RB1* somatic mutations.

Background

Hepatocellular carcinoma (HCC) is the third leading cause of cancer-related death world-wide [1]. Although early detection has been associated with improved overall survival [2], most patients present in later stages, which prevents curative therapies such as hepatic resection and liver transplantation, resulting in a 5-year survival of only 18% [3]. Previous studies have demonstrated that tumor heterogeneity is common in HCC [4], which may explain some of the differences in survival outcomes and responses to therapies [5, 6]. Sub-classification of HCCs by molecular and cellular characteristics could help guide biomarker discovery and treatment options, especially in NAFLD-related HCCs, which remain poorly understood and underrepresented in most transcriptomic HCC studies.

Single-cell RNA sequencing (scRNA-seq) has advanced the study of complex admixtures of cells, shedding light on cellular functions at the single cell level in unprecedented ways [7–10]. However, applying scRNA-seq technology to precious, archived human tissues, such as liver biopsies or resections, has proven to be challenging as it is not possible to dissociate intact cells from these existing biopsies of solid tissues. Single nucleus RNA sequencing (snRNA-seq) techniques [11] have overcome these technical challenges [12] and enabled cell-type level characterization of frozen solid tissues [13–16]. As scRNA-seq and snRNA-seq technologies improve, their use for solid tissues, such as liver, has expanded [17, 18]. However, studies integrating data from multiple single cell level cohorts are needed to improve power of small individual cohorts.

In the field of tumor biology, scRNA-seq and snRNA-seq have helped elucidate the presence of tumor heterogeneity, which is commonly observed at the molecular and clinical level in HCC [19–21]. ScRNA-seq and snRNA-seq have provided ways to further identify and characterize cell-types at finer resolutions [14–18, 21], which was not possible using bulk RNA-seq. In addition, many scRNA-seq studies have investigated tumor microenvironment by immune cells as this has been shown to be an important target in HCC treatment in the era of immunotherapy, with potential prognostic utilities [22, 23]. The importance of understanding tumor heterogeneity is further illustrated by the clinical observation that NAFLD-related HCC cases may be more resistant to new systemic immunotherapies [24], as shown at the molecular level both in human studies and murine models [22]. Thus, given the changing landscape of HCC etiologies and the observed clinical heterogeneity, additional cell-type level transcriptomics studies of HCC are warranted.

We hypothesized that snRNA-seq can complement the existing scRNA and bulk expression data from liver HCC and normal liver cohorts and that these data can be integrated to identify currently unknown HCC-associated cell-types that affect survival when their proportions expand in the tumor tissue. To this end, we first used a liver snRNA-seq data set that we previously generated from HCC tumor and adjacent non-tumor liver biopsies from patients with NAFLD-related HCC [25], and then integrated these data with two existing liver scRNA-seq data sets, representing multiple etiologies of HCC and normal liver [7, 8]. Thus, we generated a powerful reference data set, comprising both viral and non-viral origin HCC, adjacent non-tumor, and normal liver samples at the single cell resolution. We then leveraged the cell-type marker genes identified in these three reference data sets to decompose cell-type proportions in liver bulk RNA-seq data from the well-phenotyped Cancer Genome Atlas (TCGA) cohort [26] (361 HCC tumor and 49 adjacent non-tumor biopsies) to first accurately estimate the tumor/non-tumor cell-type proportions and then test the effects of the identified HCC-enriched cell-types on survival outcomes. To replicate and further validate the results, we used the Liver Cancer Institute (LCI) cohort [27] (221 HCCs and 209 adjacent non-tumor tissue biopsies), collected by the Liver Cancer Institute and Zhongshan Hospital of Fudan University, which consists predominantly of chronic hepatitis B-HCCs. Using these two independent HCC cohorts, we discovered a replicated, proliferative cell-type, Prol, characterized by 656 mitosis and cell-cycle enriched cell-type marker genes, that is significantly more present in the HCC cases than in adjacent non-tumor liver tissue both in TCGA and LCI, in line with our single cell level data. Previous studies have not identified HCC cell-types associated with survival. Thus, our discovery that HCCs with a high Prol cell-type content have significantly worse survival outcomes advances the field by elucidating a key HCC risk cell-type. Importantly, we observed this same result both in TCGA and LCI, which increases the scientific rigor of our finding.

Multiple cancer genes and mutations have been identified in HCC, including mutations in tumor suppressors, such as tumor protein P53 (*TP53*) [6]. However, it is not known whether these somatic mutations are also associated with cell-type changes in HCC. To address this knowledge gap and elucidate the molecular mechanisms of the identified cell-types, we investigated the HCC risk cell-type, Prol, for accumulation of known somatic cancer mutations [6, 28]. Using somatic mutation of origin analysis, we discovered that somatic *TP53* and *RB1* mutations are linked to the identified increase of Prol in HCC.

Methods

Study design

To identify cell-types associated with HCC and its survival outcomes, we first analyzed three liver single cell level data sets from an existing snRNA-seq cohort of NAFLD-related HCC [25], an existing scRNA-seq cohort of HCC from various etiologies [8], and a healthy liver scRNA-seq cohort [7] to identify and characterize their cell-types. Next, we leveraged these liver cell-type reference data to decompose cell-type proportions in the liver bulk RNA-seq data from the Cancer Genome Atlas (TCGA) cohort [26] and subsequently tested the estimated cell-type proportions for associations with HCC and survival outcomes. Then, the HCC and survival associated cell-types identified in TCGA were tested for replication in independent liver bulk microarray expression data from the previously published LCI cohort [27]. Finally, we searched for associations between cell-type proportions and somatic mutations in the TCGA cohort.

snRNA-seq cohort

We identified NAFLD-related HCC cases among patients undergoing surgical resection for HCC treatment at the Dumont-UCLA Liver Cancer Center [25]. The 3 patients with NAFLD-related HCC were women with a mean age of 77.9 ± 3.1 years and a mean body mass index of $25.3 \pm 2.9\,kg/m^2$, who had components of the metabolic syndrome (hypertension, dyslipidemia and insulin resistance). All patients exhibited features of nonalcoholic steatohepatitis (NASH) on liver histopathology (steatosis, ballooning and lobular inflammation [29]), and none had cirrhosis, as assessed by the METAVIR fibrosis score [30] (Additional file 1: Fig. S1). All patients also presented with clinically heterogeneous tumors, based on sizes, histological stages of differentiation (moderate to poorly differentiated), and serum alpha fetoprotein (AFP) levels, with one patient exhibiting an AFP of >400 ng/mL.

Tissues were characterized by a pathologist using H&E and immunohistochemical stains, which confirmed the diagnoses of HCC ($n=3$) and adjacent non-tumor ($n=3$). Samples were snap frozen and kept at -80^0C until extraction of the nuclei. All histopathology slides were reviewed by the same pathologist. We abstracted clinical data and other demographics from the electronic health records. The study was approved by the UCLA IRB, and all participants provided a written informed consent.

Two existing scRNA-seq cohorts

Along with the snRNA-seq data [25], we also incorporated liver scRNA-seq data from two previously published cohorts into our single cell level analysis [1]: HCC patients with viral origin of HCC ($n=4$), HCC patients with unspecific origin of HCC ($n=10$), and adjacent control liver samples ($n=14$), as well as a healthy normal donor, collected in the Singapore General Hospital and National University Hospital [8], and [2] normal liver samples ($n=9$), collected in the Strasbourg University Hospital [7]. Data from Sharma et al. [8] were downloaded from https://data.mendeley.com/datasets/6wmzcskt6k/1. Read counts for filtered droplets ($n=73{,}589$) from the 14 HCC patients and 1 control were extracted from the downloaded HCC.h5ad file. Read counts for the 9 normal liver samples from Aizarani et al. [7] were downloaded from GEO under the accession number GSE124395. We used the filtered set of droplets provided by the authors ($n=10{,}372$) for analysis.

Processing of The Cancer Genome Atlas (TCGA) bulk RNA-seq, mutation, and clinical data

To expand our cell-type composition analysis to a larger number of HCC samples, we leveraged data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) (referenced as TCGA in the text) [26]. The TCGA-LIHC cohort includes 361 cases with primary tumors. We included only those cases that were designated as non-recurrent primary HCC and excluded cholangiocarcinomas, HCC-cholangiocarcinoma mixed tumors, and other rarer types of HCC, such as fibrolamellar, as these have different pathogenesis and clinical outcomes. We integrated bulk RNA-seq, mutation, clinical, and survival data with our single cell level RNA-seq data to identify HCC-associated cell-types.

Clinical data were downloaded from Genomics Data Commons (GDC) portal [31] (https://portal.gdc.cancer.gov/projects/TCGA-LIHC). We abstracted the available clinical and biospecimen data from Genomic Common Data portal, which included age, sex, ethnicity, and HCC tumor size, as well as node and metastatic American Joint Committee on Cancer (AJCC) TNM staging, and RNA integration number (RIN). Some other clinical characteristics were missing in TCGA, and thus, we had no data on cirrhosis status, the Model for End-Stage Liver Disease (MELD), serum AFP levels, or additional

clinical phenotypes (e.g., diabetes and medication). Endpoint data for the survival analysis were downloaded from Table S1 from Liu et al. [32], and redacted cases were removed before analysis.

Liver bulk RNA-seq expression data were downloaded from the GDC portal [31] as HTSeq counts for all TCGA-LIHC individuals. We included counts for the 361 primary tumor samples, as well as for 49 matched non-tumor samples. For downstream analysis, the counts were Trimmed Mean of M-values (TMM) normalized with edgeR [33] and log10 transformed after adding a prior count of 1. Finally, RIN was regressed out to obtain the final normalized expression data.

Somatic mutation data collected from whole exome sequencing of tumor biopsies for the TCGA-LIHC were downloaded from the Broad Genome Data Analysis Center (GDAC) (http://gdac.broadinstitute.org). The Analysis Results file from the MutSig2CV under Mutation Analyses were downloaded on May 18, 2021. These included a MAF file of somatic mutations for each sample (LIHC-TP.final_analysis_set.maf), as well as a list of 69 significantly frequently mutated HCC genes ($q < 0.1$) (sig_genes.txt).

The Liver Cancer Institute (LCI) cohort used for replication analyses

To validate the results obtained in TCGA, we analyzed a previously published HCC microarray dataset [27, 34]. This study recruited the HCC patients from the Liver Cancer Institute (LCI) and Zhongshan Hospital of Fudan University, most of whom had a history of chronic hepatitis B (HBV) infection. We obtained tumor microarray expression, clinical, and overall survival (OS) outcome data for a total of 221 patients. Additionally, 209 of these patients had expression data for adjacent non-tumor liver biopsies. RMA-normalized microarray expression data in log space were directly downloaded from GSE14520 in GEO. The clinical data, including OS endpoints, were downloaded as the extra endpoint text file from GSE14520. The expression data had been normalized by the authors [27], and thus, we used them directly for downstream analysis.

Liver single nucleus extraction for snRNA-seq

For the snRNA-seq of the 3 NAFLD-related HCC and 3 adjacent non-tumor control biopsies, we cut the frozen samples over dry ice and placed them in glass tubes, as described earlier [25]. Briefly, we added 4 ml of lysis buffer consisting of 0.1% IGEPAL, 10 mM Tris-HCl, 10 mM NaCl, and 3 mM MgCl2 to the tissue. After 10 min on ice, we mechanically homogenized the tissue using a Dounce homogenizer, and then filtered them through a 70-µm MACS smart strainer (Miltenyi Biotec #130-098-462) to remove debris. We isolated the nuclei by spinning the homogenate at 500 x g for 5 minutes at 4 °C and washed the nuclei in 1 ml of resuspension buffer (RSB) consisting of 1X PBS, 1.0% BSA, and 0.2 U/µl RNase inhibitor. We filtered the nuclei a second time using 40 µm Flowmi cell strainer (Sigma Aldrich # BAH136800040) and centrifuged them at $500 \times g$ for 5 min at 4 °C. We resuspended the nuclei in the wash buffer and kept them on ice. To assess nuclei isolation (for clumping and intact membrane), we labeled the nuclei with Hoechst stain and counted them using BZ-X710 fluorescent microscope. Nuclei were immediately processed them with the 10X Chromium platform following the Single Cell 3′ v2 protocol. We generated libraries with the 10X platform and sequenced the nuclei on an Illumina NovaSeq S2 at a sequencing depth of 300–400 million reads per sample.

Processing of the snRNA-seq data

Before read alignment, we trimmed template switch oligos, primers, and polyA sequences greater than 20 base pairs from the fastq reads using cutadapt (https://cutadapt.readthedocs.io/en/stable/). We aligned reads to the GRCh38 human genome reference and Gencode v26 [35] gene annotations using STARSolo in STAR v2.7.3a [36]. Gene counts were taken from the full pre-mRNA transcript using the "—soloFeatures GeneFull" option. We filtered empty and contaminated droplets using Debris Identification using Expectation Maximization (DIEM) [13], where we further adapted estimation of the multinomial mixture model parameters by adding a prior count of 1 to the gene mean estimates and the cluster membership estimates to avoid overfitting. To further remove doublets and contaminated clusters from the snRNA-seq data, we separately clustered parenchymal hepatocytes and non-parenchymal nuclei. Nuclei were clustered in a first pass and assigned to hepatocyte and non-hepatocyte cell-types. Each group was clustered again separately. Then, we removed nuclei belonging to clusters expressing markers from multiple cell-types, leaving the filtered set of nuclei ($n = 39{,}995$).

Integration and clustering of the single cell level data from the three cohorts

To analyze the single-cell level data across the cohorts, we first removed cohort- and experiment-specific effects by performing data integration. Counts were first normalized using sctransform [37] and integrated using canonical correlation analysis (CCA) [38]. Integrations were performed across the 6 NASH-HCC samples, the 15 patients (14 HCC and 1 healthy control) in the Sharma [8] data set, and the single combined set of 9 samples in the Aizarani [7] data set. The 22 samples across the 3 cohorts were used for independent samples

during normalization and integration. Each of the 22 samples were normalized with sctransform using 3000 genes for the number of variable features. To reduce the time required for integration, we selected a subset of the 22 samples for use as a reference during the FindIntegrationAnchors step. We selected 11 samples, including the combined sample from Aizarani et al. [7] to serve as a healthy control, and 10 additional randomly selected samples. After anchor identification, all 22 samples from the 3 cohorts were integrated with the IntegrateData function in Seurat [38] using 30 dimensions. This resulted in corrected counts for the 123,956 droplets. Finally, we performed clustering on these corrected counts for downstream cell-type assignment. We ran principal component analysis (PCA) and constructed the shared nearest neighbor (SNN) graph with 30 PCs. This graph was used as the input to Louvain clustering by running the FindClusters function with a resolution of 1 [38]. We chose a resolution of 1 to accommodate the large number of cells and nuclei and better identify subtypes. To evaluate the effect of integration, we also clustered cells and nuclei in the three cohorts by clustering the merged data without CCA integration. Sctransform was run on the merged counts as described above, treating the cells and nuclei from the three cohorts as a single sample. PCA and clustering were performed on the sctransformed counts in the same manner as for the integrated data.

Marker gene identification and cell-type assignment of single cell level data

For cell-type classification, we obtained the upregulated marker genes and log-fold changes for each cluster using the uncorrected, log-normalized counts. Raw counts for all droplets were multiplied by a scaling factor to sum to 1,000 as this was the approximate median across all droplets, and then log-transformed. To identify marker genes, we performed a logistic regression test using the FindAllMarkers function in Seurat [38] and kept marker genes with an average $\log_2$ fold change of at least 0.1 and Bonferroni-adjusted p-value < 0.05 corrected for the total number of genes in the data set. For the pathway enrichment analysis, we also obtained the log fold changes for all expressed genes by calculating the difference in $\log_2$ means between the counts of droplets classified within and outside of the cluster. Cell-types were assigned based on manual curation of known marker genes [26]. Throughout the manuscript, we call the 25 assigned clusters the subcell-types. We further merged the subcell-types into the 8 main cell-types based on their common lineage, expressed genes, and enriched pathways.

Pathway enrichment analyses of the single cell level data

To gain insight into cell-type functions in the liver single cell level data, we performed pathway enrichment analysis of upregulated marker genes for each liver subcell-type. We used the clusterProfiler [39] R package to run gene set enrichment analysis (GSEA) [40]. We tested for enrichments of the pathways in the Reactome database [41, 42]. For each subcell-type, its log fold changes were used to rank the gene set as input to the gsePathway function, using 10,000 permutations and an epsilon of 1×10^{-50}. p-values were corrected for multiple testing using FDR.

Clustering of Prol cells and nuclei

The Prol cluster that we identified in the integrated analysis expressed markers involved in cell division; however, our integrated analysis did not further separate these cells/nuclei, so we subclustered the 1,743 Prol cells/nuclei to identify its composition. We ran a clustering pipeline similar to the whole data set, with modifications to account for the lower number of cells. The Prol cells/nuclei were first split by cohort, and sctransform was run on the raw counts for each of the three samples. We then ran CCA integration with the k.filter and k.weight parameters set to 75 to account for the small number of cells/nuclei in each cohort, as only 92 Prol cells were present in the healthy liver tissues from the Aizarani data set [7]. Cluster assignments and UMAPs were generated using 30 PCs with a resolution of 0.2 to accommodate the lower number of cells/nuclei and to match clusters with the main cell-types.

To assign Prol cells/nuclei to the main liver cell-types, we used SingleR [43]. For classification, we first generated a reference of the 7 main cell-types (excluding the Prol cluster) from the integrated liver data. Briefly, pairwise T-tests were performed across the 7 main cell-types and the top 100 markers were extracted. A reference was derived on the log-normalized counts using these top markers with the trainSingleR function. To account for the single-cell level nature of the reference, the counts were aggregated to pseudobulk samples with the aggr.ref parameter. Finally, we ran the classifySingleR function on the droplets in the Prol cluster and assigned their cell-type to the pruned labels.

Estimating cell-type proportions and correlation analyses of the cell-type marker genes in the liver bulk RNA-seq from TCGA and microarray data from LCI

To estimate cell-type proportions in the bulk liver expression data in the TCGA-LIHC cohort [26], we used a co-expression based approach implemented in Bisque [14]. Briefly, this approach performs PCA on the top cell-type

marker genes for each cell-type. We used normalized RNA-seq expression and cell-type markers as input, requiring a minimum of 20 genes and a maximum of 300 genes for the set of markers for PCA. The marker genes were obtained from our single cell level reference data. Decomposition was performed for the 8 main cell-types and 25 subcell-types. As we observed high correlation ($R > 0.9$) between proportion estimates of subcell-types within their main classification, we used only the proportion estimates for the main cell-types for downstream analysis.

In order to replicate our results with the decomposed proportion estimates observed in TCGA, we ran the same decomposition in the LCI cohort. We ran Bisque on the normalized microarray expression data using the same parameters and marker gene input described above. To assess the reliability of the TCGA and LCI proportion estimates, we analyzed the co-expression patterns of the marker genes in each cohort. We found that for the LCI cohort, the B cell marker genes did not show positive correlations across their expression. As our decomposition approach relies on co-expression of marker genes, we excluded B cells from the LCI main cell-type proportion estimates.

Cell-type proportion differences between tumor and non-tumor in the single cell level and bulk data

To identify tumor-enriched or depleted cell-types, we performed paired Wilcoxon signed-rank tests between tumor and non-tumor samples. In the single-cell-level data, we calculated differences in the observed proportions between paired tumor and non-tumor samples in the 17 patients with matched biopsies. Differences were calculated for each subcell-type. The observed subcell-type proportions for each tumor or non-tumor sample were calculated by dividing the number of cells/nuclei in the subcell-type by the total number in the sample. For the tumor samples in the Sharma data set [8], the core and peripheral tumor droplets were combined. *p*-values were corrected for testing 25 subcell-types using FDR.

For calculating differences in cell-type proportion estimates between tumor and adjacent non-tumor samples in the TCGA and LCI bulk tissue cohorts, we performed a paired Wilcoxon test in TCGA ($n = 49$) and LCI ($n = 209$). *p*-values were corrected for testing 8 and 7 cell-types in the TCGA and LCI cohorts, respectively, using FDR.

Survival outcome associations with cell-type proportion estimates

To investigate the effect of cell-types on survival outcomes, we performed associations between survival outcomes and cell-type proportion estimates. Associations were carried out with Cox proportional hazard regressions for overall survival (OS) and progression free interval (PFI) in TCGA, and OS in the LCI validation cohort. We included age, sex, and ethnicity as covariates in TCGA, and age and sex in the LCI cohort, as most patients from this cohort were of Asian descent. In addition, we included tumor stage as a binary covariate where specified, where patients with stage I and II were grouped into the low group and those with stage III and IV were grouped into the high group. Patients with any missing covariate data were excluded. All *p*-values were corrected for multiple testing using false discovery rate (FDR). All survival analyses were performed with the survival package in R [44]. We tested survival differences between low vs. high proportion groups, splitting the participants by median or quartile. In the median analysis, tumor samples with proportion estimates below and above the median were grouped into low and high, respectively. Similarly, the quartile analysis was performed using the lower and upper 25% quartiles of the cell-type proportion estimates as cutoffs. Plots were generated using the Kaplan-Meier method without any covariates. Unless otherwise specified, all cell-type effects were corrected for testing 8 and 7 cell-types in the TCGA and LCI cohorts, respectively, using FDR.

Mutation analyses in TCGA-LIHC

We hypothesized that mutations in distinct genes would lead to increased Prol proportions in HCC tumor samples. We thus tested for differences in proportions between tumor samples with and without a somatic mutation in TCGA-LIHC, as LCI did not profile tumor mutations. Somatic mutations in TCGA-LIHC were collected from exome sequencing data processed by GDAC (http://gdac.broadinstitute.org). We restricted our analysis to 69 genes frequently and significantly mutated in HCC, as reported previously in the TCGA-LIHC cohort (http://gdac.broadinstitute.org) [45]. A gene was considered significantly mutated if its *q*-value was less than 0.1, as determined by MutSig2CV [46]. Tumor samples with at least one synonymous, nonsense, in frame, splice site, missense, or frame shift variant were considered as having a somatic mutation (mut.). Tumor samples without any somatic mutation detected were considered as wildtype (WT). For each gene and each main cell-type, we used a Wilcoxon test to assess the difference in cell-type proportion estimates between tumor samples with a somatic mutation detected and tumor samples without a somatic mutation. For *TP53*, we also tested for tumor proportion differences between wildtype (WT) cases and each of the somatic mutation types listed previously. Wilcoxon *p*-values were adjusted for multiple testing across all gene-main-cell-type pairs using FDR.

Bulk liver differential expression (DE) analyses

In addition to estimating cell-type proportions, we also reciprocally evaluated the significance of cell-types in HCC by assessing single cell expression of genome-wide significant bulk tumor-elevated, survival decreasing, and mutation-elevated genes. To first obtain the genome-wide tumor-elevated genes in TCGA and LCI, we ran differential expression (DE) genome-wide in both the TCGA and LCI bulk expression cohorts. DE was run on the 49 and 209 paired samples in the TCGA and LCI, respectively, that contained the matched tumor and adjacent non-tumor samples. For the TCGA RNA-seq data, we first filtered for expressed genes by removing those with an average number of reads less than 10 across the 98 samples. We then ran edgeR [33] on the TMM normalized counts with the generalized linear model (GLM) framework (glmFit and glmLRT functions) and setting a prior count of 1. For the LCI microarray data, we used the gene-filtered and normalized data provided. We then ran limma [47] to fit a linear model (lmFit function) and compute test statistics with empirical Bayes shrinkage of variances (eBayes function). For both the TCGA and LCI, we accounted for the paired status of the samples by including the patient as an indicator covariate.

Next, to obtain the genome-wide survival-decreasing genes in the bulk expression cohorts, i.e., OS- and PFI-decreasing genes in TCGA and OS-decreasing genes in LCI, we performed Cox proportional hazard regressions for OS and PFI in TCGA, and OS in the LCI validation cohort. As with the proportion analyses, we included age, sex, and ethnicity as covariates in TCGA, and age and sex in the LCI cohort. Patients with any missing covariate data were excluded. We then ran Cox proportional hazards regression testing normalized gene expression values as a quantitative predictor against survival outcomes. The statistical significance of these survival-decreasing genes was corrected for genome-wide testing using FDR. Regressions were performed with the survival package in R [44].

Similarly, to identify genes upregulated in the context of a somatic mutation in *TP53* and *RB1*, we also performed genome-wide DE between somatic mutation (mut.) and wildtype (WT) carriers in the TCGA cohort. We broadly included genes with greater than 0 counts in at least 50% of the 410 samples. To test for DE, we ran the GLM framework in edgeR [33] using TMM normalization. DE was run on the 357 primary tumor samples with both mutation and RNA-seq data. We tested for differences in bulk liver expression between participants that were wildtype (WT) and those that had a somatic mutation (mut.) in the particular gene. A genome-wide DE analysis was performed for both *TP53* and *RB1*.

Scoring of the cell-cycle, tumor-elevated, OS- and PFI-decreasing, and mutation upregulated bulk genes in the single cell level data

To assess cell/nuclei expression of the cell-cycle genes [48] (42 S phase genes and 54 G2 and M phase genes) as well as the tumor-elevated, OS- and PFI-decreasing, and mutation upregulated genes identified in our bulk DE analyses (see above), we assigned module scores with the AddModuleScore function implemented in the Seurat package [38]. Briefly, module scores are derived by calculating the average expression of the gene set and subtracting the average expression of gene sets. Control gene sets are randomly selected from bins based on average expression. The expression data of cells/nuclei for module scoring were calculated by multiplying raw read counts to sum to 1,000 and log transforming them.

For cell cycle scoring, the gene sets included 42 S phase genes and 54 G2 and M phase genes provided in the Seurat package [38, 48]. For tumor-elevated scores, we used the genes identified in the bulk liver DE analysis that had a log fold change greater than 1 of tumor over non-tumor and an FDR-corrected p-value < 0.05. The tumor-elevated gene set included 1065 genes in TCGA and 335 genes in LCI. For the OS- and PFI-decreasing gene sets, we analyzed the genes identified in the genome-wide survival analysis of the bulk liver data that had a hazard ratio > 1 (increased expression leading to a worse prognosis) and an FDR-corrected p-value < 0.05. There were 740 OS-decreasing genes and 528 PFI-decreasing genes in TCGA and 36 OS-decreasing genes in LCI. For the mutation upregulated genes, we included those from the genome-wide DE mutation analysis for *TP53* and *RB1* that had a log fold change greater than 0.5 and an FDR adjusted p-value < 0.05. This resulted in a set of 1358 *TP53* mut. upregulated genes and a set of 774 *RB1* mut. upregulated genes.

Differences in tumor-elevated, OS-decreasing, PFI-decreasing, and mutation upregulated gene scores between the Prol and all other clusters were assessed by running a Wilcoxon test between droplet scores within and outside of the Prol cluster.

Results

Overview of study design

HCCs are poorly characterized at the cell-type level. To address this scientific and biomedical knowledge gap, we utilized the following three single cell level RNA-seq data sets to produce a comprehensive cell-type reference for HCC tumor, adjacent non-tumor tissue, and normal livers [1]: liver snRNA-seq data that we previously generated from HCC samples ($n = 3$) and adjacent non-tumor control tissue samples ($n = 3$) from patients with NAFLD-related HCC undergoing hepatic resection [25]

[2]; existing liver scRNA-seq data from HCC patients with viral origin of HCC ($n=4$), HCC patients with unknown etiology of origin of HCC ($n=10$), adjacent non-tumor control tissue samples ($n=14$), and a healthy control liver sample ($n=1$) [8]; and [3] existing liver scRNA-seq data from normal liver samples ($n=9$) [7]. After integrating these data and identifying the cell-types, we leveraged the cell-type transcriptional profiles to estimate cell-type proportions (decompose) in bulk liver RNA-seq samples from the well-established TCGA cohort [26] (361 patients with primary HCC tumors, 49 of whom have paired adjacent non-tumor tissue samples) and searched for cell-types that are significantly enriched in HCC. To replicate these findings, we used the LCI cohort [27] with microarray data from 221 patients with primary HCC tumors, of whom 209 have paired adjacent non-tumor biopsies. Next, we tested the effect of the HCC-increased cell-type on survival outcomes in the TCGA and LCI cohorts. Finally, we searched for associations between somatic mutations and the increased cell-type proportion estimates in HCC (for the overall study design, see Additional file 1: Fig. S2).

Data integration, clustering, and cell-type assignment in three single cell level RNA-seq cohorts

To decompose liver bulk RNA-seq cell-types in the TCGA and LCI cohorts, we first set up a single cell level reference data set. We utilized three single cell level cohorts generated using either snRNA-seq or scRNA-seq to build a powerful liver cell-type reference data set with a large number of cells and multiple HCC etiologies represented. Briefly, the included cohorts consist of both viral and non-viral origin HCC biopsy samples, adjacent non-tumor control samples, and normal liver samples (for cohort descriptions see Methods). Merging of the three data sets without integration resulted in cohort-specific clustering, indicating the presence of batch effects (Additional file 1: Fig. S3). When merging without integration, we also observed evidence of inter-patient heterogeneity across the 17 paired HCC samples (Additional file 1: Fig. S3). In order to identify shared cell-types and correct for these batch effects, we integrated these single cell level expression data using the CCA approach [38, 49] that should retain biologically meaningful signals while reducing technical variance (Fig. 1a,b). The integrated data were clustered using Seurat [38], resulting in the identification of 25 cell-types (Fig. 1a,b).

Discovery of an HCC-associated, cell-cycle-related cell-type in the single cell level data

Clustering of the integrated single cell level data (123,956 analyzed nuclei/cells) identified 25 subcell-types in total (Fig. 1b), which we merged and classified into 8 main cell-types based on their lineage (Fig. 1a). Subcell-types and main-cell types were classified based on expression of known marker genes and enriched pathways (Fig. 1c; Additional file 1: Fig. S4; Additional file 2: Table S1 and Additional file 3: Table S2). We then searched for subcell-types/cell-types enriched or depleted in HCC tumor cells (Fig. 1d). We observed a significant enrichment of tumor cells (80.4%) in a new cell-type cluster that we named Proliferative (Prol) cell-type (Fig. 1d,e). The pathway analysis of its marker genes suggested that this tumor-enriched cell-type consists of mitotic cells (Fig. 1c, see below). We also observed a significantly increased number of tumor cells in T, myeloid, and hepatocyte subcell-types and a decreased number of tumor cells in natural killer T subcell-type (Fig. 1d). Thus, our multi-cohort integration of both snRNA-seq and scRNA-seq data allowed us to identify the tumor cell-enriched Prol cell-type that had not been identified previously. The top pathway enrichments of the Prol marker genes were oxidative phosphorylation and cell cycle, suggesting that their functions are related to growth and cell division (Fig. 1c). To further investigate the proliferative capacity of Prol, we assigned G2M and S module scores based on average expression of G2M and S cell cycle genes [48] (see Methods),

(See figure on next page.)

Fig. 1 Multi-cohort integration of three liver HCC single cell level data sets identifies and characterizes an HCC-associated cell-type. We assessed liver cell-types and HCC-related cell-type changes by integrating Aizarani et al. [7] scRNA-seq data ($n=9$ non-HCC samples), Sharma et al. [8] scRNA-seq data ($n=1$ non-HCC, $n=14$ HCC-tumor, and $n=14$ adjacent non-tumor samples), and Rao et al. [25] snRNA-seq data ($n=3$ HCC-tumor and $n=3$ non-tumor samples). **a**, **b** Uniform Manifold Approximation and Projection (UMAP) visualization of 123,956 cells and nuclei integrated to remove cohort-specific effects. Clusters were assigned to (**a**) 8 major cell-types and (**b**) 25 subcell-types. **c** Pathway gene set enrichment analysis of the expression profiles for each subcell-type using the Reactome pathway database. The enr values indicate normalized enrichment scores and *q*-values denote Benjamini-Hochberg-adjusted *p*-values. Full pathway names are shown in Additional file 3: Table S2. **d** The bar plot shows the proportion of cells/nuclei in the full set of 123,956 cells/nuclei originating from HCC tumor and non-tumor samples separated by subcell-type. Darker fills indicate an FDR-adjusted *p*-value < 0.05 from a paired Wilcoxon test between proportions of HCC tumor and non-tumor samples. **e** Proportions of the Proliferative (Prol) cell-type are significantly higher in the 17 HCC tumor samples than in their 17 adjacent paired non-tumor samples after correcting for multiple testing with FDR, as assessed by a paired Wilcoxon test. **f**, **g** UMAP plots with cells/nuclei colored by their cell cycle score in the full single-cell level RNA-seq data of 123,956 droplets show that the Prol cluster consists of droplets with higher expression of (**f**) G2M phase genes and (**g**) S phase genes. The asterisks denote the significance of a difference between G2M and S phase gene scores between Prol and non-Prol cells/nuclei. Significance levels for *p*-values in (**e**–**g**) *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$. B indicates B cells; Chol, cholangiocytes; Endo, endothelial cells; Hep, hepatocytes; Myel, myeloid cells; Stell, stellate cells; and T, T cells

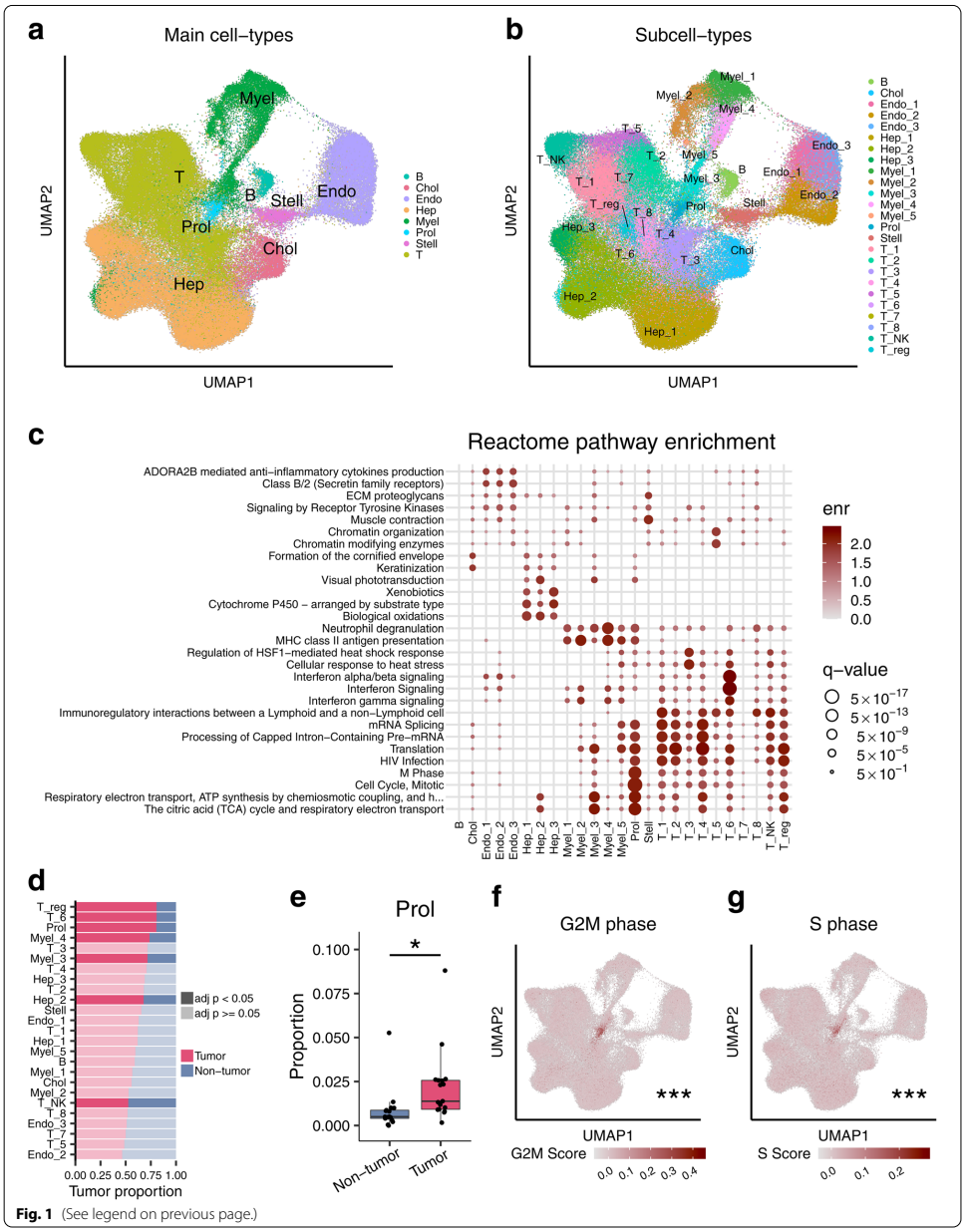


Fig. 1 (See legend on previous page.)

respectively. We found that cells/nuclei from Prol demonstrated significantly higher S and G2M phase module scores when compared to other cell-types (Fig. 1f,g). The higher cell cycle scores imply that Prol consists of actively dividing cells. To determine the cell-type composition of these proliferating cells, we re-classified the Prol cells/nuclei into main cell-types using a reference trained on the non-Prol cells/nuclei. In addition to hepatocytes, all non-parenchymal cell-types were observed in this tumor-enriched cluster (Additional file 1: Fig. S5). This presence of dividing non-hepatocyte cells observed in the tumor-enriched Prol cluster highlights the importance of the microenvironment in supporting HCC growth [50].

We next explored the marker genes within the Prol cell-type to further understand its biology. We identified 656 protein-coding marker genes in Prol, of which 15 had a log fold change >1 for differential expression between the Prol and other cell-types (Additional file 2: Table S2). Most of these 15 strongest Prol marker genes (12/15; 80%) had previously been identified in HCC pathogenesis or associated with clinical features of the disease [51–58]. Consistent with our findings, liver bulk expression of the histone protein, *H2AFZ*, a marker gene in Prol, was also identified in an independent HCC study to be associated with cell cycle genes regulated by TP53 [59]. However, among the 15, we discovered three genes, *HMGN2*, *RARRES2*, and *HIST1H4C*, which have previously been described in other malignancies [60, 61] but not in HCC. Two of these, *HIST1H4C* and *HMGN2*, are nuclear proteins that bind to nucleosomal DNA, consistent with Prol having higher S and G2M scores (Fig. 1f, g).

Overall, the single cell level reference data suggest that the Prol cell-type is associated with HCC. Therefore, we next used this single cell level reference data set to decompose cell-type proportions in the liver bulk RNA-seq HCC cohorts, TCGA and LCI, and then tested them for cell-type proportion differences between the HCC tumor and adjacent non-tumor control tissues.

Decomposition of cell-type proportions in HCC and adjacent non-tumor samples discovers high proportions of the proliferative cell-type Prol in HCC

Next, we sought to determine whether cell-type composition changes observed in our single cell level reference data were conserved and universally present in HCCs. Therefore, we estimated cell-type proportions for the 8 main cell-types and 25 subcell-types from bulk liver RNA-seq data in the TCGA Liver Hepatocellular Carcinoma (TCGA-LIHC) cohort, consisting of 361 non-recurrent primary tumors and 49 paired adjacent non-tumor samples (total $n=410$). We investigated the proportion estimates only for the 8 main cell-types as we found that estimates of the 25 subcell-types showed high intra-group correlation within their broader classifications (Additional file 1: Fig. S6), and these types of high correlations typically prevent accurate decomposition of subcell-types in bulk tissues [14]. For cell-type decomposition, we utilized Bisque [14], as described in detail in the Methods, resulting in proportion estimates for the 8 main cell-types. The marker genes of these main cell-types used for decomposition in Bisque (Additional file 4: Table S3) show high intra-cell-type co-expression and correlation with their respective proportion estimates (Additional file 1: Fig. S7a), suggesting their validity for estimating proportions. We then searched for differences in the abundance of these 8 cell-types between the paired HCC tumor and non-tumor tissue in TCGA. Of the 8 cell-types, we found that only Prol was significantly increased (Wilcoxon adjusted $p=5.68\times10^{-14}$) in the 49 HCC tumors when compared to the paired adjacent non-tumor samples in TCGA, while 5 cell-types significantly decreased in tumors (Fig. 2a and Additional file 5: Table S4). This increase in Prol abundance was consistent with our observations in the single cell level data (Fig. 2a and Fig. 1d).

To replicate the cell-type differences we identified in the TCGA cohort, we investigated the LCI cohort that consists of mainly Asian HCC patients with HBV-HCC. We estimated the proportions of 7 of the 8 main cell-types in the liver microarray data from tumor ($n=221$) and adjacent non-tumor ($n=209$) biopsies (see Methods). We excluded B cells, as its marker genes showed little to no co-expression in the microarray data of this cohort, and thus the proportions could not be estimated reliably (Additional file 1: Fig. S7b). All of the other 7 main cell-types demonstrated higher intra-cell-type co-expression and correlations with their respective proportion estimates (Additional file 1: Fig. S7b). Then, we tested for differences between the tumor and adjacent non-tumor biopsies. We found strikingly similar cell-type changes between the tumor and non-tumor tissues in the LCI and TCGA cohorts (Fig. 2b and Additional file 5: Table S4). Only the Prol cell-type was significantly increased in HCC in the LCI cohort (Fig. 2b), while the myeloid, T, and Hep clusters were significantly decreased in both TCGA and LCI, with Hep showing the largest decrease (Fig. 2b). These replicated results show that Prol is the only consistently upregulated cell-type in HCC tumors using both the TCGA and LCI cohorts.

We then sought to validate the observed increase in the Prol proportion estimates in HCC tumors by analyzing gene-level differential expression between tumors and adjacent non-tumors from the bulk liver data. We first took the most specific cell-type marker genes with a log fold change >0.5 in the single-cell level data and searched for differences in expression between the tumor and

non-tumors in the bulk. The marker genes for the Prol cluster had the highest average log fold changes in both the TCGA and LCI cohorts when compared to all other cell types (Fig. 2c,d). We then performed a reciprocal analysis by taking all tumor upregulated genes with a log fold change greater than 1 in the bulk data and scoring cells/nuclei in the single-cell level data for their average expression using the module score option in Seurat [38]. We found that cells/nuclei from Prol had the highest bulk tumor scores when using the strongest tumor-upregulated genes from both TCGA (Wilcoxon $p < 2.2 \times 10^{-16}$) and LCI (Wilcoxon $p < 2.2 \times 10^{-16}$) (Fig. 2e–h). Taken together, the significantly increased expression of Prol marker genes at the bulk HCC tissue level, and vice versa the highest expression of the bulk tumor-upregulated genes in the Prol cell-type, support an increased abundance of the Prol cell-type itself in HCCs.

The Prol cell-type is associated with HCC survival outcomes in TCGA and LCI

To determine the clinical significance of the Prol cell-type on survival outcomes in TCGA [32], we assessed its impact on overall survival (OS) and progression-free interval (PFI) in the 361 HCC patients. We hypothesized that an increased proportion of the tumor-associated Prol cell-type may be associated with poorer OS and PFI outcomes. To investigate this, we first associated the Prol cell-type proportions with survival outcomes in TCGA. We stratified the HCC patients into low and high cell-type proportion groups using the median (see Methods) and performed a Cox proportional hazards regression adjusting for age, sex, and ethnicity (Additional file 6: Table S5). Noteworthy, in TCGA, Prol had a statistically significant hazard ratio above 1 for both OS (HR = 1.76; $p = 4.77 \times 10^{-3}$) and PFI (HR = 1.89; $p = 1.25 \times 10^{-4}$) (Table 1, Fig. 3a,b). The Prol survival associations were even more pronounced after stratifying by quartile and remained significant after adjusting for tumor stage (Table 1). As expected, the other cell-types did not significantly decrease OS or PFI in TCGA (Additional file 6: Table S5). These results suggest that a high estimated Prol cell-type proportion is associated with poor survival outcomes and plays a key role in HCC tumor aggressiveness.

Next, we sought to replicate the effect of the HCC risk cell-type Prol on survival in the LCI cohort. As OS was the only available overlapping outcome in LCI, we used OS for our validation analysis. We performed Cox proportional hazards regression adjusting for age and sex. Testing the effect of the Prol cell-type on OS in LCI resulted in a significant hazard ratio (HR = 1.79; $p = 8.79 \times 10^{-3}$) (Table 1 and Fig. 3c) and remained significant after adjusting for stage (HR = 1.67; $p = 2.34 \times 10^{-2}$). This result replicated our finding observed in TCGA, demonstrating that a higher Prol is associated with a worse OS outcome in LCI as well. Taken together, the negative link between the tumor Prol cell-type and survival is robust and reproducible across independent HCC cohorts.

We again sought to validate our proportion-based results at the gene level. To do so, we analyzed the relationship between survival outcomes and gene expression of individual cell-type markers. We first performed Cox proportional hazards regression adjusting for age, sex, and ethnicity for all expressed genes in the TCGA HCC liver expression data for OS and PFI as outcomes. We observed that a higher number of Prol-specific marker genes (log fold change > 0.5) had a hazard ratio over 1 for OS (71.7%) and PFI (58.7%) compared to those of all other cell-types (Fig. 3d,e). Additionally, 23.9% and 17.4% of Prol markers had a genome-wide significant hazard ratio for OS and PFI, respectively, all of which were associated with a worse prognosis (Additional file 1: Fig. S8a,b). To replicate these findings, we performed Cox proportional hazards regression for OS in the LCI

(See figure on next page.)

Fig. 2 Among all cell-types decomposed in the TCGA and LCI bulk liver cohorts, Prol has the highest enrichment in HCC when compared to adjacent non-tumor tissue. The Prol cell-type shows consistent upregulation in HCC tumors in two independent liver bulk cohorts. **a**, **b** Proportions were estimated in the liver bulk RNA-seq data for the major cell-types identified in the single-cell level data and then tested for differential abundance between the tumor and non-tumor samples. The upper panel shows the T-statistic from a paired *t*-test between tumor and adjacent non-tumor tissue, with FDR-adjusted *p*-values calculated from a paired Wilcoxon test. The bottom panel shows a bar plot of the proportion estimates separated by tumor status. The differential abundance tests highlight the Prol cell-type as upregulated in the (**a)** TCGA ($n = 49$) and (**b)** LCI ($n = 209$) cohorts. B cell proportions were not estimated for LCI (**b**) as its marker genes did not show evidence of co-expression. **c**, **d** Association of the Prol cell-type with HCC tumors is highlighted by the $\log_2$ fold-changes ($\log_2$FC) of tumor over adjacent non-tumor samples for the marker genes of the cell-types that are indicated on the y-axis. Log_2FC values were derived from a paired differential expression (DE) analysis in (**c)** TCGA ($n = 49$) and (**d)** LCI ($n = 209$) cohorts. **e**–**h** The Prol cells/nuclei significantly express tumor-elevated genes, as shown by droplet scores in the single-cell level data for tumor-elevated genes derived from the TCGA and LCI cohorts. Genome-wide DE analysis was performed between the paired tumor and non-tumor samples, and genes with an FDR-adjusted *p*-value less than 0.05 and a $\log_2$FC greater than 1 were considered tumor-elevated genes. Module scores of the tumor-elevated genes for each droplet were calculated based on their expression compared to a background set. **e**, **f** UMAP plots for (**e)** TCGA and (**f)** LCI are shown with cells and nuclei colored by their tumor module score. **g**, **h** Bar plots show the droplet tumor scores calculated from (**g)** TCGA and (**h)** LCI tumor-elevated genes separated by major cell-type. **e**–**h** Asterisks denote a significant difference in gene scores between Prol and non-Prol cells/nuclei as assessed by a Wilcoxon test. Significance levels for *p*-values: $^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$

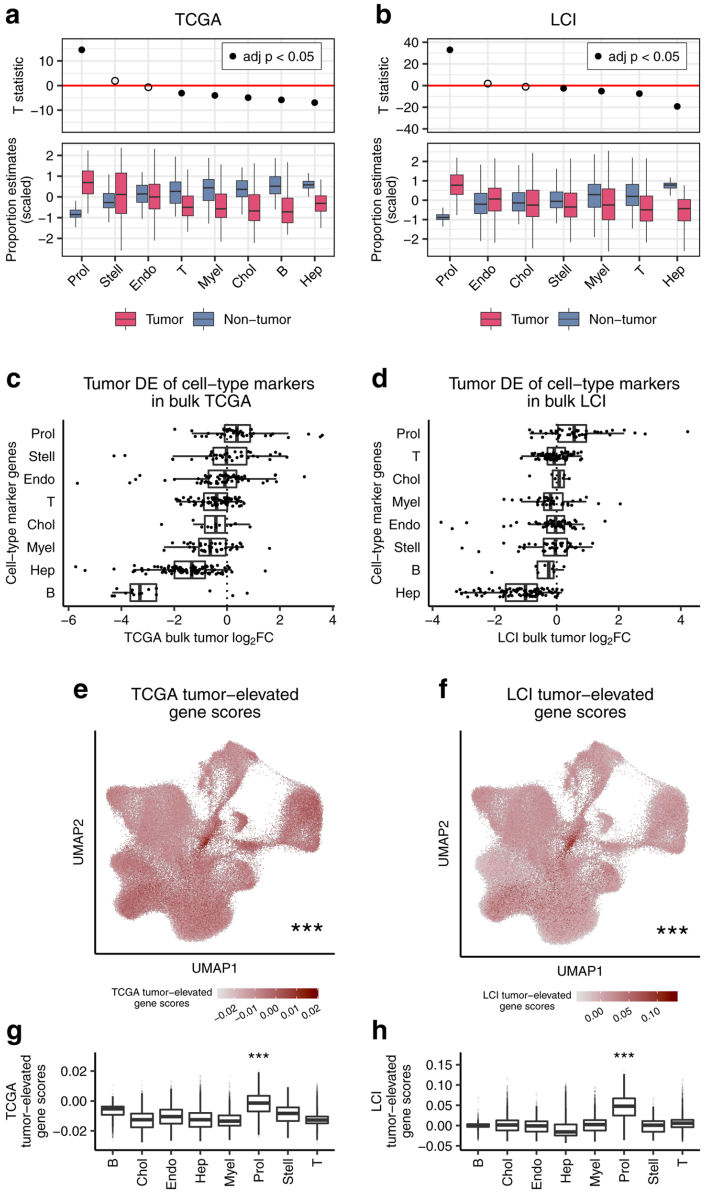


Fig. 2 (See legend on previous page.)

Table 1 Increased abundance of the tumor-associated cell-type Prol is associated with a worse prognosis both in the TCGA and LCI cohorts

Cohort	Event	Prol model	Multivariable HR	95% CI	*p*-value
TCGA	OS	Median	1.76	1.19–2.61	4.77×10^{-3}
TCGA	OS	Median adj. stage	1.52	1.02–2.26	4.20×10^{-2}
TCGA	OS	Quartile	3.25	1.84–5.72	4.62×10^{-5}
TCGA	PFI	Median	1.89	1.37–2.63	1.25×10^{-4}
TCGA	PFI	Median adj. stage	1.73	1.24–2.41	1.14×10^{-3}
TCGA	PFI	Quartile	2.85	1.76–4.63	2.14×10^{-5}
LCI	OS	Median	1.79	1.16–2.76	8.79×10^{-3}
LCI	OS	Median adj. stage	1.67	1.07–2.60	2.34×10^{-2}

Hazard ratios of overall survival and progression free interval based on the Prol cell-type proportion in the TCGA HCC cases ($n = 361$) and hazard ratios of overall survival in the LCI HCC cases ($n = 221$) show that an increased abundance of Prol is associated with decreased survival. Cox proportional hazard regression was performed for the event and model indicated. The Prol model indicates the predictor tested. The median model stratifies the cases into low and high abundance groups based on whether the individual's estimated Prol proportion was below or above the median, respectively. The median adjusted (adj.) stage results are obtained by including in the median model a covariate for the low and high AJCC tumor stage status, where stage I and II form the low stage and stage III and IV form the high stage. The quartile model tests low and high abundance groups by splitting participants below and above the 25th and 75th percentile of Prol proportion estimates, respectively. All tests in TCGA were adjusted for age, sex, and ethnicity. All tests in LCI were adjusted for age and sex. Unadjusted *p*-values are shown. *HR* indicates hazard ratio, *CI* confidence interval, *OS* overall survival, *PFI* progression free interval

cohort. Although we did not observe a notable number of genome-wide significant effects, we found a similar enrichment of Prol marker genes with a hazard ratio over 1 for OS (63.6%) (Fig. 3f and Additional file 1: Fig. S8c). These marker gene results further support the conclusion that the Prol cell-type itself is associated with survival.

Finally, we evaluated the cell-type enrichment for all genes with a significant association with OS and PFI in the bulk TCGA cohort and with OS in the LCI cohort. Cells/nuclei in the single-cell level data were assigned survival-decreasing module scores using Seurat for expression of the 740 and 528 genes with a significant hazard ratio above 1 for OS and PFI in the TCGA bulk RNA-seq data, respectively (FDR adjusted $p < 0.05$). We found that the Prol nuclei/cells had the highest average OS-decreasing (Wilcoxon $p < 2.2 \times 10^{-16}$) and PFI-decreasing (Wilcoxon $p < 2.2 \times 10^{-16}$) scores (Fig. 3g, h and Fig. 3j, k), indicating that Prol over-expresses DE genes associated with poor survival outcomes in TCGA more prominently than all other cell-types. In order to replicate these results in the LCI cohort, we scored cells/nuclei for expression of the 36 genes that among all genes had a significant hazard ratio above 1 for OS in LCI (FDR adjusted $p < 0.05$). Again, the Prol cluster had the highest average OS-decreasing score from the LCI association results (Wilcoxon $p = 2.98 \times 10^{-151}$) (Fig. 3i, l). Thus, by taking all genome-wide significant results in an unbiased manner, we highlight the Prol cell-type in poor survival outcomes. Overall, our bulk-based single cell level findings, showing that Prol nuclei/cells significantly over-express both tumor-elevated bulk DE genes (Fig. 2e–h) and survival-decreasing bulk DE genes (Fig. 3g–l), support the association of the Prol cell-type with HCC and worse survival independently from our decomposition analysis.

Somatic TP53 mutations are associated with increased proportions of the Prol cell-type in HCC

Somatic mutations in HCC have been characterized in several cohorts, and although heterogeneous, these studies have identified commonly mutated driver genes [5]. However, it has remained elusive whether somatic mutations can lead to specific tumor cell-type expansions or depletions. Therefore, we performed associations between cell-type profiles against mutations in

(See figure on next page.)

Fig. 3 The HCC-enriched Prol cell-type associates with overall survival (OS) and progression free interval (PFI) in TCGA and with OS in LCI. Increased Prol cell-type proportion estimates are associated with poor survival outcomes in TCGA and LCI. **a**–**c** Kaplan-Meier survival curves for (**a**) overall survival (OS) and (**b**) progression free interval (PFI) in TCGA and (**c**) OS in LCI show worse survival outcomes for patients with high liver Prol cell-type frequency estimates. Patients with Prol frequency (freq.) estimates above and below the median were classified into high and low groups, respectively. The "+" signs on the line indicate right censoring of the event. The hazard ratios (HR) and FDR adjusted *p*-values were calculated from a Cox proportional hazards regression adjusting for age, sex, and for TCGA, race. **d**–**f** Association of the Prol cell-type with poor survival outcomes is highlighted by the HR values for cell-type marker genes calculated from a Cox proportional hazards regression of their expression in TCGA and LCI. Survival tests were performed for (**d**) OS and (**e**) PFI in TCGA and (**f**) OS in LCI. Each dot indicates a gene, with its HR on the *x*-axis and its cell-type on the *y*-axis. **g**–**l** Module scores of survival-decreasing genes in the single-cell level data are significantly higher in cells/nuclei from the Prol cell-type. Survival-decreasing genes were derived from genome-wide Cox proportional hazards regression analyses of all genes for the indicated event and cohort and taking the genes with FDR-adjusted *p*-values less than 0.05 and HR values greater than 1.0 into the module score analyses in (**g–l**). **g**–**i** UMAP plots show cells/nuclei colored by (**g**) TCGA OS score, (**h**) TCGA PFI score, and (**i**) LCI OS scores. **j**, **l** Bar plots of survival-decreasing module scores for (**j**) TCGA OS, (**k**) TCGA PFI, and (**l**) LCI OS separated by the cell-type. **g**–**l** Asterisks denote a significant difference in survival-decreasing gene scores between Prol and non-Prol cells/nuclei as assessed by a Wilcoxon test. Significance levels for *p*-values: $*p < 0.05$, $**p < 0.005$, $***p < 0.0005$

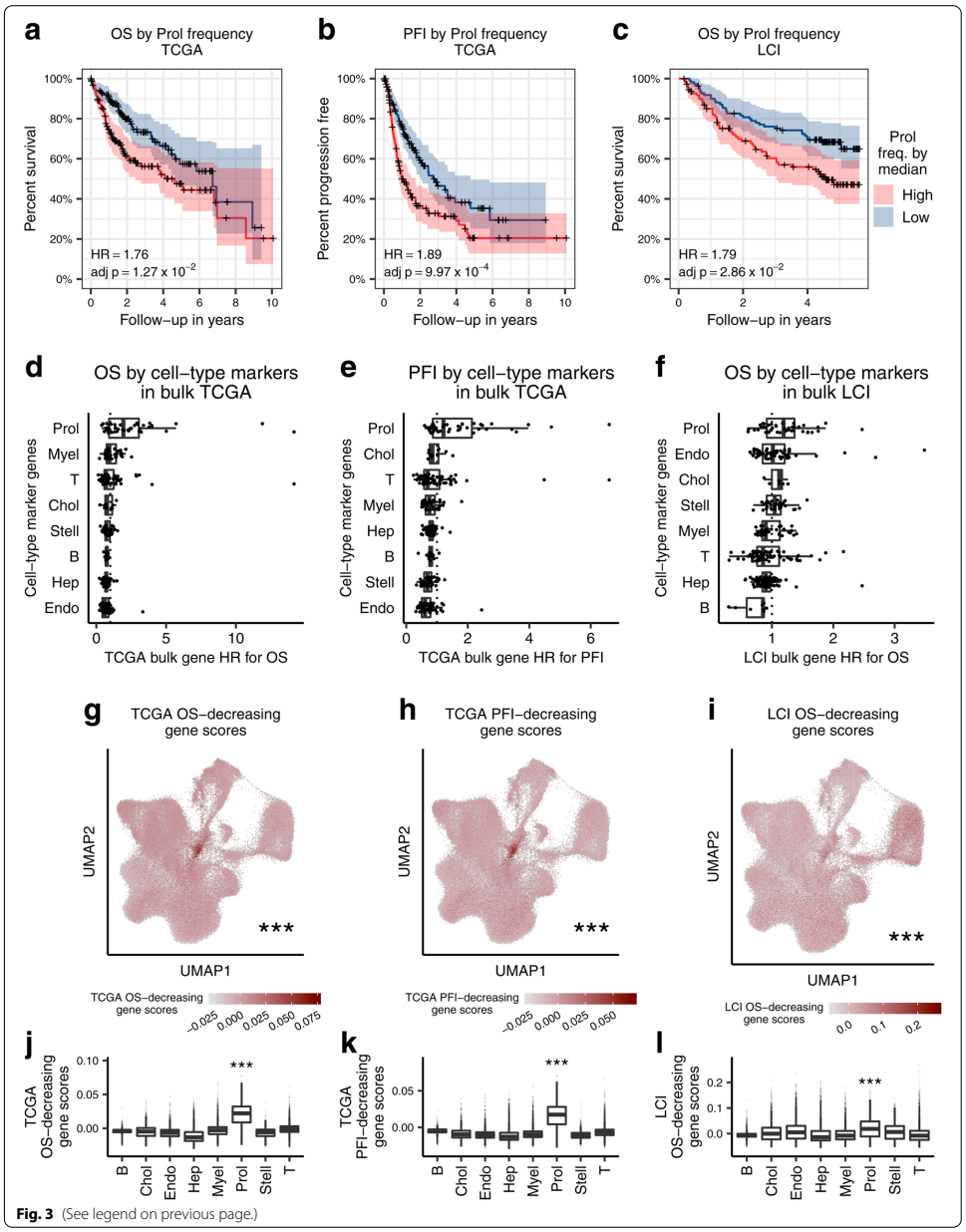


Fig. 3 (See legend on previous page.)

the 69 significantly mutated genes that have previously been characterized in TCGA HCCs (https://gdac.broadinstitute.org). We identified 3 genes associated with a higher cell-type abundance (Wilcoxon adjusted $p < 0.05$) (Additional file 7: Table S6). Among these, mutations in *TP53* (Wilcoxon adjusted $p = 7.58 \times 10^{-9}$) and in *RB1* (Wilcoxon adjusted $p = 9.45 \times 10^{-3}$) led to a significant increase in the estimated proportions of the Prol tumor cell-type (Fig. 4a and Additional file 1: Fig. S9). Furthermore, Prol was the only significantly increased cell-type in individuals with *TP53* mutations or *RB1* mutations (Fig. 4b). Interestingly, we also observed that *BAP1* mutations are associated with an increase in cholangiocyte proportion estimates. *BAP1* has been shown to be frequently inactivated in cholangiocarcinomas [62].

Many mutations in *TP53* are known to lead to a loss of the tumor suppressor function of p53 and consequently uncontrolled cell growth [63]. We therefore tested the effect of distinct *TP53* mutation types on Prol abundance. We observed that *TP53* missense, frame shift, and nonsense mutations led to significantly higher proportions of Prol (Fig. 4c). While frame shift and nonsense mutations are likely to lead to a total loss of function, missense mutations in *TP53* have been found to occur mainly in the DNA-binding domain of the protein, also leading to a loss of its tumor suppressor function [63]. To further investigate whether the Prol cell-type is the main consequence of *TP53* mutations, we first identified 1358 mut. over-expressed genes with significant log fold changes greater than 0.5 between mutation (mut.) carriers and wildtype (WT) cases in TCGA. We then assigned module scores to droplets in the single-cell level data based on expression of these mut. upregulated genes. We found that the Prol cells/nuclei had significantly higher *TP53* mutation scores than all other cells/nuclei (Wilcoxon $p < 2.2 \times 10^{-16}$) (Fig. 4d, f). We performed the same analysis for 774 *RB1* DE genes and found a similar enrichment of mut. upregulated gene scores in Prol droplets (Wilcoxon $p < 2.2 \times 10^{-16}$) (Fig. 4e, g). Overall, these results suggest that distinct somatic mutations can lead to a tumor cell-type expansion and highlight the role of *TP53* mutations in proliferation and uncontrolled cell growth.

Discussion

We developed a new framework using comprehensive single cell level reference data from multiple etiologies of HCC, adjacent non-tumor, and normal liver tissue to decompose cell-types in liver bulk RNA-seq and microarray expression data generated from HCC and adjacent non-tumor tissue in the TCGA and LCI cohorts. This integrative transcriptomics framework identified an HCC-associated proliferative cell-type, Prol, the high proportion of which in HCC tumors is associated with significantly worse survival outcomes. Noteworthy, we first observed this survival effect in TCGA, and then replicated our finding in LCI. Our results should be robust not only because we replicated our findings in an independent cohort, but also because they do not depend on the technology used to measure single cell and tissue-level gene expression in the liver, given that both scRNA-seq and snRNA-seq were used to build the reference data set and both bulk RNA-seq in TCGA and microarray technology in LCI were used to decompose the cell-types in the liver tissue. Furthermore, our reciprocal module score analyses show that Prol nuclei/cells significantly over-express both tumor-elevated DE genes and survival-decreasing DE genes obtained from the bulk expression data in the TCGA and LCI cohorts. Thus, these bulk-based single cell level results further support the association of the Prol cell-type with HCC and worse survival independently from the decomposition analysis. When searching for mutated driver genes of the HCC cell-types, we found that among 69 genes with somatic mutations catalogued in TCGA earlier (https://gdac.broadinstitute.org), Prol is the only significantly increased cell-type in individuals with *TP53* and *RB1* mutations. Thus, we

(See figure on next page.)

Fig. 4 Associations between estimated cell-type proportions and somatic mutations in the TCGA cohort link *TP53* and *RB1* mutations to increased Prol abundance. Mutations associated with changes in the bulk TCGA liver proportion estimates of the Prol cell-type. **a** Prol proportion estimates are significantly higher in the HCC cases harboring a mutation (Mut) in *TP53* (left panel) and *RB1* (right panel) compared to those with both wildtype (WT) alleles. **b** The Prol cell type is highlighted as the only cell-type significantly increased in HCC cases with Mut *TP53* and Mut *RB1*. Differential abundance for the 8 cell-types testing for differences in proportions between Mut vs. WT *TP53* (top panel) and *RB1* (bottom panel) cases. Differential abundance was performed with a Wilcoxon test ($n = 357$ tumor samples). The difference in means of the scaled proportions is plotted in the *x*-axis and the -log10 *p*-value in the *y*-axis. The vertical red line ($x = 0$) indicates no difference. **c** Prol proportion estimates are plotted against no *TP53* mutation (None) and different *TP53* mutation types. Prol estimates are significantly increased in individuals with loss of function (LOF) mutations in *TP53*. **d**–**g** The cells/nuclei in the Prol cell-type significantly express mutation-upregulated genes, as shown by the droplet module scores of mutation upregulated genes for the indicated mutation in TCGA. Mutation upregulated genes were derived by running genome-wide differential expression (DE) between patients with and without a somatic mutation in the indicated gene and taking those over-expressed in HCC patients harboring a mutation and with an FDR-adjusted *p* value less than 0.05. Droplet module scores were calculated by comparing the average expression of mutation upregulated genes to a background set of genes. **d**, **e** UMAP of the single-cell-level data showing droplets colored by scores for genes upregulated in patients with (**d**) *TP53* and (**e**) *RB1* mutations. **f**, **g** Bar plots of the (**e**) *TP53* mutation upregulated scores and (**g**) *RB1* mutation upregulated scores separated by cell-type. **d**, **g** Asterisks denote a significant increase in mutation upregulated gene scores between Prol and non-Prol cells/nuclei as assessed by a Wilcoxon test. Significance levels for nominal *p*-values in (**a**, **c**, **d**-**g**): $*p < 0.05$, $**p < 0.005$, $***p < 0.0005$

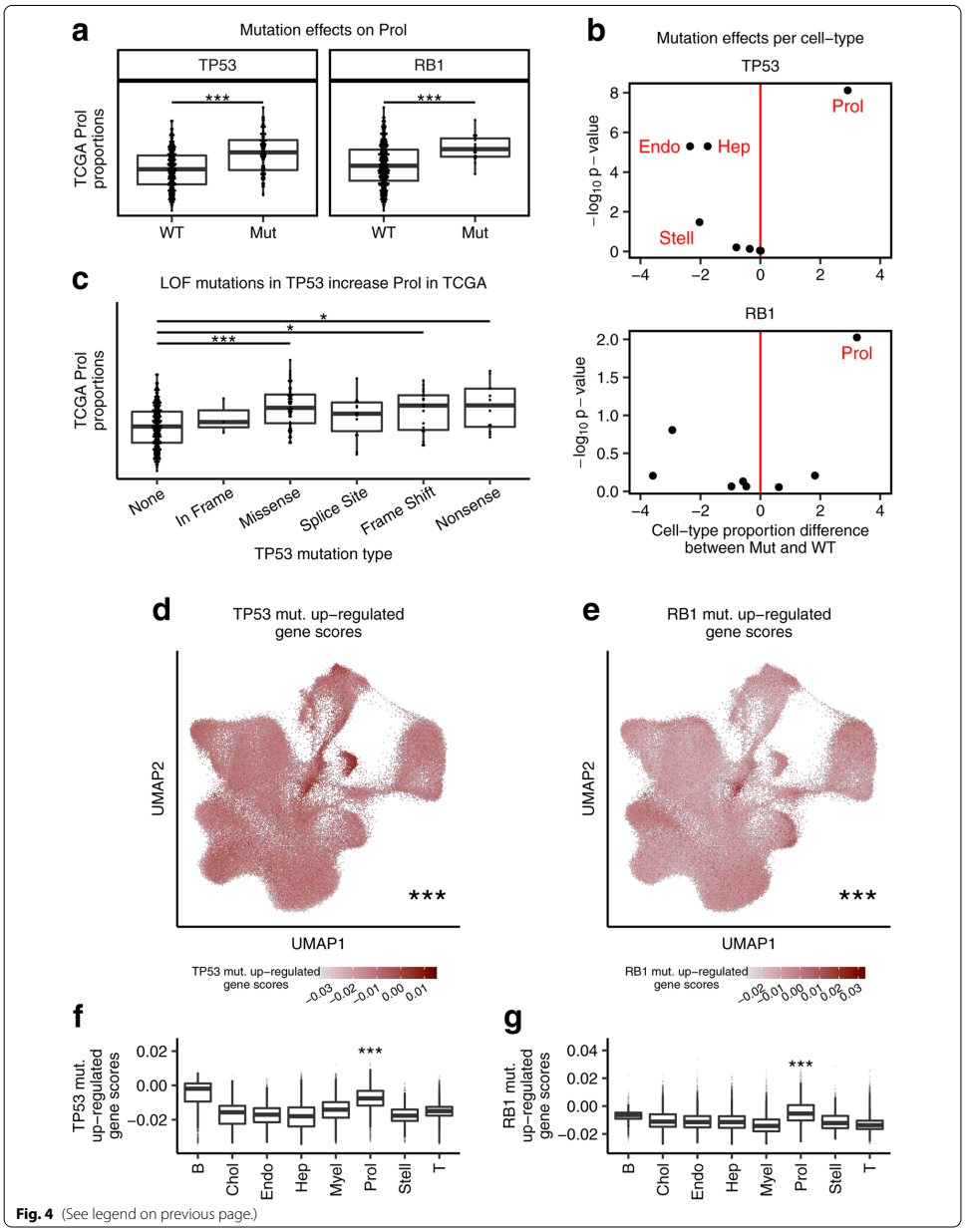


Fig. 4 (See legend on previous page.)

show that mutations in these tumor suppressor genes are associated with the expansion of the tumor-associated Prol cell-type in HCC.

Exploring cell-type heterogeneity provides a novel avenue to study microenvironment in cancer cells. The notion that tumor microenvironment, specifically of immune cells, may affect tumor progression and affect survival first stemmed from non-HCC studies, such as ovarian tumors [64]. This was subsequently investigated in HCC [19, 65], which has clinical implications given that only ~18% of patients responded to therapies targeted to immune-dependent pathways with checkpoint inhibitors (programmed death 1 or PD1) in early clinical trials [66]. Losic and colleagues used scRNA-seq from 2 patients across multiple regions within the same tumor to demonstrate tumor heterogeneity, which provided early evidence that the immune microenvironment is heterogeneous between patients and within samples [19]. In line with these studies, we observed similar heterogeneity in cell-type composition within HCC patients from multiple etiologies. Using single cell level data generated by both scRNA-seq and snRNA-seq, we found patient-specific clusters when not correcting for this heterogeneity with integration (Additional file 1: Fig. S3).

Large existing cohorts, such as TCGA [26] and LCI [27], provide invaluable tools to the research community. Accordingly, we leveraged our integrated liver single cell level data to identify cell-types associated with HCC and their clinically significant outcomes in TCGA and LCI, both with long-term follow-up data. The systematic identification of the Prol cell-type across the single cell level reference data with multiple etiologies of HCC, the TCGA cohort (mostly viral etiologies with HBV and HCV), and the LCI cohort (HBV-predominant origin of HCC) suggests universal points of convergence in HCC pathogenesis that can be further investigated at the single cell level. Understanding tumor biology at the cell-type level instead of the bulk tissue level provides more insight into the underlying tumor biology [67]. Several of the Prol marker genes have previously been associated with poor survival outcomes [52–59, 68]; however, our study discovered that these genes form a distinct HCC-associated cell-type. Furthermore, we discovered that somatic mutations in *TP53* and *RB1* are associated with increased Prol proportions in HCC. Interestingly, differences in somatic mutations have also been observed in various etiologies of HCC, with the *TP53* mutations being linked to viral and alcohol etiologies of HCC [69] (similar to the patient composition of the TCGA and LCI cohorts), while *ACVR2A* (activin A receptor type 2A) mutations have been more commonly found in NASH-HCC [69].

Previous studies have identified molecular sub-classes of HCC that correlate with tumor phenotypes and clinical outcomes [6, 26, 70–72]. About half of all HCCs consist of the proliferative sub-class that predominantly have *TP53* mutations [6], which we also identified as significant mutations in our cell-type analyses. Our data suggest that somatic mutations in the tumor suppressor gene, *TP53*, result in dysregulation of mitosis and cell-cycle pathways, in line with their enrichment in the Prol cell-type. Consistent with our findings, in an independent study, the histone protein, H2AFZ that we identified as a marker gene in Prol, was associated with cell cycle genes and reported to be regulated by TP53 in HCC [59]. Overall, our integrative approach identified a cell-type with somatic mutations in a tumor suppressor gene that is significantly associated with worse overall survival. These results may improve current HCC subclassification and provide insight into co-dependent biological mechanisms of HCC.

Several of the genes identified in the Prol cell-type have previously been associated with poor overall or recurrence-free survival outcomes in HCC, including *PTMA* [52, 68], *HMGB2* [53], *HMGB1* [54], *H2AFZ* [59], *GAPDH* [55], *TUBB* [57], *STMN1* [56], and *TUBA1B* [58]. However, despite this growing body of literature identifying individual HCC genes with prognostic potential in the TCGA and other cohorts, our study used a single cell level-based decomposition approach to identify an HCC-associated cell-type, the proportion of which is significantly increased in HCC tumors with poor survival. The Prol cell-type suggests uncontrolled mitosis and cell-cycle dysregulation as converging mechanisms for worse survival. Furthermore, the Prol cell-type not only contains previously known HCC genes [52–59, 68], but also provides new targets, including *HMGN2*, *RARRES2*, and *HIST1H4C* that have not been explored yet. Overall, our integrative multi-cohort approach provided hundreds of Prol cell-type marker genes, which can be used to advance our understanding of the complex HCC biology in future studies.

Given the poor survival outcomes in patients diagnosed with HCC [3, 66], it is critical to further our understanding of factors affecting survival. We demonstrate that the use of cell-type markers could be of clinical utility as a potential future biomarker to guide treatment options and determining clinical outcomes. Current clinical prognostic tools of HCC mostly rely on the number and size of tumors, AFP, the presence of underlying chronic liver disease, and the patient's medical status. The use of cell-type markers as a tool to understand tumor biology can improve current clinical practice. Our Prol marker genes could serve as a basis for developing new expression-based prognostic technologies. For example, quantitative PCR could be used to rapidly perform predictive gene expression panel tests [73]. As RNA sequencing

matures, clinical labs can detect global gene expression patterns with prognostic value [74]. Assays such as these could measure Prol markers to evaluate the abundance of the two cell types and test if they predict clinical outcomes. Whether this cell-type is prognostic for HCC recurrence post resections or liver transplantation would also need to be determined. Our pipeline utilizing single cell level reference data to decompose cell-types in bulk RNA-seq can also be applied to other malignancies that have an admixture of heterogeneous cells to identify predominant cancer cell-types.

Although this study improves our understanding of new HCC cell-types with a potential for clinical implications, it is not without limitations. As HCC prevalence continues to rise and liver transplantation allocation policies are changing [75], larger studies with different HCC etiologies are needed in cirrhosis and non-cirrhosis backgrounds, especially given the observed differences in treatment responses [22]. In addition, cell-type changes in recurrent HCCs would have to be investigated in future studies. It should also be noted that although our survival analyses in TCGA discovered the significance of the Prol cell-type in OS and PFI, even after adjusting for tumor stage, other clinically relevant factors in HCC outcomes, including AFP levels, extent of chronic liver disease, presence of lymph vascular invasion on histopathology, and tumor size could not be explored in our models because up to 35% of the 361 individuals had missing data for these parameters. Thus, future studies are warranted to assess their correlations with the Prol tumor-associated cell-type.

Conclusions

In conclusion, using comprehensive single cell level reference data to decompose cell-types in the TCGA and LCI liver bulk tissue cohorts, we discover the important role of the previously unknown Prol cell-type in HCC and survival outcomes in TCGA, which replicated in LCI. We also linked somatic mutations in the tumor suppressors *TP53* and *RB1* to Prol cell-type expansion in HCC. Our integrative transcriptomics pipeline can be extrapolated to other cancer cohorts to identify key tumor cell-types using single cell level samples as the cell-type reference data. The detection of tissue-specific and cancer-associated cell-types can advance our understanding of tumor biology with a great potential for biomarker discovery in larger, prospective validation studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-022-01055-5.

Additional file 1: Supplementary figures (Fig. S1-S9).

Additional file 2: Table S1. Subcell-type marker genes.

Additional file 3: Table S2. Reactome gene set enrichment analysis of the subcell-type marker genes.

Additional file 4: Table S3. Bisque marker genes used for decomposition.

Additional file 5: Table S4. Differential proportion analysis between HCC tumor and adjacent non-tumor tissue in the TCGA and LCI liver bulk tissue cohorts.

Additional file 6: Table S5. Associations of main cell-type proportions with survival outcomes in TCGA.

Additional file 7: Table S6. Associations between main cell-type proportions and somatic mutations in TCGA.

Acknowledgments

We would like to thank the patients and investigators who participated in the TCGA Research Network (http://www.cancer.gov/tcga) and the Liver Cancer Institute. We also thank the participation of the patients from the Dumont-UCLA Liver Cancer Center. We would like to thank the participants of the two scRNA-seq cohorts from Sharma et al. [8] and Aizarani et al. [7]. We also thank the UCLA Technology Center for Genomics and Bioinformatics (TCGB) for sequencing of the human snRNA-seq samples. We would also like to acknowledge the *Hmisc* and *corrplot* soft wares used for our pair-wise correlation analyses.

Authors' contributions

Marcus Alvarez: conceptualization, methodology, investigation, computational analyses, validation, writing—original draft. Jihane N Benhammou: conceptualization, methodology, investigation, writing—original draft, funding acquisition. Nicholas Darci-Maher: computational analyses, validation, writing—original draft. Samuel W French: review of pathology slides from NAFLD-HCC cohort, writing—review, editing. Steven B Han: funding acquisition, writing—review, editing. Janet S Sinsheimer: statistical approach, writing—review, editing. Vatche G Agopian: provided all NAFLD-HCC patient samples for sequencing, writing—review, editing. Joseph R Pisegna: conceptualization, methodology, writing—original draft, supervision. Päivi Pajukanta: conceptualization, methodology, validation, writing—original draft, supervision, funding acquisition. All authors read and approved the final manuscript.

Funding

MA was supported by the HHMI Gilliam Fellowship. The sequencing work was funded by the American Association of the Study of Liver Diseases (AASLD) Advanced Transplant Hepatology grant and the CURE: Digestive Diseases Research Center (DKP304131) for JNB. JNB is a junior investigator in the NCI Translational Liver Cancer (TLC) Consortium (U01 CA230997-1). Tissue banking efforts were supported by the NIH/NCI grants R21 CA216807 and R01 CA246304 to VGA. Tissues were provided by the Translational Pathology Core Laboratory (TPCL) at UCLA, supported by the NIH Cancer Center Support Grant P30CA016042-45 (PD/PI Michael Teitell) to SWF. JSS was supported by the NIH grants R01HG006139, R01HG009120, and R35GM141798. This study was supported by the NIH grant R01HG010505.

Availability of data and materials

Raw snRNA-seq counts for the NAFLD-related HCC cohort [25] are available from NIH GEO under accession number GSE189175. Liver HCC scRNA-seq data from Sharma et al. [8] are available from https://data.mendeley.com/datasets/6wmzcskt6k/1. Read counts for the scRNA-seq of the 9 normal liver samples from Aizarani et al. [7] are available from NIH GEO under the accession number GSE124395. The TCGA [26] data are available for downloading at (https://portal.gdc.cancer.gov/projects/TCGA-LIHC) and (http://gdac.broadinstitute.org). The LCI [27] data are available for downloading from NIH GEO under accession number GSE14520. The code used for the analyses in this manuscript can be found in https://github.com/marcalva/hcc_sc_2022 [76].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [2]Vatche and Tamar Manoukian Division of Digestive Diseases, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [3]Division of Gastroenterology, Hepatology and Parenteral Nutrition, Department of Medicine, VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA. [4]Department of Pathology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [5]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [6]Department of Computational Medicine, UCLA, Los Angeles, CA, USA. [7]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA. [8]Dumont-UCLA Transplant and Liver Cancer Centers, Department of Surgery, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [9]Institute for Precision Health, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA.

Received: 31 January 2022 Accepted: 5 May 2022
Published online: 17 May 2022

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.
2. Singal AG, Pillai A, Tiro J. Early detection, curative treatment, and survival rates for hepatocellular carcinoma surveillance in patients with cirrhosis: a meta-analysis. PLoS Med. 2014;11:e1001624.
3. Jemal A, Ward EM, Johnson CJ, et al. Annual Report to the Nation on the Status of Cancer, 1975-2014, Featuring Survival. J Natl Cancer Inst. 2017;109:djx030.
4. Caruso S, Calatayud AL, Pilet J, et al. Analysis of liver cancer cell lines identifies agents with likely efficacy against hepatocellular carcinoma and markers of response. Gastroenterology. 2019;157:760–76.
5. Llovet JM, Zucman-Rossi J, Pikarsky E, et al. Hepatocellular carcinoma. Nat Rev Dis Primers. 2016;2:16018.
6. Llovet JM, Kelley RK, Villanueva A, et al. Hepatocellular carcinoma. Nat Rev Dis Primers. 2021;7:6.
7. Aizarani N, Saviano A, Sagar, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature. 2019;572:199–204.
8. Sharma A, Seow JJW, Dutertre CA, et al. Onco-fetal reprogramming of endothelial cells drives immunosuppressive macrophages in hepatocellular carcinoma. Cell. 2020;183:377–394.e21.
9. Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343:776–9.
10. Zong C, Lu S, Chapman AR, et al. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338:1622–6.
11. Grindberg RV, Yee-Greenbaum JL, McConnell MJ, et al. RNA-sequencing from single nuclei. Proc Natl Acad Sci U S A. 2013;110:19802–7.
12. Habib N, Avraham-Davidi I, Basu A, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14:955–8.
13. Alvarez M, Rahmani E, Jew B, et al. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. Sci Rep. 2020;10:11019.
14. Jew B, Alvarez M, Rahmani E, et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun. 2020;11:1971.
15. Karunakaran D, Turner AW, Duchez AC, et al. RIPK1 gene variants associate with obesity in humans and can be therapeutically silenced to reduce obesity in mice. Nat Metab. 2020;2:1113–25.
16. Miao Z, Alvarez M, Ko A, et al. The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance. PLoS Genet. 2020;16:e1009018.
17. Denisenko E, Guo BB, Jones M, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21:130.
18. Slyper M, Porter CBM, Ashenberg O, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat Med. 2020;26:792–802.
19. Losic B, Craig AJ, Villacorta-Martin C, et al. Intratumoral heterogeneity and clonal evolution in liver cancer. Nat Commun. 2020;11:291.
20. Ma L, Hernandez MO, Zhao Y, et al. Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. Cancer Cell. 2019;36:418–430.e6.
21. Horning AM, Wang Y, Lin CK, et al. Single-cell RNA-seq reveals a subpopulation of prostate cancer cells with enhanced cell-cycle-related transcription and attenuated androgen response. Cancer Res. 2018;78:853–64.
22. Pfister D, Nunez NG, Pinyol R, et al. NASH limits anti-tumour surveillance in immunotherapy-treated HCC. Nature. 2021;592:450–6.
23. Winograd P, Hou S, Court CM, et al. Hepatocellular carcinoma-circulating tumor cells expressing PD-L1 are prognostic and potentially associated with response to checkpoint inhibitors. Hepatol Commun. 2020;4:1527–40.
24. Finn RS, Qin S, Ikeda M, et al. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. N Engl J Med. 2020;382:1894–905.
25. Rao S, Yang X, Ohshiro K, et al. beta2-spectrin (SPTBN1) as a therapeutic target for diet-induced liver disease and preventing cancer development. Sci Transl Med. 2021;13:eabk2267.
26. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169:1327–1341.e23.
27. Roessler S, Jia HL, Budhu A, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res. 2010;70:10202–12.
28. Rebouissou S, Nault JC. Advances in molecular classification and precision oncology in hepatocellular carcinoma. J Hepatol. 2020;72:215–29.
29. Brunt EM, Tiniakos DG. Histopathology of nonalcoholic fatty liver disease. World J Gastroenterol. 2010;16:5286–96.
30. The French METAVIR Cooperative Study Group. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. Hepatology. 1994;20:15–20.
31. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375:1109–12.
32. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173:400–416.e11.
33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
34. Ye QH, Qin LX, Forgues M, et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med. 2003;9:416–23.
35. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.
36. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
37. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.
38. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell. 2019;177:1888–1902.e21.
39. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
40. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
41. Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12:477–9.
42. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48:D498–503.
43. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72.

44. Therneau TM, Grambsch PM. Modeling survival data: Extending the Cox Model. Springer New York. 2000. p. 1–350.
45. Broad Institute TCGA Genome Data Analysis Center (2016): Firehose 2016_01_28 run. http://gdac.broadinstitute.org/runs/stddata__2016_01_28/. Accessed 18 May 2021.
46. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505:495–501.
47. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
48. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96.
49. Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
50. Massalha H, Bahar Halpern K, Abu-Gazala S, et al. A single cell atlas of the human liver tumor microenvironment. Mol Syst Biol. 2020;16:e9682.
51. Yu B, Ding Y, Liao X, et al. Overexpression of PARPBP correlates with tumor progression and poor prognosis in hepatocellular carcinoma. Dig Dis Sci. 2019;64:2878–92.
52. Ha SY, Song DH, Hwang SH, et al. Expression of prothymosin alpha predicts early recurrence and poor prognosis of hepatocellular carcinoma. Hepatobiliary Pancreat Dis Int. 2015;14:171–7.
53. Kwon JH, Kim J, Park JY, et al. Overexpression of high-mobility group box 2 is associated with tumor aggressiveness and prognosis of hepatocellular carcinoma. Clin Cancer Res. 2010;16:5511–21.
54. Zhang L, Han J, Wu H, et al. The association of HMGB1 expression with clinicopathological significance and prognosis in hepatocellular carcinoma: a meta-analysis and literature review. PLoS One. 2014;9:e110626.
55. Gong Y, Zou B, Peng S, et al. Nuclear GAPDH is vital for hypoxia-induced hepatic stellate cell apoptosis and is indicative of aggressive hepatocellular carcinoma behavior. Cancer Manag Res. 2019;11:4947–56.
56. Zhang R, Gao X, Zuo J, et al. STMN1 upregulation mediates hepatocellular carcinoma and hepatic stellate cell crosstalk to aggravate cancer by triggering the MET pathway. Cancer Sci. 2020;111:406–17.
57. Ma C, Xu T, Sun X, et al. Network pharmacology and bioinformatics approach reveals the therapeutic mechanism of action of baicalein in hepatocellular carcinoma. Evid Based Complement Alternat Med. 2019;2019:7518374.
58. Lu C, Zhang J, He S, et al. Increased alpha-tubulin1b expression indicates poor prognosis and resistance to chemotherapy in hepatocellular carcinoma. Dig Dis Sci. 2013;58:2713–20.
59. Dong M, Chen J, Deng Y, et al. H2AFZ is a prognostic biomarker correlated to TP53 mutation and immune infiltration in hepatocellular carcinoma. Front Oncol. 2021;11:701736.
60. Liu-Chittenden Y, Jain M, Gaskins K, et al. RARRES2 functions as a tumor suppressor by promoting beta-catenin phosphorylation/degradation and inhibiting p38 phosphorylation in adrenocortical carcinoma. Oncogene. 2017;36:3541–52.
61. Xie W, Zhang J, Zhong P, et al. Expression and potential prognostic value of histone family gene signature in breast cancer. Exp Ther Med. 2019;18:4893–903.
62. Artegiani B, van Voorthuijsen L, Lindeboom RGH, et al. Probing the tumor suppressor function of BAP1 in CRISPR-engineered human liver organoids. Cell Stem Cell. 2019;24:927–943.e6.
63. Monti P, Menichini P, Speciale A, et al. Heterogeneity of TP53 mutations and P53 protein residual function in cancer: does it matter? Front Oncol. 2020;10:593383.
64. Zhang AW, McPherson A, Milne K, et al. Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. Cell. 2018;173:1755–1769.e22.
65. Zhang Q, He Y, Luo N, et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. Cell. 2019;179:829–845.e20.
66. Villanueva A. Hepatocellular carcinoma. N Engl J Med. 2019;380:1450–62.
67. Suva ML, Tirosh I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. Mol Cell. 2019;75:7–12.
68. Wu CG, Habib NA, Mitry RR, et al. Overexpression of hepatic prothymosin alpha, a novel marker for human hepatocellular carcinoma. Br J Cancer. 1997;76:1199–204.
69. Pinyol R, Torrecilla S, Wang H, et al. Molecular characterisation of hepatocellular carcinoma in patients with non-alcoholic steatohepatitis. J Hepatol. 2021;75:865–78.
70. Boyault S, Rickman DS, de Reynies A, et al. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology. 2007;45:42–52.
71. Hoshida Y, Villanueva A, Kobayashi M, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. N Engl J Med. 2008;359:1995–2004.
72. Caruso S, O'Brien DR, Cleary SP, et al. Genetics of hepatocellular carcinoma: approaches to explore molecular diversity. Hepatology. 2021;73(Suppl 1):14–26.
73. Gerami P, Cook RW, Wilkinson J, et al. Development of a prognostic genetic signature to predict the metastatic risk associated with cutaneous melanoma. Clin Cancer Res. 2015;21:175–83.
74. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17:257–71.
75. Kwong AJ, Ghaziani TT, Mehta N. Decreased urgency among liver transplantation candidates with hepatocellular carcinoma in the United States. Liver Transpl. 2021;28(4):725–7.
76. Alvarez M, Benhammou JN, Darci-Maher N, French SW, Han SB, Sinsheimer JS, et al. Code for single-cell level HCC cell-type survival analysis. Github. 2022. https://github.com/marcalva/hcc_sc_2022.

Publisher's Note

Supplementary Figures

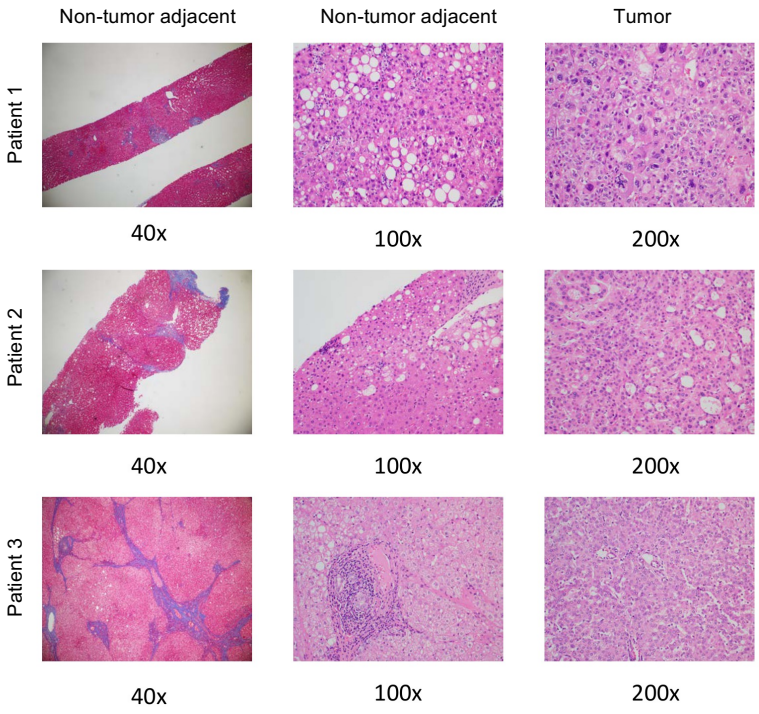


Fig. S1. Histopathology of tumor and adjacent non-tumor biopsies in the 3 NAFLD-related HCC cases.

Histopathology slides using hematoxylin and eosin and trichrome stains demonstrate tumor and patient heterogeneity.

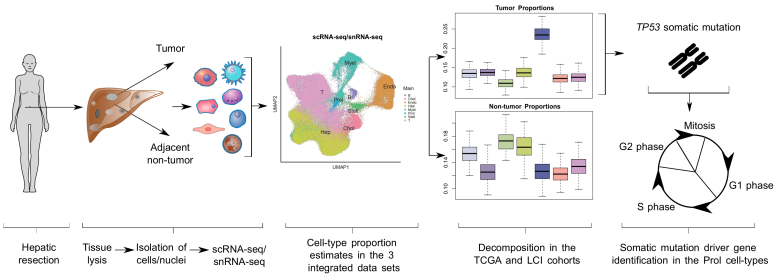


Fig. S2. Overview of study design to profile cell composition changes in HCC.

Single-cell and single-nucleus RNA-seq (scRNA-seq and snRNA-seq) were used to profile cell-type transcriptomes in human livers from non-HCC, HCC tumor, and adjacent non-tumor tissue. We performed snRNA-seq on tumor and adjacent non-tumor biopsies from three patients with fatty liver related HCC. Our snRNA-seq was integrated with two single-cell RNA-seq data sets from Aizarani *et al*. [7] and Sharma *et al.* [8] to characterize transcriptional profiles across various etiologies of HCC. The identified cell-types and their gene expression were used to estimate their proportions in larger bulk liver HCC RNA-seq cohorts with survival outcome data. These analyses highlighted the role of a tumor-associated mitotic cell-type Prol, associated with survival outcomes and *TP53* mutations.

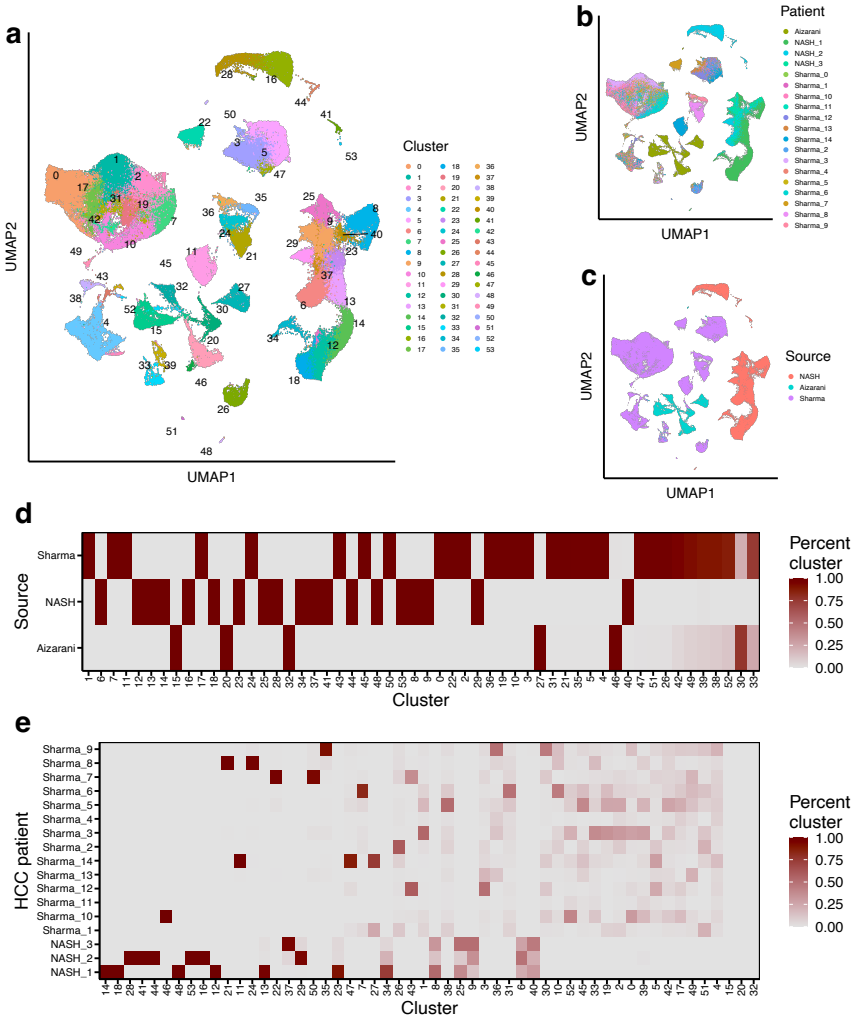


Fig. S3. Un-integrated merging of the three single cell level cohorts results in cohort- and patient-specific batch effects.

a-c, UMAP plots of the three single cell level cohorts after merging without integration. Raw counts were normalized with sctransform [37], and clustering was performed on the PCs with a resolution of 1.0. Cells and nuclei are colored by **a,** cluster, **b,** patient, and **c,** cohort (source). **d,e,** The heatmap plots show the prevalence of cohort and patient effects in the merged data without integration. Each heatmap indicates the proportion of droplets in a cluster that originate from **d,** cohort (source) and **e,** HCC patient (excluding the Aizarani *et al.* cohort [7] that comprises only healthy controls and the healthy control from the Sharma *et al*. data [8]). For each of the 54 clusters, the column proportions sum to 1. Cells and nuclei from a cohort cluster together, indicating the presence batch effects, while several clusters show patient-specific effects and suggest inter-patient heterogeneity.

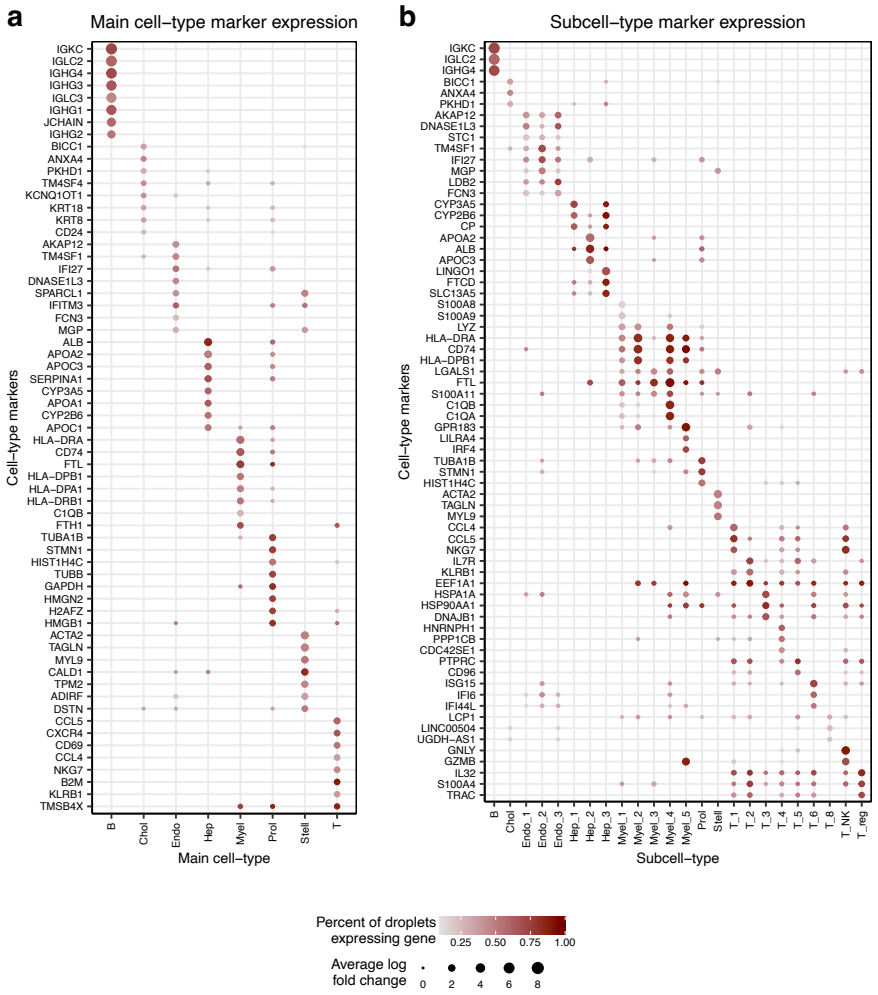


Fig. S4. Expression of top up-regulated marker genes across cell-types in the integrated single cell level data supports the functional identity of the assigned cell-types.

a,b, Expression of the top marker genes for **a,** main cell-types and **b,** subcell-types supports the functional identity of the assigned cell-types. The **a,** top 8 marker genes per main cell-type and **b,** top 3 marker genes per subcell-type are shown. A logistic regression in Seurat [38] was used to test the difference in expression between droplets in the indicated main cell-type/subcell-type and all other droplets. The percent of droplets expressing the marker gene indicates the percent which have at least one UMI aligned to the gene. The average log fold change indicates the $\log_2$ fold change of the average expression of the main cell-type/subcell-type droplets over the average expression of all other droplets. Main cell-types were assigned by merging subcell-types based on their major lineage. Cells and nuclei from T_7 contain no statistically significant marker genes.

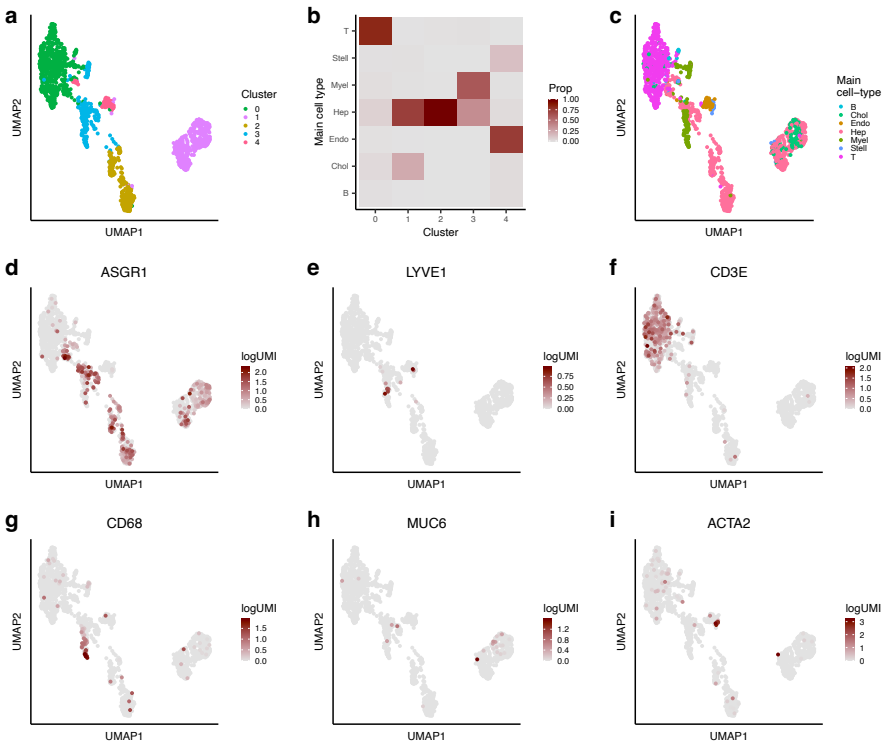


Fig. S5. Cells and nuclei from the Prol cell-type subcluster into main liver cell-types.

a, UMAP of cells and nuclei from Prol colored by subcluster. The 1,743 droplets from the Prol cluster identified in the full single-cell-level data set were subclustered after sctransform [37] and CCA integration by cohort using a resolution of 0.2 [38]. **b,** Proportion of cells/nuclei in the Prol cell-type classified into all other major cell-types. Classifications were performed using SingleR [43] with a reference trained on the full data set that excluded the Prol cluster. **c,** UMAP of Prol cells and nuclei colored by SingleR classification to all other main cell-types (consisting of 41.7%

hepatocyte, 33.8% T, 9.9% myeloid, 7.7% cholangiocyte, 4.4% endothelial, 1.6% stellate, and 0.9% B cells). **d-i,** UMAP of Prol cells/nuclei colored by log-normalized gene expression. Expression of the marker genes **d,** *ASGR1* (Hepatocyte), **e,** *LYVE1* (Endothelial), **f,** *CD3E* (T), **g,** *CD68* (Macrophage), **h,** *MUC6* (Cholangiocyte), **i,** *ACTA2* (Stellate) are shown in subclusters.

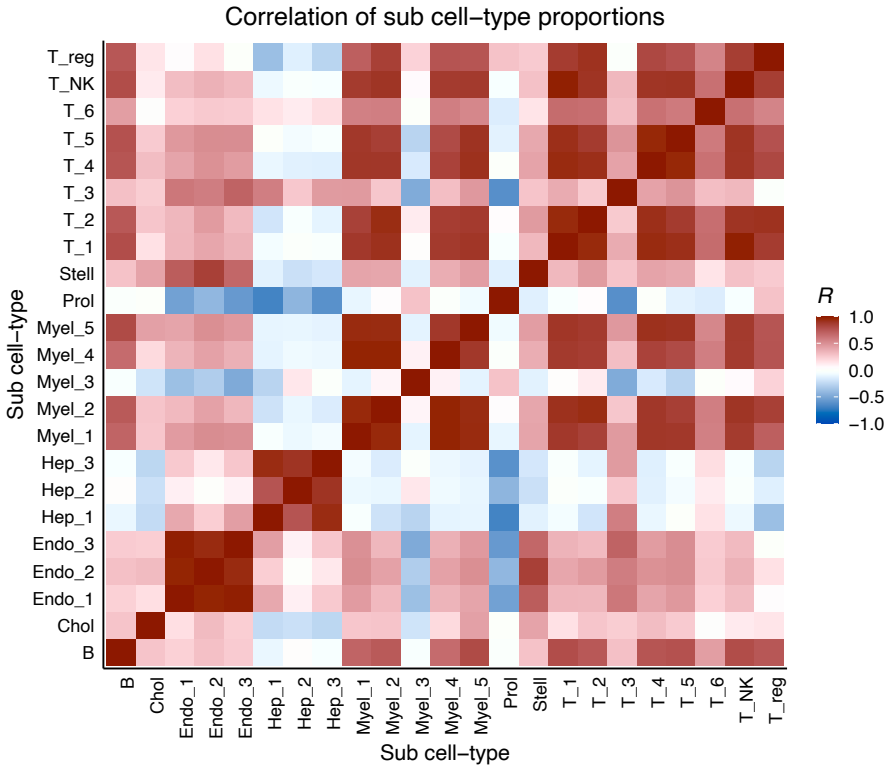


Fig. S6. Proportion estimates for sub cell-types within a main group show high correlation in TCGA.

The heatmap shows the pairwise Pearson correlation coefficients (*R*) between sub cell-type proportion estimates in TCGA. Proportions were estimated in the 410 bulk liver RNA-seq liver samples using Bisque [14]. Cell-types from the same main group (for example, hepatocytes) show high correlations.

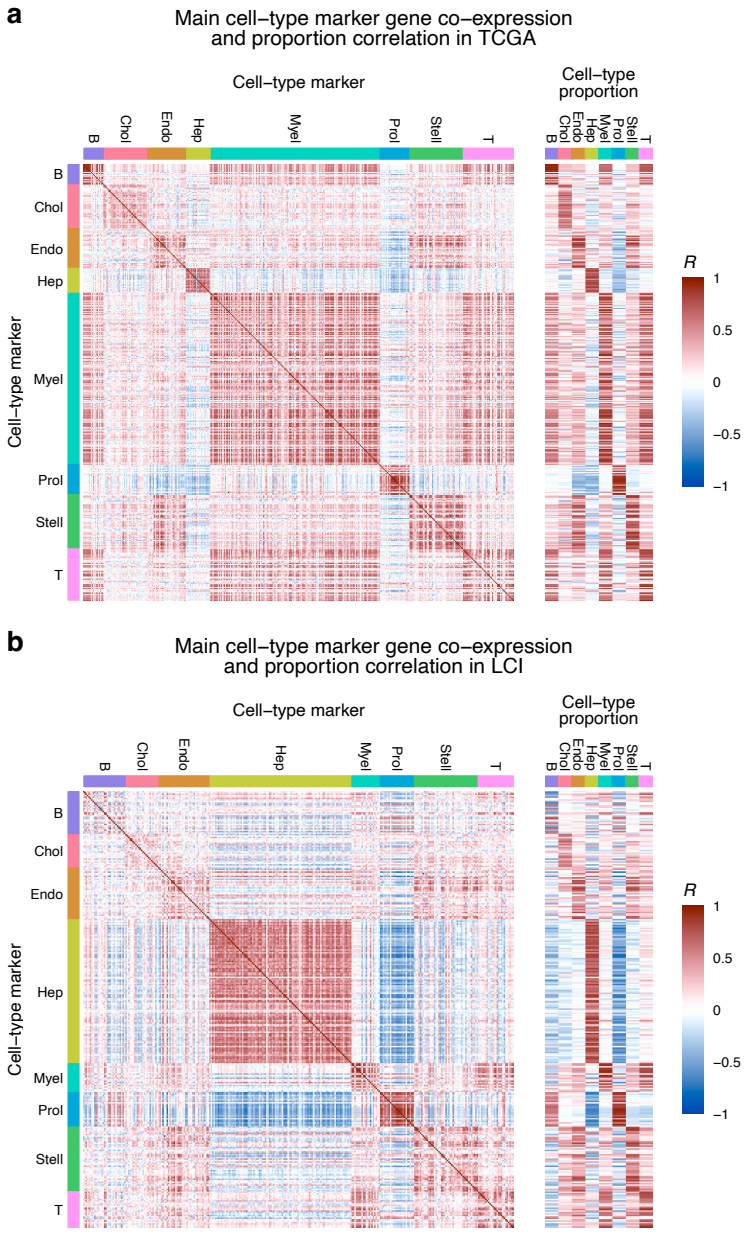

a
Main cell–type marker gene co–expression
and proportion correlation in TCGA
Cell–type marker
Cell–type
proportion
B
Chol
Endo
Hep
Myel
Prol
Stell
T
Cell–type marker
R
1
0.5
0
−0.5
−1
b
Main cell–type marker gene co–expression
and proportion correlation in LCI
Cell–type marker
Cell–type
proportion
B
Chol
Endo
Hep
Myel
Prol
Stell
T
Cell–type marker
R
1
0.5
0
−0.5
−1

Fig. S7. High intra-cell-type co-expression of main cell-type markers supports decomposed proportion estimates in TCGA and LCI.

Marker gene co-expression and proportion correlation for the main cell-types validates the reference-free approach to decompose cell-type frequency estimates. The plots show the co-expression of the top subset of marker genes ordered by cell-type as well as the expression-proportion correlations in **a,** TCGA (n=410) and **b,** LCI (n=430). Each tile displays the Pearson correlation coefficient (*R*). The left panel shows the correlation between of the expression of marker pairs, where marker genes within the same cell-type display higher co-expression than outside the cell-type. The right panel shows the correlation between the expression of marker genes and proportion estimates. The co-expressed marker genes show high correlations with their cell-type proportion estimates, validating that the proportion estimates are reflective of marker gene RNA abundance. The top subset of single cell markers and proportion estimates were calculated by Bisque [14] in the reference-free decomposition procedure. Marker genes for the B cell-type in the LCI cohort show lower intra-correlations when compared to marker gene sets of the other main cell-types, indicating that their expression is not indicative of B cell abundance.

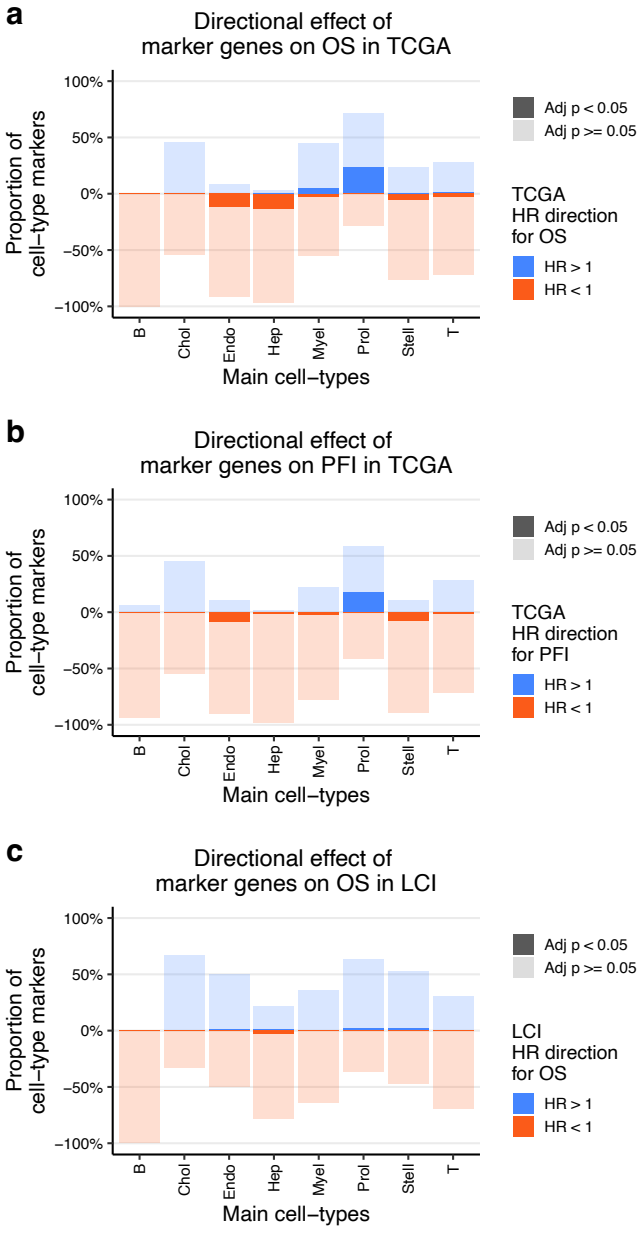

a
Directional effect of
marker genes on OS in TCGA
Proportion of
cell−type markers
100%
50%
0%
−50%
−100%
B
Chol
Endo
Hep
Myel
Prol
Stell
T
Main cell−types
Adj p < 0.05
Adj p >= 0.05
TCGA
HR direction
for OS
HR > 1
HR < 1
b
Directional effect of
marker genes on PFI in TCGA
Proportion of
cell−type markers
100%
50%
0%
−50%
−100%
B
Chol
Endo
Hep
Myel
Prol
Stell
T
Main cell−types
Adj p < 0.05
Adj p >= 0.05
TCGA
HR direction
for PFI
HR > 1
HR < 1
c
Directional effect of
marker genes on OS in LCI
Proportion of
cell−type markers
100%
50%
0%
−50%
−100%
B
Chol
Endo
Hep
Myel
Prol
Stell
T
Main cell−types
Adj p < 0.05
Adj p >= 0.05
LCI
HR direction
for OS
HR > 1
HR < 1

Fig. S8. Expression of Prol marker genes are associated with poor survival outcomes in TCGA and LCI.

a-c, The bar plots show the percent of marker genes that are positively and significantly associated with **a,** overall survival (OS) and **b,** progression free interval (PFI) in TCGA and **c,** OS in LCI. We considered marker genes as those with a log2 fold change (logFC) greater than 0.5 and an FDR-adjusted p-value less than 0.05. For each main cell-type, the percent of its marker genes that decrease survival outcomes (HR > 1) and increase survival outcomes (HR < 1) are shown by color. The percent of these genes that pass genome-wide multiple testing with an FDR-adjusted p-value less than 0.05 are shown by the darker fill for each direction.

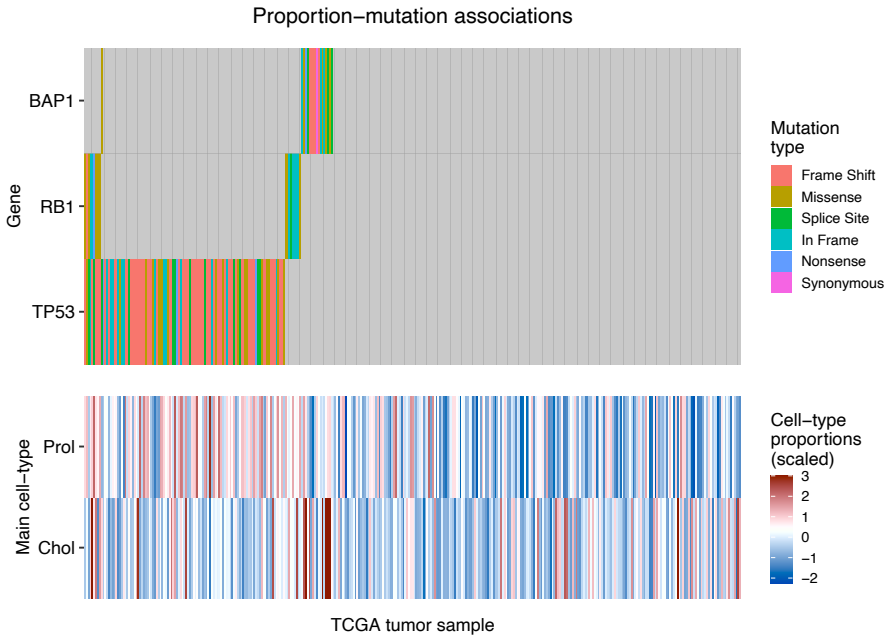


Fig. S9. Prol proportions are increased with *TP53* and *RB1* mutations.

The plot shows the proportion estimates of significantly increased cell-types (bottom) by somatic mutation (top). Proportions were tested for differences between individuals with and without a somatic mutation in significantly mutated HCC genes with a Wilcoxon test (n=357). Significant gene-cell-type pairs with an increase in proportions are shown (FDR-adjusted $p < 0.05$). The top panel shows the somatic mutation (colored by type) present in each of the 357 primary tumor samples, while the bottom panel shows their estimated cell-type proportions (scaled).

CHAPTER 5

Human adipose single nucleus RNA-sequencing reveals an adipocyte axis associated with cardiometabolic disease

5.1 Introduction

This chapter presents the current results section of our manuscript in preparation (Alvarez *et al*. manuscript in preparation). The aim of this study is to characterize how the cell-type composition of subcutaneous adipose tissue changes in obesity, insulin resistance, and hyperlipidemia that are the key risk factors for T2D and CVD. All experiments have been completed and a large portion of the analysis is finished. Not included in these results are bulk and snRNA integrations which provide further evidence to substantiate our findings. In addition, we have identified expression quantitative trait loci (eQTLs) that affect gene expression in a cell-type-specific manner, which we will investigate for the relevance to GWAS associations. These results will then form the final manuscript for publication.

5.2 Results

5.2.1 Description of cohorts

Cell-type composition in adipose tissue is altered in obesity and insulin [1, 10]. To investigate this heterogeneity in an unbiased manner, we performed single-nucleus RNA-seq on subcutaneous adipose tissue biopsies from four cohorts: the METabolic Syndrome In Men (METSIM) cohort

(n=84) [3]; a BMI-discordant monozygotic twin study (n = 13) [4, 5]; a weight-loss study [6, 7, 8] (n = 8); and random samples from individuals who underwent liposuction (n = 6). Samples from the METSIM cohort were multiplexed in random sets of 4 for snRNA-seq to maximize the final sample size. All cohorts except the liposuction cohort had data available for clinical traits. In addition, we had 335 samples from the METSIM cohort with subcutaneous adipose tissue bulk RNA-seq data. The 84 METSIM snRNA samples overlapped fully with this adipose bulk RNA-seq cohort. Findings from the snRNA-seq analyses were validated via cell-type decomposition of the bulk tissue data.

5.2.2 Single-nucleus RNA-seq of human subcutaneous fat tissue

To investigate cell-type heterogeneity in subcutaneous adipose tissue, we performed snRNA-seq. After filtering, we obtained 104,669 nuclei across a total of 111 individuals from the 4 cohorts. We normalized UMI counts with SCTransform [4] and ran CCA integration in Seurat [10] to remove batch effects. Using a coarse resolution setting, we identified 11 clusters (Figure 1A), which we manually annotated into major cell-types based on up-regulated marker genes. Cell-type assignments agreed with expression of known markers (Figure 1B) as well as pathway enrichment analyses. Perivascular and endothelial cells were assigned based on expression of *PDGFRB* and *VWF*, respectively [11, 12] (Figure 1B). The stromal cluster expressed genes involved in extracellular matrix pathways, such as *FBN1*. In addition, we identified a cluster, NTM, that expressed immune-response genes. The Ribo cluster highly expressed genes involved in translation, although this cluster also expressed immune-related genes. Finally, we observed three distinct myeloid clusters (Figure 1A,B).

Inflammation in adipose tissue occurs in obesity and insulin resistance and is thought to play a role in adipocyte function [9, 7]. To gain a better understanding of the identity of the immune clusters in our data set, we integrated the known and possible immune cells from the fat snRNA-seq data with peripheral blood mononuclear cells (PBMCs) [15]. The T, B, myeloid as well as NTM+ and Ribo cell-types were included. First, we performed CCA label transfer to assign PBMC cell-

type identities to the fat tissue immune nuclei [10]. Second, we integrated the fat and PBMC data to visualize co-clustering with UMAP. The T cells from both tissues clustered together with subsets of the fat nuclei mapping to T subcell-types, such as CD4 and CD8 (Figure 1C,D). The B cells and nuclei from both tissues also co-clustered (Figure 1C,D). The MYO1F myeloid nuclei mapped with blood CD14 and CD16 monocytes, neutrophils, and dendritic cells (Figure 1C,D). The NTM and Ribo clusters were mostly predicted to consist of blood T and NK cells (Figure 1C), although they clustered separately after integration (Figure 1D), agreeing with the lower prediction scores. Similarly, the *RBPJ* and *TPRG* expressing myeloid nuclei mapped to PBMC cell-types weakly (Figure 1C) and clustered separately (Figure 1D). In conclusion, immune cells in subcutaneous adipose tissue showed characteristics of both blood-like and tissue-resident-like cell-types.

5.2.3 Cell-type proportion associations with traits

To identify cell-types relevant to cardiometabolic traits, we performed differential abundance analysis. We ran negative binomial regression with batch and total number of nuclei as covariates. We tested 8 traits that reflect obesity, insulin resistance, and serum lipid levels. Of these, 5 cell-types were significantly associated with at least one trait (FDR-adjusted $p<0.05$) (Figure 2A). The most significant correlations were observed with tissue-resident-like *RBPJ*+ and *TPRG1*+ myeloid cell-type abundances and insulin resistance and obesity (Figure 2B). These results validate earlier findings of macrophage infiltration in obesity [10]. Adipocytes correlated negatively with obesity (Figure 2A), likely due to volume expansion of these cells in obesity [15]. Interestingly, we also observed glycated hemoglobin correlations with endothelial and *MYO1F* expressing myeloid cell-type abundances (Figure 2A). These results highlight the relevance of adipose tissue myeloid cell infiltration in metabolic health.

Given the associations between major cell-type proportions and cardiometabolic traits, we hypothesized that subcell-type variation is correlated as well. Differential expression (DE) within a major cell-type could reflect a difference in subtype composition. To this end, we performed cell-type-specific differential expression using edgeR [17]. A total of 227 genes passed multiple testing

(FDR-adjusted p<0.05), with 193 differentially expressed in adipocytes (Figure 2B). Endothelial cells had the next highest number, with 15 significant DE genes. Overall, adipocyte heterogeneity showed the strongest associations with obesity and insulin resistance.

To further investigate whether subtypes of these major cell-types are relevant to metabolic traits, we associated subtype composition with traits. Nuclei were re-clustered within each major cell-type lineage to obtain subcell-types. Then, subtype proportions were calculated within each major group separately. Finally, we associated these intra-cluster subtype proportions with traits. Of the 41 subcell-types, 8 were differentially abundant across 7 cardiometabolic traits (Figure 2C). Adipocyte subtypes showed the most significant correlations (FDR-adjusted p<0.05) for obesity and insulin resistance (Figure 2C). Additionally, two myeloid subcell-types correlated with waist circumference in opposing directions. Two stromal subcell-types also associated significantly with waist circumference. Finally, a perivascular cell-type was significantly associated with serum triglyceride levels. These results show that subcell-type composition differs in CMD, and these patterns are distinct from those observed at the major cell-type level.

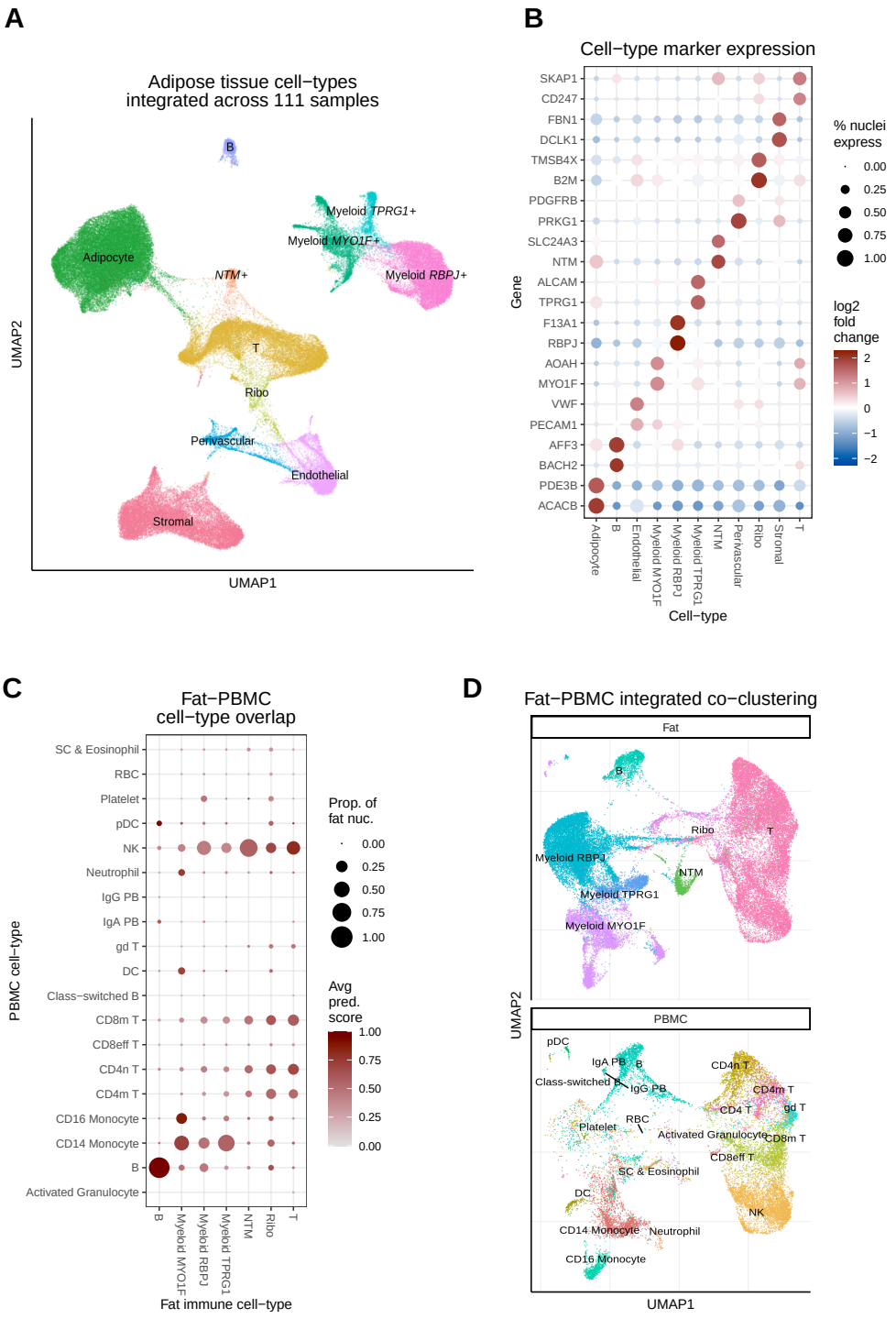


Figure 1. Integration of human subcutaneous adipose snRNA-seq identifies 11 major cell-types. We performed snRNA-seq on 111 individuals across 4 cohorts. **A,B,** Classification of

104,669 nuclei into 11 major cell-types identified using CCA integration and clustering with Seurat [10]. **A,** Uniform Manifold Approximation and Projection (UMAP) visualization of 104,669 nuclei colored by assigned cell-type. **B**, Expression of key marker genes per cell-type cluster show support and specificity of classifications. Cell-types and gene markers are indicated on the x-axis and y-axis, respectively. The color of the points shows the average $\log_2$ fold change of nuclei in the respective cell-type compared to all others. The size of the points expresses the percent of nuclei in the cell-type with non-zero counts. **C,D,** Integration of immune and immune-like fat tissue nuclei with PBMC scRNA-seq data from Wilk *et al.* [15]. **C,** Label transfer of query fat tissue nuclei onto reference PBMC cell-types assigned by Wilk *et al.* [15]. Larger and darker points indicate a higher proportion of fat tissue nuclei (x-axis) labeled to the PBMC cell-type (y-axis) with higher prediction scores. **D,** CCA integration and UMAP visualization show co-clustering of fat tissue nuclei (top) and PBMCs (bottom). The UMAPs have the same axes limits to allow visualization of nuclei adjacent in the reduced space.

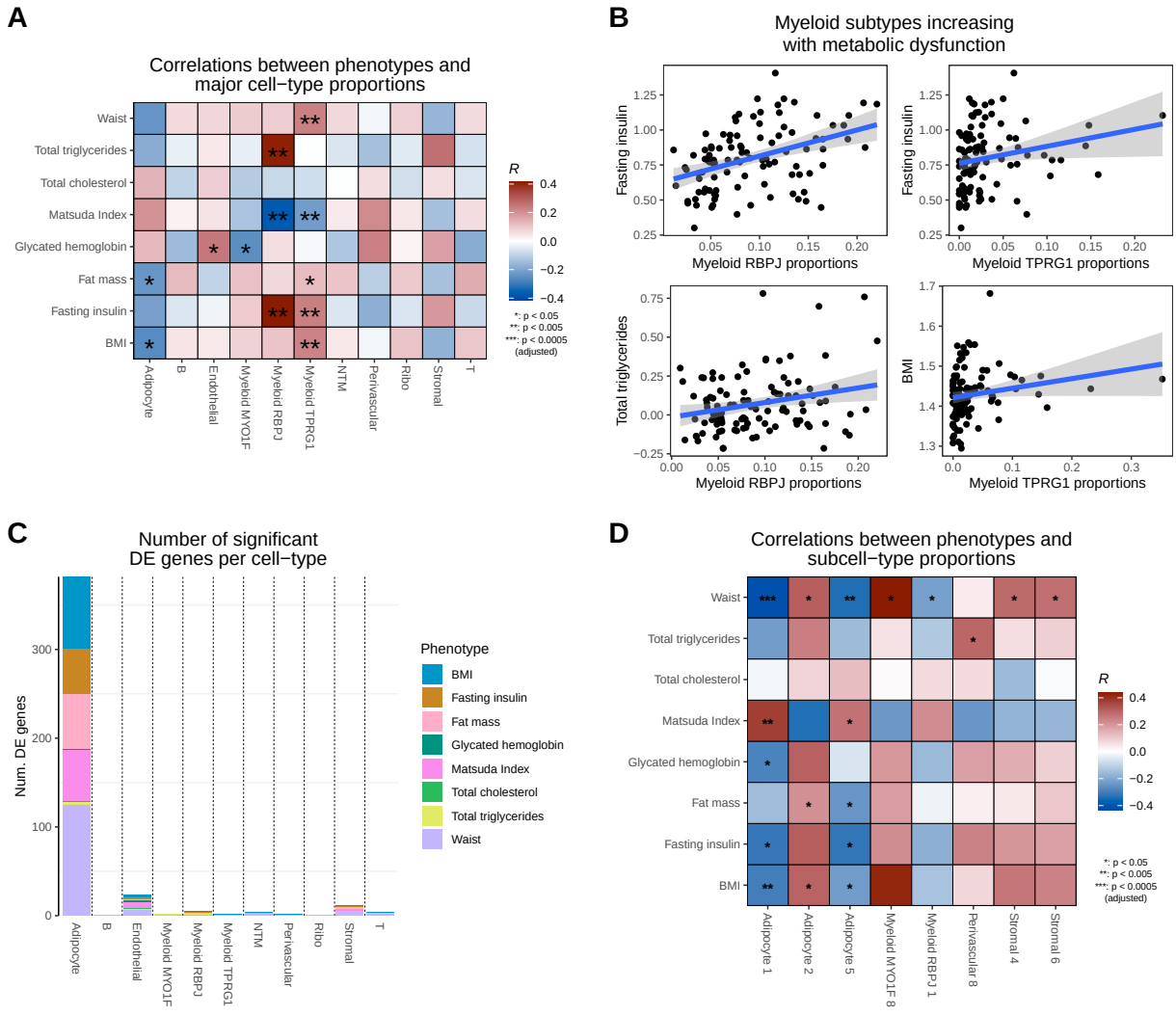


Figure 2. Subcutaneous adipose cell-type and subcell-type proportions associate with metabolic traits. Associations between adipose cell-type and subcell-type abundance from snRNA-seq and obesity, insulin resistance, and lipids. **A,B,** Correlations between proportions of the 11 adipose major cell-types and metabolic traits highlight myeloid infiltration. **A,** The heatmap shows the strength of correlation between traits and cell-type proportion. Colors denote the Pearson correlation coefficient while the asterisks indicate the significant associations determined by a negative binomial regression after correcting for multiple testing with FDR. **B,** Scatter plots showing myeloid *RBPJ+* and *TPRG1+* cell-type proportions (x-axis) increase with obesity, lipids, and insulin re-

sistance (y-axis). **C,** The bar plot gives the number of cell-type-specific differentially expressed (DE) genes per major cell-type. Cell-type pseudo-bulk gene counts were summed for each person, and edgeR was used to test for DE in the 8 traits. **D,** The strongest subcell-type-trait correlations originate from adipocytes. Subcell-types were identified by clustering within each major cell-type, and proportions of subcell-type nuclei within the major cell-type were used for differential abundance analysis. Colors denote the Pearson correlation coefficient while the asterisks indicate the significant associations determined by a negative binomial regression after correcting for multiple testing with FDR.

REFERENCES

[1] Miao, Z. *et al.* The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance. *PLOS Genetics* **16**, e1009018 (2020). URL https://doi.org/10.1371/journal.pgen.1009018.

[2] Weisberg, S. P. *et al.* Obesity is associated with macrophage accumulation in adipose tissue. *The Journal of Clinical Investigation* **112**, 1796–1808 (2003). URL https://doi.org/10.1172/JCI19246.

[3] Laakso, M. *et al.* The metabolic syndrome in men study: a resource for studies of metabolic and cardiovascular diseases. *Journal of Lipid Research* **58**, 481–493 (2017). URL https://doi.org/10.1194/jlr.O072629.

[4] Kaprio, J. Twin studies in finland 2006. *Twin Research and Human Genetics* **9**, 772–777 (2006). URL https://doi.org/10.1375/twin.9.6.772.

[5] Naukkarinen, J. *et al.* Characterising metabolically healthy obesity in weight-discordant monozygotic twins. *Diabetologia* **57**, 167–176 (2014). URL https://doi.org/10.1007/s00125-013-3066-y.

[6] Jokinen, R. *et al.* Adipose tissue mitochondrial capacity associates with long-term weight loss success. *International Journal of Obesity* **42**, 817–825 (2018). URL https://doi.org/10.1038/ijo.2017.299.

[7] Rappou, E. *et al.* Weight loss is associated with increased nad+/sirt1 expression but reduced parp activity in white adipose tissue. *The Journal of Clinical Endocrinology & Metabolism* **101**, 1263–1273 (2016). URL https://doi.org/10.1210/jc.2015-3054.

[8] Pietiläinen, K. H. *et al.* Agreement of bioelectrical impedance with dual-energy x-ray absorptiometry and mri to estimate changes in body fat, skeletal muscle and visceral fat during a 12-month weight loss intervention. *British Journal of Nutrition* **109**, 1910–1916 (2013). URL https://doi.org/10.1017/S0007114512003698.

[9] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019). URL https://doi.org/10.1186/s13059-019-1874-1.

[10] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019). URL https://doi.org/10.1016/j.cell.2019.05.031.

[11] Armulik, A. *et al.* Pericytes regulate the blood–brain barrier. *Nature* **468**, 557–561 (2010). URL https://doi.org/10.1038/nature09522.

[12] Chen, J. & López, J. A. Interactions of platelets with subendothelium and endothelium. *Microcirculation* **12**, 235–246 (2005). URL https://doi.org/10.1080/10739680590925484.

[13] Hotamisligil, G. S., Shargill, N. S. & Spiegelman, B. M. Adipose expression of tumor necrosis factor-α: Direct role in obesity-linked insulin resistance. *Science* **259**, 87–91 (1993). URL https://doi.org/10.1126/science.7678183.

[14] Longo, M. *et al.* Adipose tissue dysfunction as determinant of obesity-associated metabolic complications. *International Journal of Molecular Sciences* **20** (2019). URL https://doi.org/10.3390/ijms20092358.

[15] Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe covid-19. *Nature Medicine* **26**, 1070–1076 (2020). URL https://doi.org/10.1038/s41591-020-0944-y.

[16] Stenkula, K. G. & Erlanson-Albertsson, C. Adipose cell size: importance in health and disease. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **315**, R284–R295 (2018). URL https://doi.org/10.1152/ajpregu.00257.2017.

[17] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). URL https://doi.org/10.1093/bioinformatics/btp616.

CHAPTER 6

Conclusions, limitations, and future directions

Obesity and related CMDs are complex diseases characterized by networks of interacting cells and tissues. Consequently, obtaining a comprehensive picture of these disorders requires investigating functions within and between cells, free from the confounding effects of cell-type composition. In this thesis, I presented our work on developing and applying tools to study gene expression with cell-type resolution.

Chapter 2 presents our tool, DIEM [1], for processing droplet-based snRNA-seq or scRNA-seq data. We show that frozen tissues, common in biobanks, can result in contaminated cell and nucleus suspensions. This contamination leads to difficulties in removing empty droplets and uncertainty in downstream results. DIEM can effectively remove problematic droplets that consist of mostly contaminating RNA.

In Chapter 3, I present our approach, Bisque [2], that estimates cell-type proportions in RNA-seq data originating from bulk-tissue samples. Given the typically smaller sample sizes of single-cell data, cell-type proportion estimates in larger bulk-tissue RNA-seq cohorts can allow for higher-powered association tests. Our results recover expected correlations between cell-type abundance and traits in two independent settings and propose novel candidate cell-type-trait connections.

I present our HCC cell-type findings in Chapter 4 [3]. We integrated our snRNA-seq data with two additional studies to create an extensive cell-type reference for decomposition. Then, we used this reference to estimate cell-type abundance in bulk-tissue gene expression in two HCC cohorts. Our results revealed that cycling, i.e., proliferating, cells were enriched in tumor tissue, predictive of poor survival outcomes, and associated with mutations in *TP53*.

I describe our results from snRNA-seq analysis of adipose tissue in Chapter 5 (Alvarez *et al.* manuscript in preparation). Genetic multiplexing of samples allowed us to produce an in-depth subcutaneous adipose tissue snRNA-seq data set of over 100 individuals. Our findings included associations between subcell-type variation and cardiometabolic health, specifically obesity and insulin resistance.

The approaches developed here carry certain limitations. The main shortcoming of both DIEM and Bisque is the employment of relatively simple models that may not accurately reflect the underlying complex physical processes. In DIEM, we use a multinomial model for gene expression. However, others have shown that negative binomial models better account for the high variances observed in gene count distributions [4]. Similarly, Bisque assumes a linear relationship between read counts and cell-type abundance. More data and additional analyses are required to determine if this assumption is valid, or whether a non-linear model is more appropriate.

Other limitations exist in our HCC study. A considerable challenge in the liver cancer field is inter- and intra-tumoral heterogeneity [5]. In our work, we used a total of 17 HCC samples to extrapolate cell-type information on roughly 350 and 200 HCC samples in two cohorts. While this study was sufficiently powered to detect more common aberrations and convergent mechanisms, given the heterogeneity of HCC tumors, it is reasonable to consider that our reference may have lacked cell-types and cell states existing in other liver cancer patients. Another limitation was the inability to investigate the identified proliferating cell-type in more depth. The lower numbers of proliferating cells in our data made subcell-type discovery more challenging. Nevertheless, our results provide new insight into factors predicting poor HCC survival that can be further verified in future experimental and clinical studies.

Limitations also exist in the adipose snRNA-seq experiment. Perhaps the most significant limitation was the lack of diverse populations, as only Finnish individuals were included. Therefore, we cannot extrapolate and generalize our findings to non-European populations, such as Latinos, who exhibit a higher risk of developing multiple common CMDs [6]. A second limitation is the absence of a replication cohort. Reproducing our findings in an independent group would bolster

the validity of our results.

Several possibilities exist to develop improved methods in the future. In DIEM, modeling single-cell level gene counts with a negative binomial distribution, as an alternative for a multinomial, could further improve the accuracy of droplet classification. As exact inference is computationally intensive, variational methods could appropriately approximate maximum likelihood estimation [7]. In Bisque, the assumption of a linear relationship between gene expression and cell-type abundance may turn out to be invalid when more single cell level reference data will become available. Thus, the application of non-linear models could help improve the accuracy of cell-type proportion estimates in bulk. Neural networks provide a promising approach to fit non-linear models if the expression-abundance relationship is overly complicated [8].

Various follow-up experiments and analyses are possible for our snRNA-seq studies. The HCC results offer a global picture of tumor compositional changes. However, profiling cell states would expand our understanding of tumorigenesis even further. Persister cells, i.e., cell subpopulations that resist treatment, were recently shown to be heterogeneous [9], emphasizing the importance of studying rare cell-types. Flow sorting of cycling cells, using DNA copy number, followed by single-cell level RNA-seq could reveal distinct cell lineages with unprecedented resolution. We could further follow up our adipose snRNA-seq studies to address the limitations described above. Applying single-cell tools to diverse populations would clarify how ancestry affects cell-type and cell state composition and elucidate to what extent the single cell level results in Europeans can be generalized to non-European populations. Specifically, Latinos and African Americans would benefit most given their higher predisposition to obesity and T2D.

Additionally, complementing our RNA assays with epigenomic and spatial single-cell level data would help us prioritize cell-types. Single-nucleus ATAC-seq has the ability to reveal chromatin accessibility changes in regulatory regions that parallel gene expression differences [10]. Identifying cell-type-specific accessible chromatin regions could help fine-map obesity-associated GWAS loci by pinpointing variants more likely to bind to nuclear proteins. Single-cell technologies that powerfully map physical locations can also discover cell-type interactions based on proximity

[11]. Furthermore, detailed knowledge of the lymphoid and myeloid cells bordering the tumor will likely help us understand immune evasion in HCC. Similarly, spatial scRNA-seq in the adipose tissue will facilitate the discovery of cell-types neighboring the CMD-associated adipocyte subtypes. Overall, advances in single-cell genomic technologies are expected to provide high-resolution tools to study spatial cell composition in future experiments.

Collectively, this thesis presents single-cell RNA-seq methods and applications that expand the resolution of gene expression to the cell-type level. Our analytical tools permitted us to discover and quantify cell-types and subcell-types in tissues. Finally, we uncovered novel associations between cell abundance and CMDs. These results contribute to our understanding of obesity-related CMDs and HCC by elucidating relevant cell-types, genes, and pathways.

REFERENCES

[1] Alvarez, M. *et al.* Enhancing droplet-based single-nucleus rna-seq resolution using the semi-supervised machine learning classifier diem. *Scientific Reports* **10**, 11019 (2020). URL https://doi.org/10.1038/s41598-020-67513-5.

[2] Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications* **11**, 1971 (2020). URL https://doi.org/10.1038/s41467-020-15816-6.

[3] Alvarez, M. *et al.* Human liver single nucleus and single?cell rna sequencing identify a hepatocellular carcinoma-associated cell-type affecting survival. *Genome Medicine* **14**, 50 (2022). URL https://doi.org/10.1186/s13073-022-01055-5.

[4] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019). URL https://doi.org/10.1186/s13059-019-1874-1.

[5] Losic, B. *et al.* Intratumoral heterogeneity and clonal evolution in liver cancer. *Nature Communications* **11**, 291 (2020). URL https://doi.org/10.1038/s41467-019-14050-z.

[6] Young, K. L., Graff, M., Fernandez-Rhodes, L. & North, K. E. Genetics of obesity in diverse populations. *Current Diabetes Reports* **18**, 145 (2018). URL https://doi.org/10.1007/s11892-018-1107-0.

[7] Yang, S. *et al.* Decontamination of ambient rna in single-cell rna-seq with decontx. *Genome Biology* **21**, 57 (2020). URL https://doi.org/10.1186/s13059-020-1950-6.

[8] Alzubaidi, L. *et al.* Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data* **8**, 53 (2021). URL https://doi.org/10.1186/s40537-021-00444-8.

[9] Oren, Y. *et al.* Cycling cancer persister cells arise from lineages with distinct programs. *Nature* **596**, 576–582 (2021). URL https://doi.org/10.1038/s41586-021-03796-6.

[10] Guilhamon, P. *et al.* Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. *eLife* **10**, e64090 (2021). URL https://doi.org/10.7554/eLife.64090. ELife 2021;10:e64090.

[11] Qi, J. *et al.* Single-cell and spatial analysis reveal interaction of fap+ fibroblasts and spp1+ macrophages in colorectal cancer. *Nature Communications* **13**, 1742 (2022). URL https://doi.org/10.1038/s41467-022-29366-6.