UNIVERSITY OF CALIFORNIA

Los Angeles

Utilizing Gene Expression Data to Estimate Cell Type Abundances

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in Bioinformatics

by

Brian Benjamin Nadel

2020

ABSTRACT OF THE DISSERTATION

Utilizing Gene Expression Data to Estimate Cell Type Abundances

by

Brian Benjamin Nadel

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2020

Professor Matteo Pellegrini, Chair

The healthy function of complex tissues is dependent on a complex combination of cell types properly working together to maintain homeostasis. Diseases or stressful conditions frequently alter the normal mix of cell types found in a healthy tissue, either directly or by eliciting an immune response. The cell type composition of these tissues is, therefore, of natural interest to both researchers and clinicians.

However, quantifying cell type populations has proven to be a challenging and often expensive task. Traditional methods suffer from several limitations and have potential to introduce bias. FACS sorting has been a common approach for many years, but remains slow and expensive, making it difficult to apply to large studies. Single-cell methods are emerging and may become more cost effective in the future, but still present a prohibitive financial barrier for many labs. Moreover, both these technologies fail to capture cells with unusual morphologies. Neurons, myocytes, and adipocytes are too large, unusually shaped, or fragile to be reliably estimated by these methods.

As gene expression data has become more ubiquitous, interest in computational cell type quantification methods have gained interest and popularity. These approaches, termed cell type deconvolution, utilize knowledge of cell type specific gene expression to estimate cell type abundances in samples of unknown composition. However, gene expression deconvolution is a challenging problem, and accurate predictions are sensitive to a number of factors. Many approaches have emerged, but struggle to maintain accurate predictions when faced with novel data from varying platforms, tissue types, or species.

I have developed the Gene Expression Deconvolution Tool (GEDIT), a flexible, robust deconvolution tool that aims to overcome limitations still present in the field. GEDIT is designed to be flexible applicable to a wide range of cell types, platforms, and species. GEDIT utilizes novel techniques for selecting signature genes, which identifies genes with cell type specific expression patterns and improves the speed and accuracy of results. A transformation is also applied, in order to control for the effect of highly expressed genes and further improve quality of results. Lastly, GEDIT applies a linear regression to model the observed. I have applied GEDIT to a number of datasets, including the entire GTEx database.

In addition, I am also performing a large-scale benchmarking project, in which I compare 8 current benchmarking tools (with more to be added) on several datasets of known proportions. This includes a large clinical dataset, with over 5,000 blood samples taken directly from healthy individuals. Cell type quantification for this data has been carried out by physical means, specifically cell electrical impedance counting. This project is comprehensively evaluating the performance of these tools when used with several mixture and reference datasets.

The dissertation of Brian Benjamin Nadel is approved.

Jason Ernst

Paivi Pajukanta

Janet Sinsheimer

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2020

Table of Contents

Acknowledgements

I would like to thank Matteo Pellegrini for serving as my advisor for four years. He has displayed enormous helpfulness, patience, and understanding throughout my tenure. It is because of him that I was first exposed to computational research at UCLA, almost 7 years ago now when I was still in undergrad. He has proven a terrific guide, both in terms of research as well as career and academic choices.

I would also like to thank my other three committee members, Jason Ernst, Paivi Pajukanta, and Janet Sinsheimer. They have provided excellent advice, and I also thank them for opening their busy schedules to meet with me on many occasions. I have also had the pleasure of taking several classes taught by Dr. Ernst and Dr Sinsheimer, all three of which were wonderful learning experiences. In addition, I took one of my rotations in my first year in Dr. Pajukanta's lab, which I also very much enjoyed.

I would like to thank my co-workers in the Pellegrini lab. Collin Farrell and Feiyang Ma have been good friends the past several years, and it has been very helpful having them to consult on various matters. Dennis Montoya also provided welcome company and advice during our time in the lab together. Lastly, a special thanks to David Lopez, who helped me get started in the lab. He spent large amounts of his time helping me "learn the ropes" so to speak, and greatly eased my transition into the lab.

A final thanks to my funding sources. I was funded for two years by the Biomedical Big Data Grant, and I thank the caretakers and funding sources of that grant for selecting me and financially helping me through my time at UCLA. The other years I was funded either by my lab, or the Bioinformatics Program. I would like to thank the chair of the program, Grace Xiou, for managing the program. I would also like to thank Gene Grey for serving as SAO, and overseeing many matters including my paycheck.

Chapter 2 of this dissertation is the manuscript recently submitted to Giga Science. I was the primary researcher on this manuscript. The co-authors provided contributions in many ways, including the assembly of necessary data.

Chapter 3 of this dissertation is the manuscript in preparation for a benchmarking project, comparing many current deconvolution tools. A special thanks here to Serghei Mangul, who will be advising my postdoctoral position after I finish at UCLA. The first task in this new position is to finish and submit this project. The coauthors have provided a combination of datasets to be used in the benchmark, in addition to installation of several tools.

Chapter 4 of this dissertation is the manuscript under review at Nature. I thank Sarah Hellmuth for leading this project, and the other authors for their substantial contributions. My role in this project was applying deconvolution via GEDIT to the entire GTEX database. My contributions to the project are described in the supplementary materials.

Vita

Education:

2015-present	Ph.D. Candidate University of California, Los Angeles Bioinformatics Interdepartmental Program Los Angeles, CA
2010-2014	B.A. Biology and Computer Science (double major) Swarthmore College, Swarthmore PA Departments of Biology and Computer Science

Work Experience:

Summer 2018	Internship at Fulgent Genetics Temple City, CA
2014-2015	Research Assistant Jeff Long Lang University of California, Los Angeles Department of Molecular, Cellular, and Developmental Biology
Summer 2013	Research Assistant Matteo Pellegrini Lab University of California, Los Angeles Department of Molecular, Cellular, and Developmental Biology

Honors and Awards

2017-2018	Biomedical Big Data Training Grant Recipient UCLA Bioinformatics IDP Los Angeles, CA
2018 and 2019	Best Abstract Award UCLA Bioinformatics IDP Annual Student Retreat

Publications:

Nadel BB, Lopez D, Montoya DJ, Waddel H, Khan MM, Mangul S, Pellegrini M. The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data. Submitted to Giga Science

Nadel BB, Moutin A, Shou B, Ma F., Montoyo DJ, Pellegrini M, Mangul S. Systematic Evaluation of Current Deconvolution Tools and References. In preparation

Kim-Hellmuth S., Aguet F., Oliva M., …, **Brian Nadel**, …, Stranger B.E., Ardlie KG, Lappalainen T. Cell Type Specific Genetic Regulation of Gene Expression Across Human Tissues. In final stages of review at Science.

Nomoto, H., Pei, L., Montemurro, C., Roseberger M., Furterer A., Coppola G., **Nadel B.**, Pellegrini M., Gurlo T., Butler P.C., Tudzarova S. Activation of the HIF1α/PFKFB3 stress response pathway in beta cells in type 1 diabetes. *Diabetologia* 63, 149–161 (2020). https://doi.org/10.1007/s00125-019-05030-5

Catherine S. Grasso, Marios Giannakis, …, **Brian B. Nadel**, …, Antoni Ribas, Shuji Ogino, and Ulrike Peters. Genetic Mechanisms of Immune Evasion in Colorectal Cancer Cancer Discov May 9 2018 DOI: 10.1158/2159-8290.CD-17-1327

Brian B. Nadel, Shawn J. Cokus, Marco Morselli, Laura J. Marinelli, Robert L. Modlin, Joe Distefano III, Matteo Pellegrini. A Novel Uropathogenic Escherichia Coli Genome (strain D3) and Comparative Analysis with Other Uropathogenic and Nonpathogenic Strains. bioRxiv 2017 (not peer reviewed); doi: https://doi.org/10.1101/197533

Chapter 1

Introduction

The cell type composition of living tissues is a complex environment that is frequently of interest to researchers and clinicians alike. When researchers observe gene expression differences between sets of samples, interpretation can become difficult without knowledge of the underlying cell type populations. If a gene, set of genes, or pathway are up- or down-regulated, this may represent cells modulating their expression profiles. However, similar patterns can be observed as a result of changes in cell type populations, even if the expression profile of each cell remains roughly constant. These two scenarios can carry dramatically different implications, and researchers require tools to distinguish between them in order to advance our understanding of a wide range of biological processes.

Cell type composition is also of enormous interest in disease research. Many diseases are associated with changes in cell type profiles, and tracking these changes with greater resolution and on larger sample sizes can be key to understanding the biology of diseases. Immune cell populations are often heavily studied, since knowledge of cell type compositions can lead to valuable insights into the state of the immune system.

Cell type composition is a variable of particular interest in cancer research and treatment. The tumor microenviromnent (TME) is a complex system that can vary dramatically depending on the site and nature of the cancer. The exact state of the patient's immune system can provide important information about survivability or response to treatment (Gentles et al, 2015, Fridman et al 2012). For example, studies have shown that high levels of TH17 and CD8 T cells are associated with high survival, whereas patients with high numbers of Th2 cells or Tregs have a poorer prognosis (Senbabaoglu et al, 2016). Studies that account for cell type heterogeneity have elucidated potential drug targets (Li et al, 2016).

Depending on resources and capabilities available, researchers have used a wide variety of physical means, ranging from manual counting of cells on a plate to newer single cell methods. However, both traditional and newer methods suffer from serious limitations. Plate counting is inherently imprecise and time consuming. FACS sorting and single cell methods remain expensive and difficult to apply to larger studies.

Moreover, many of these methods introduce biases, and the predictions they over represent some cell type while underestimating others. It has been shown that subtle differences in sample preparation can dramatically change the numbers of cells captured by these technologies. Separating masses of cells into a single cell suspension is a process prone to sampling bias. The treatment necessary to create this single cell suspension frequently destroys some cells while leaving others still attached. Cells that are grouped together are often not captured by these technologies.

Therefore, computational approaches for cell type quantification offer a promising alternative. One approach is to use gene expression data from heterogeneous tissues to infer cell type composition. Each cell type generally has a specific profile, and if these profiles are known cell type composition in a mixed sample can be inferred.

Relatively simple gene set algorithms can produce cell type scores that can be compared between samples. For example, each sample will receive a score for monocytes, and the user can be confident that the sample with the highest score has the highest concentration of monocytes. However, these outputs are not comparable in an inter-cellular fashion; a higher score for monocytes compared to neutrophils does not necessarily mean that there are more monocytes than neutrophils in the sample.

Researchers have taken many approaches to solve the problem of quantitatively decomposing bulk data. Some tools take simple approaches, such as computing the mean or log mean of a set of marker genes, and treating that as the expression score for a cell type (Lopez et al, 2017, Becht *et al*, 2016). A recent approach takes this one step further by applying a transformation to these scores, in order to convert them into predicted fractions (Aran et al, 2017). Several other tools apply some form of regression to model the combination of cell types present in a mixture. These range from linear regression (Racle et al, 2017) to support vector regression (Newman et al, 2015).

Reference data is a requirement for most deconvolution tools. These data quantify the expression profiles of individual cell types, such that those cell types can be estimated in more complex mixtures. Reference data generally comes in two forms: lists of signature genes, or matrices of expression values. Studies have shown that choice of reference matrix is an important step for producing accurate results, and that some reference matrices suffer from serious flaws or biases (Vallania *et. al.*, 2018). A substantial fraction of my work has been devoted to assembling and testing reference matrices, which is documented in chapters 2 and 3.

Many deconvolution tools utilize signature genes to facilitate faster and more accurate results. Current sequencing experiments frequently capture tens of thousands of genes simultaneously, but the majority are not informative for deconvolution. A large number of genes are either “housekeeping” genes, which are expressed at roughly constant levels across all cell types. Other genes may have very low expression in all cell types present in a tissue sample. Several tools require the user to supply a list of signature genes (), but there is no general consensus on how these should be defined. Other tools do not explicitly require a list of signature genes, but their performance falters when given large reference matrices that include non-signature genes.

Over the course of my doctorate, I have developed the Gene Expression Deconvolution Tool, a new option for users that overcomes many limitations of previous tools. GEDIT is designed to be versatile, such that it returns accurate predictions for data from a variety of cell types, platforms, and even species. Moreover, it produces estimates of cell type fractions, rather than scores, which can be compared in an inter-cellular manner.

I have also assembled a library of reference data, co-published with GEDIT, which provides users of any reference-based tool with a wide range of options. These reference profiles cover a wide range of cell types, including immune and stromal. Most cell types are represented by at least 2 reference matrices. Users can therefore run deconvolution using multiple reference matrices and compare results for consistency. This serves as an effective means of verify accuracy of results, as it tests for effects specific to particular reference matrices.

Signature gene selection is a task central to many deconvolution approaches, and GEDIT applies a novel approach utilizing information entropy. Information entropy is a measure of the randomness of a probability distribution. As part of the GEDIT pipeline, the expression vector of each gene is converted to probabilities, and the information entropy is calculated. Genes with uniform expression will have very high entropy, whereas genes with expression specific to particular cell types will have low entropy. By selecting genes with low entropy, GEDIT isolates a set of signature genes that are the most informative for deconvolution.

Another problem that can arise when performing regressions on gene expression data is the drastically different scales at which different genes are expressed. Without proper correction, this can result in highly expressed genes dominating the regression and leaving lowly expressed (though often informative) genes essentially ignored. As a means of dealing with this issue, GEDIT applies to the input data a transformation which we term "row scaling". This transformation converts expression values such that each gene has the same range of expression values. In effect, this means all signature genes have the same impact on the linear regression solution.

GEDIT offers greater versatility, relative to other tools, in terms of supported platforms, tissue types, and species. We demonstrate that GEDIT provides accurate results for both microarray and RNA-seq data, blood and stromal samples, and when applied to mouse or human. In addition, we provide reference data for all these scenarios. Among this is single cell data from the Tabula Muris, which provides high quality reference data for a variety of mouse cells (The Tabula Muris Consortium, 2018)

In addition, I am also completing a project in which we evaluate the accuracy of several deconvolution tools a large datasets of known cell type proportions. In this benchmark, we comprehensively evaluate the effect of a number of reference matrices when used for each tool and each mixture. Our suite of reference data includes data from 7 sources spanning multiple platforms and a wide range of cell types. We exhaustively test every possible combination of tool and reference and compare and discuss the results. We demonstrate that proper selection of reference matrix is a non-trivial problem, and that some matrices reliably capture certain cell types but not others. The LM22 reference matrix, on average, produces the best results for simulated blood data. However, other references outperform LM22 when applied to simulated stromal data or to real blood data. No single matrix produces the best results for all tools.

An additional challenge in quantifying cell type populations is the imprecise definitions of some cell types, and distinctions between subtypes. For instance, monocytes develop into macrophages as part of the immune process. However, this does not happen instantaneously, but incrementally over a period of time. When these cells are observed during the transition, experts do not always agree on what these cells should be called. Fibroblasts, on the other hand, are a very broad group that often encompass cells derived from entirely different origins. Situations such as these frequently arise during cell type quantification.

Single cell deconvolution methods are becoming increasingly prevalent, and may drastically change the landscape of deconvolution, as well as cell type categorization in general. If single cell data continues to become more widely available, deconvolution methods that utilize single cell data (e.g.

Bisque) will offer greater resolution (Jew *et. al*., 2020). For the time being, however, bulk RNA-seq and microarray data remains more prevalent and affordable, and single cell data is not always available.

Chapter 2 of this dissertation is the manuscript for GEDIT, a tool I have developed for estimating the cell type abundances of heterogeneous tissue samples. I demonstrate in this manuscript many superior qualities of GEDIT, relative to other tools. This manuscript has recently been submitted to Giga Science.

Chapter 3 of this dissertation is the manuscript of a benchmarking paper in development. This project compares several deconvolution tools on datasets with known cell type proportions. Moreover, we also include a large set of possible reference data sources, and test every possible combination. There are two more tools we will be adding to this study (quaNTiseq and dtange; Finotello et al, 2019, Hunt et al, 2019), and we may be adding additional datasets, as well. I will be continuing this project into my next position, which is a research position at the Serghei Mangul Lab at USC.

Chapter 4 is a manuscript in the final stages of review at Science. My contribution was the application of GEDIT to the entire GTEX database, which is documented in the supplementary materials.

References

Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18: 220. doi:10.1186/s13059-017-1349-1

Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17: 218. doi:10.1186/s13059-016-1070-5

Finotello, F., Mayer, C., Plattner, C. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med* 11, 34 (2019). https://doi.org/10.1186/s13073-019-0638-6

Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours:impact on clinical outcome. Nat Rev Cancer. 2012;12: 298–306. doi:10.1038/nrc3245

Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21: 938–945. doi:10.1038/nm.3909

Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. dtangle: accurate and robust cell type deconvolution. Bioinformatics [Internet]. 2018 Nov 8 [cited 2019 Jan 17]; Available from: https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty926/5165376

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* **11,** 1971 (2020). https://doi.org/10.1038/s41467-020-15816-6

Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17: 174. doi:10.1186/s13059-016-1028-7

Lopez D, Montoya D, Ambrose M, Lam L, Briscoe L, Adams C, et al. SaVanT: a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. BMC Genomics.

2017;18: 824. doi:10.1186/s12864-017-4167-7

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457. doi:10.1038/nmeth.3337

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., & Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *ELife*, 6. https://doi.org/10.7554/eLife.26476

Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biol. 2016;17: 231. doi:10.1186/s13059-016-1092-z

Vallania F, Tam A, Lofgren S, Schaffert S, Azad TD, Bongen E, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. Nat Commun. 2018;9: 4735. doi:10.1038/s41467-018-07242-6

Chapter 2

The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data

Brian B. Nadel[1], David Lopez[1], Dennis J. Montoya[1], Hannah Waddel[3], Misha M. Khan[4], Serghei Mangul[5,6], Matteo Pellegrini[1,2]

[1]Bioinformatics Interdepartmental Degree Program, Molecular Biology Institute, Department of Molecular Cellular and Developmental Biology, and Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, CA

[2]Department of Dermatology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA

[3]Department of Mathematics, University of Utah, Salt Lake City, UT

[4]Departments of Biology and Computer Science, Swarthmore College, Swarthmore, PA

[5] Department of Clinical Pharmacy, USC School of Pharmacy

[6] The Quantitative and Computational Biology, USC Dornsife College of Letters, Arts and Sciences, University of Southern California

Abstract

The cell type composition of heterogeneous tissue samples can be a critical variable in both clinical and laboratory settings. However, current experimental methods of cell type quantification (e.g. cell flow cytometry) are costly, time consuming, and can introduce bias. Computational approaches that infer cell type abundance from expression data offer an alternate solution. While these methods have gained popularity, most are limited to predicting hematopoietic cell types and do not produce accurate predictions for stromal cell types. Many of these methods are also limited to particular platforms, whether

RNA-seq or specific microarrays. We present the Gene Expression Deconvolution Interactive Tool (GEDIT), a tool that overcomes these limitations, compares favorably with existing methods, and provides superior versatility. Using both simulated and experimental data, we extensively evaluate the performance of GEDIT and demonstrate that it returns robust results under a wide variety of conditions. These conditions include a variety of platforms (microarray and RNA-seq), tissue types (blood and stromal), and species (human and mouse). Finally, we provide reference data from eight sources spanning a wide variety of stromal and hematopoietic types in both human and mouse. This reference database allows the user to obtain estimates for a wide variety of tissue samples without having to provide their own data. GEDIT also accepts user submitted reference data, thus allowing the estimation of any cell type or subtype, provided that reference data is available.

Author Summary

The Gene Expression Deconvolution Interactive Tool (GEDIT) is a robust and accurate tool that uses gene expression data to estimate cell type abundances. Extensive testing on a variety of tissue types and technological platforms demonstrates that GEDIT provides greater versatility than other cell type deconvolution tools. GEDIT utilizes reference data describing the expression profile of purified cell types, and we provide in the software package a library of reference matrices from various sources. GEDIT is also flexible and allows the user to supply custom reference matrices. A GUI interface for GEDIT is available at http://webtools.mcdb.ucla.edu/, and source code and reference matrices are available at https://github.com/purebrawn/GEDIT.

Introduction

Cell type composition is an important variable in biological and medical research. In laboratory experiments, cell sample heterogeneity can act as a confounding variable. Observed changes in gene expression may result from changes in the abundance of underlying cell populations, rather than changes in expression of any particular cell type [1]. In clinical applications, the cell type composition of tissue biopsies can inform treatment. For example, in cancer, the number and type of infiltrating immune cells has been shown to correlate highly with prognosis ([2], [3], [4]). Moreover, patients with a large number of infiltrating T cells are more likely to respond positively to immunotherapy [5].

For many years, cell flow cytometry via FACS sorting has been the standard method of cell type quantification. More recently, single cell RNA-seq methods such as 10x Chromium, Drop-Seq, and Seq-Well have become available [6],[7]. However, both approaches suffer from significant limitations. FACS sorting is cumbersome and expensive, and some sample types require hours of highly skilled labor to generate data. Similarly, single cell RNA-seq methods remain expensive for large sample studies. Additionally, cell types such as neurons, myocytes, and adipocytes are difficult for these technologies to capture due to cell size and morphology.

Both FACS sorting and single cell methods have the potential to introduce bias, as these technologies require that tissue samples be dissociated into single cell suspensions. Many stromal cell types are tightly connected to one another in extracellular matrices. The procedures necessary to create single cell suspensions can damage some cells, while others remain in larger clusters that are not captured or sequenced. Consequently, subtle differences in sample preparation can produce dramatically different results [8]. While FACS sorting and single cell methods can produce pure samples of each cell type, the observed cell counts may not accurately represent the cell type abundances in the original sample. Tools like Cell Population Mapping and MuSiC utilize single cell reference data to perform bulk deconvolution, but requires that single cell data be available for all the cell types of interest, which is not always the case [9,10].

During the past few years, digital means of cell type quantification, often referred to as cell type deconvolution or decomposition, have become a popular complement to FACS sorting and single cell approaches. However, these methods produce approximate results that are often limited to use on particular cell types or platforms. For example, ImmuneQuant can estimate cell type fractions for immune cells only [11]. xCell can produce estimates for the 64 cell types supported by the tool, but it does not allow the inclusion of additional cell types or subtypes [12]. CIBERSORT is specifically designed for data generated from microarrays, and provides reference data only for hematopoietic cell types [13].

To overcome some of the limitations of existing cell abundance estimation tools, we present the Gene Expression Deconvolution Interactive Tool (GEDIT). GEDIT utilizes gene expression data to accurately predict cell type composition of tissue samples. We have assembled a library of reference data from 11 distinct sources and use these data to generate thousands of synthetic mixtures. In order to produce optimal results, these synthetic mixtures are used to test and refine the approaches and parameters used by GEDIT. We compare the performance of GEDIT relative to other tools using three sets of mixtures containing known cell type proportions: 12 *in vitro* mixtures of immune cells sequenced on microarrays, six RNA-seq samples collected from ovarian cancer ascites, and eight RNA-seq samples collected from blood. We also use GEDIT to deconvolute two sets of human tissue samples: 21 skin samples from patients with skin diseases, and 17,382 samples of varied tissues from the GTEx database. Lastly, we apply GEDIT to the Mouse Body Atlas, a collection of samples collected from various mouse tissues and cell types. We find that GEDIT compares favorably to other cell type deconvolution tools and is effective across a broad range of datasets and conditions.

Results

Reference Data

Reference data profiling the expression of purified cell types is a requirement for reference-based deconvolution. Methods that do not directly require reference data, such as non-negative matrix factorization, still require knowledge of expression profiles or marker genes in order to infer the identity of the predicted components. For this study, we have assembled or downloaded a set of 11 reference matrices, each containing the expression profiles of eight to 29 cell types (Table 1). These data sources span multiple platforms, including bulk RNA-seq, microarray, and single-cell RNA-seq. Complete details on the sources and assembly of these matrices are described in the methods [13–23].

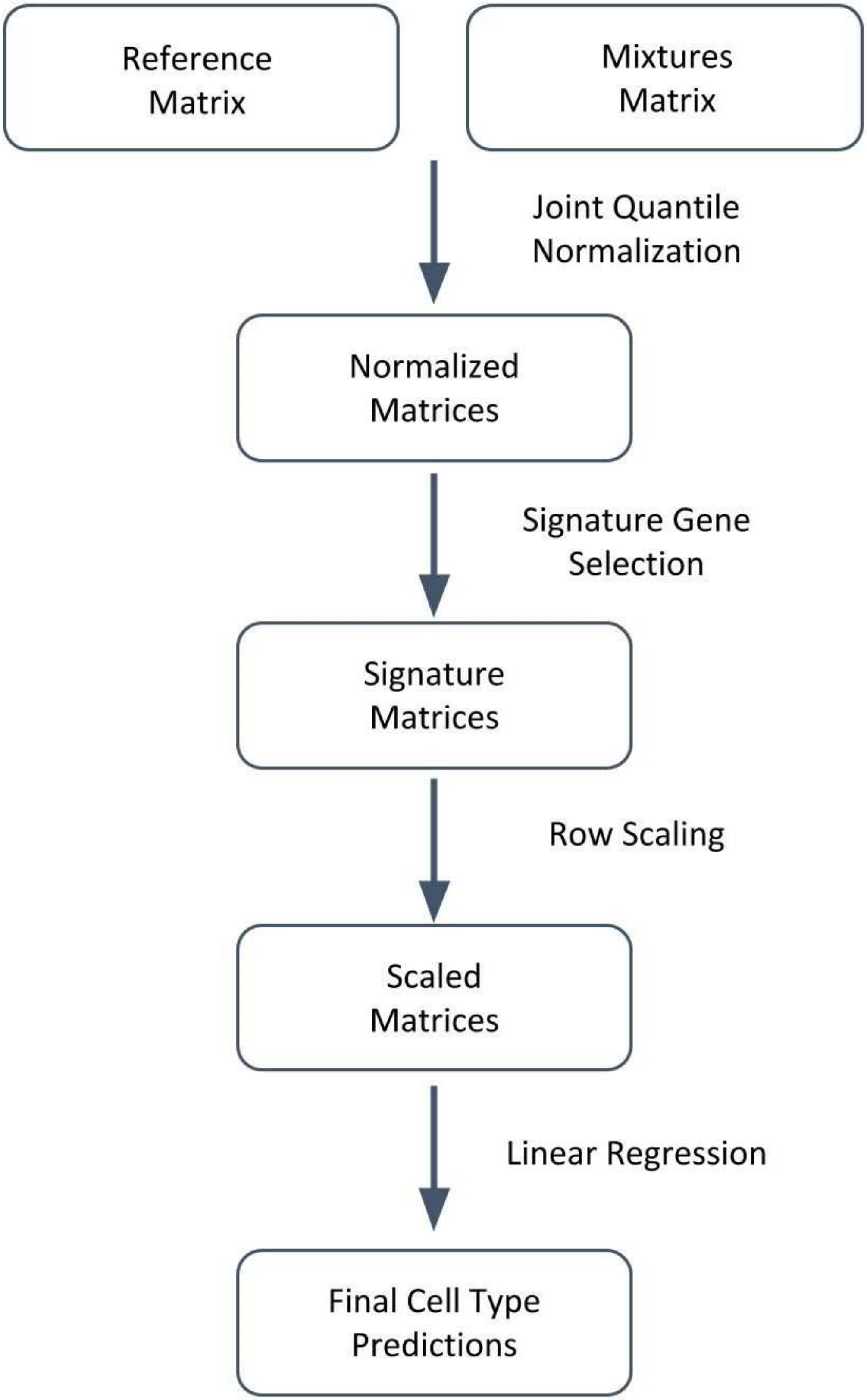
Matrix	**Species**	**Reference**	**Platform**	**# of Cell Types**	**Cell Types**
Human Skin Signatures	Human	(Swindell et al. 2013)	Multi-Microarray	21	Immune
Human Body Atlas	Human	(Su et al. 2004)	Affymetrix U133A/GNF1H	13	Immune
Human Primary Cell Atlas	Human	(Mabbott et al. 2013)	Affymetrix U133 Plus 2.0	26	Immune and Stromal
BLUEPRINT*	Human	(Martens and Stunnenberg 2013)	Bulk RNA-Seq	8	Immune
ENCODE*	Human	(ENCODE Project Consortium 2004)	Bulk RNA-Seq	29	Mostly Stromal
LM22	Human	(Newman et al. 2015)	Affymetrix Microarray	22	Immune
10x Single Cell Dataset*	Human	(Zheng et al. 2017)	Single Cell RNA-Seq	9	Immune
ImmunoStates	Human	(Vallania et. al., 2018)	Multi-Microarray	20	Immune
Tabula Muris	Mouse	(The Tabula Muris Consortium, 2018)	Single Cell RNA-seq	12	Immune and Stromal
Mouse Body Atlas	Mouse	(Lattin et al, 2008)	Affymetrix Mouse Genome 430 2.0 Array	20	Immune and Stromal
ImmGen	Mouse	(Heng et al, 2008)	Affymetrix Gene 1.0 ST	137	Immune with many subtypes

Table 1. Library of Reference Data. Asterisk denotes matrices assembled from source data as part of this project. All matrices are compatible with GEDIT and available on the GitHub repository (https://github.com/BNadel/GEDIT).

GEDIT Algorithm

GEDIT requires as input two matrices of expression values. The first is expression data is collected from the mixtures that will be deconvoluted; each column represents one mixture, and each row corresponds to a gene. The second matrix contains reference data, with each column representing a purified reference profile and each row corresponding to a gene. In a multi-step process, GEDIT utilizes the reference profiles to predict the cell type proportions of each submitted mixture (Figure 1).

Figure 1. The GEDIT pipeline. The input matrices are quantile normalized then reduced to matrices containing only signature genes. Next, a row-scaling step serves to control for the dominating effect of highly expressed genes. Lastly, linear regression is performed, and predictions of cell type abundances are reported to the user.


Reference Matrix
Mixtures Matrix
Joint Quantile Normalization
Normalized Matrices
Signature Gene Selection
Signature Matrices
Row Scaling
Scaled Matrices
Linear Regression
Final Cell Type Predictions

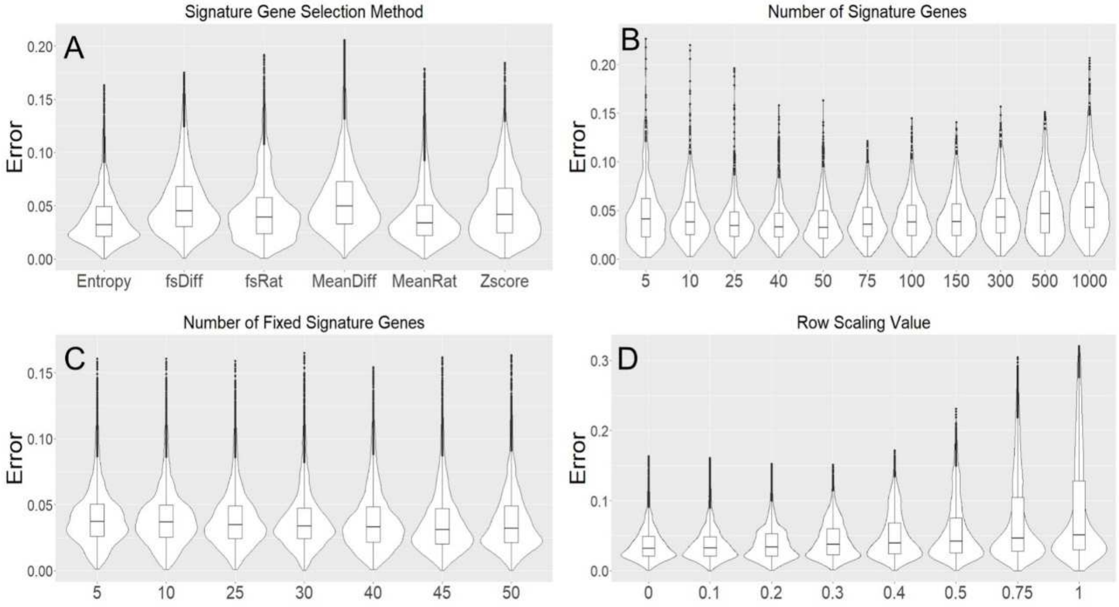
Input	Description	Allowed Values	Default Value
RefMat	Matrix of purified cell types	N by M matrix; N is number of genes, M is number of cell types	NA
MixMat	Matrix of mixtures to be deconvoluted	N by P matrix; N is number of genes, P is number of mixtures	NA
SigMeth	Method of signature gene selection	Entropy, MeanRat, MeanDiff, ZScore, fsRat, fsDiff	Entropy
NumSigs	Average number of signature genes per cell type	[1, 10,000]	50
MinSigs	Minimum number of signatures per cell type	[1,NumSigs]	=NumSigs
RowScale	Extent of per-row normalization	[0.0,1.0]	0

Table 2. GEDIT inputs include two matrices and four parameter settings. RefMat is an expression matrix documenting the expression profiles of each cell type to be estimated. MixMat is an expression matrix documenting expression values for each sample to be deconvoluted. SigMeth determines the method by which signature genes are selected. NumSigs determines the total number of signature genes, whereas MinSigs sets the minimum number of signature genes for each cell type. RowScale refers to the extent to which expression vectors are transformed to lessen the dominating effect of highly expressed genes, with a value of 0.0 representing the most extreme transformation. Default values were determined by evaluating performance on a set of synthetic mixtures (Figure 3).

Synthetic Mixture Generation and Parameter Testing

We generated a large number of synthetic mixtures *in silico* to test the efficacy of GEDIT and to assess how accuracy varies as a function of four parameter choices (SigMeth, NumSigs, MinSigs, RowScale, described in Table 2). We produced a total of 10,000 simulated mixtures of known proportions using data from four reference matrices: BLUEPRINT, The Human Primary Cell Atlas, 10x Single Cell, and Skin Signatures. We then ran GEDIT on these simulated mixtures and evaluated its performance while varying four parameter settings (Figure 2) and other design choices. Based on these results, we selected default values for each parameter (SigMeth = Entropy, NumSigs = 50, MinSigs = 50, RowScale = 0.0). Full details on the generation of these simulations are described in the supplementary materials.

Figure 2. Effect of GEDIT parameter choices on accuracy of predictions in simulated experiments. 10,000 simulated mixtures were generated, each using one of four reference matrices, with either four, five, six, or ten cell types being simulated. Deconvolution was performed using a separate expression matrix than the one used to generate the mixtures. When not otherwise noted, we use the following parameters: signature selection method = entropy; number of signatures = 50; row scaling = 0.0; and number of fixed genes = number of signatures.



Preprocessing and Quantile Normalization

The first step in the GEDIT pipeline is to render the two matrices comparable. This is done by first excluding all genes that are not shared between the two matrices. Genes that have no detected expression in any reference cell type are also excluded, as they contain no useful information for deconvolution. Both matrices are then quantile normalized, such that each column follows the same distribution as every other; this target distribution is the starting distribution of the entire reference matrix.

Signature Gene Selection

GEDIT next identifies signature genes. Gene expression experiments can simultaneously measure tens of thousands of genes, but many of these genes are uninformative for deconvolution. Specifically, genes with similar expression levels across all cell types are of little use, as observed expression values in the mixtures offer no insight into cell frequencies. Genes that are highly expressed in a subset of cell types are more informative, and we refer to these as signature genes. By using only signature genes, rather than the entire expression matrix, the problem of deconvolution becomes more tractable and less computationally intensive. Moreover, identification of signature genes can be valuable to researchers for other applications (e.g. cell type assignment for scRNA-seq data).

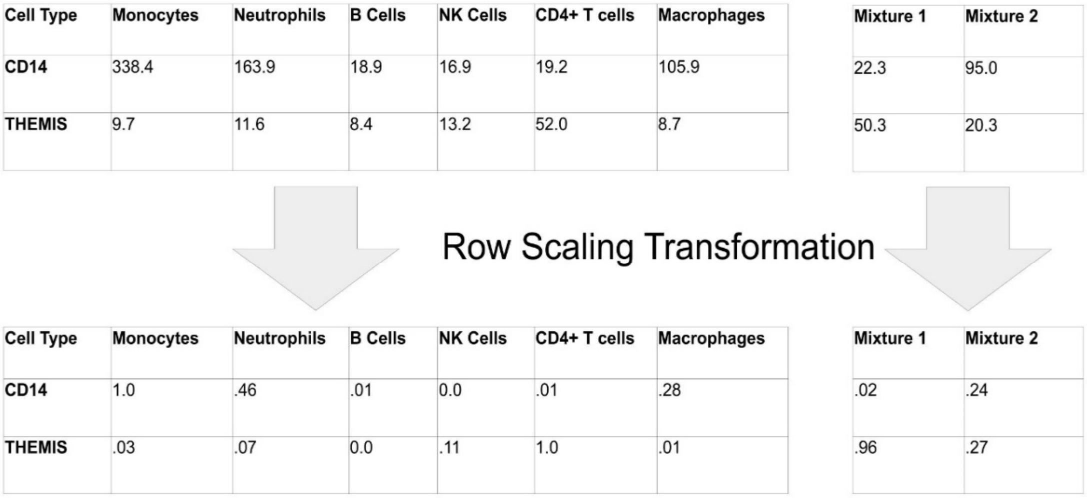
In order to identify the best signature genes in a given reference matrix, GEDIT calculates a signature score for each gene. By default, this score is computed using the concept of information entropy. Information entropy quantifies the amount of information in a probability distribution, with highly uniform distributions having the highest entropy. The expression vector for each gene (i.e. the set of expression values across all cell types in the reference) is divided by its sum, such that the entries can be interpreted as probabilities. Information entropy is then calculated according to its mathematical definition (see Methods), and genes with the lowest entropy are selected as signature genes. Entropy is minimized when expression is detected only in a single cell type and maximized when expression values are equal across all cell types. Thus, by selecting genes with low entropy, we favor genes that are expressed in a cell type specific manner. By default, 50 signature genes are selected for each cell type in the reference matrix. We chose 50 signature genes, and entropy as our scoring method, because it returned optimal results when run on 10,000 synthetic mixtures (see Figure 2).

We also evaluated the effect of accepting more signature genes for some cell types than others, depending on how many genes have low entropy. In this scheme, on average 50 signature genes are used per cell type. However, a fourth parameter is used, which specifies the minimum number of signature genes per cell type. After these have been selected, remaining signature genes are added based only on lowest entropy, regardless of cell type of maximal expression. We found that this parameter had minimal effect on accuracy, when applied to synthetic mixtures (Figure 2c). Therefore, this option is not used by default, though it can be specified by the user.

Row Scaling

One complication in the application of linear regression to gene expression data is the drastically different scale at which some genes are expressed. For example, CD14 and THEMIS (Figure 3) have both been identified as signature genes: CD14 for monocytes and THEMIS for CD4+ T cells. However, CD14 is expressed at much higher levels in most cell types and will have a larger impact on the estimation of cell type composition, relative to THEMIS. In other words, the possible penalty resulting from a poor fit of CD14 is much larger than the penalty from a poor fit of THEMIS.

Figure 3. The "row scaling" transformation, as implemented by GEDIT. CD14 and THEMIS are two examples of signature genes with drastically different magnitudes of expression. CD14 is a signature gene for monocytes, and THEMIS for CD4+ T cells. The original expression vectors are transformed, such that all values fall between 0.0 and 1.0, equalizing the effect of genes with varying magnitudes of expression.



Cell Type	Monocytes	Neutrophils	B Cells	NK Cells	CD4+ T cells	Macrophages
CD14	338.4	163.9	18.9	16.9	19.2	105.9
THEMIS	9.7	11.6	8.4	13.2	52.0	8.7

Mixture 1	Mixture 2
22.3	95.0
50.3	20.3

Cell Type	Monocytes	Neutrophils	B Cells	NK Cells	CD4+ T cells	Macrophages
CD14	1.0	.46	.01	0.0	.01	.28
THEMIS	.03	.07	0.0	.11	1.0	.01

Mixture 1	Mixture 2
.02	.24
.96	.27

In order to equalize the effect of each signature gene on the linear regression, we implement a transformation that we term row scaling. Specifically, the range of all observed values for a particular gene (including reference cell types and samples) is adjusted such that the maximum value is 1.0 and the minimum value is 0.0. As a result, all genes have a comparable influence on the calculation of the linear regression solution, regardless of overall magnitude of expression. This transformation can be modulated by adjusting the row scaling parameter. By default, the value of this parameter is 0.0, and the transformation is applied as described above. Values between 0.0 and 1.0 are also allowed, which reduces the extent of the transformation (see Methods for details).

Linear Regression:

Non-negative linear regression was performed using the glmnet package in R. The glmnet function is used with lower.limits=0, alpha=0, lambda=0, intercept=FALSE. These settings perform a linear regression where all coefficients are non-negative, and with no regularization and no intercept term.

GEDIT Compares Favorably with Other Tools

					Reference Data Provided with Tool		
Tool	**Publication**	**Custom Reference**	**Approach**	**Output**	**Number of Datasets**	**Cell Types**	**Species**
GEDIT	Nadel et. al., 2020	Yes	Deconvolution	Predicted Fractions	11	Immune and Stromal	Human, Mouse
Cibersort	Newman et. al., 2015	Yes, if marker genes specified	Deconvolution	Predicted Fractions	1	Immune	Human
xCell	Aran et. al., 2017	No	Marker Genes	Predicted Fractions	5	Immune and Stromal	Human
dtangle	Hunt et. al., 2018	Yes, if marker genes specified	Deconvolution	Predicted Fractions	0	N/A	N/A

DeconRNASeq	Gong et. al., 2013	Yes	Deconvolution	Predicted Fractions	0	N/A	N/A
DCQ/ImmQuant	Altboum et. al., 2014; Frishberg et. al., 2016	Yes	Deconvolution	Scores	3	Immune	Human, Mouse
Cibersort (absolute mode)	Newman et. al., 2015	Yes, if marker genes specified	Deconvolution	Scores	1	Immune	Human
SaVant	Lopez et. al., 2017	Yes, if marker genes specified	Marker Genes	Scores	12	Immune and Stromal	Human, Mouse
MCP-Counter	Becht et. al., 2016	No	Marker Genes	Scores	N/A	Immune and Stromal	Human

In order to assess the performance of GEDIT relative, we perform a benchmarking experiment comparing GEDIT to 4 other deconvolution tools (CIBERSORT, DeconRNASeq, dtangle and xCell; [12,13,24,25]). The authors have tried to carry out the benchmarking work in an unbiased manner, but it must be noted that this has been carried out in parallel with the development of the tool. Non-deconvolution tools like MCP-counter, SAVANT, and the DCQ algorithm are excluded from this benchmark because they do not predict cell type fractions [26–28].Tools that require single cell data, such as MuSiC and CPM, are also excluded, as this study is limited to tools that operate on bulk expression data [9,10]). See Table 3 for a summary of current bulk deconvolution methods.

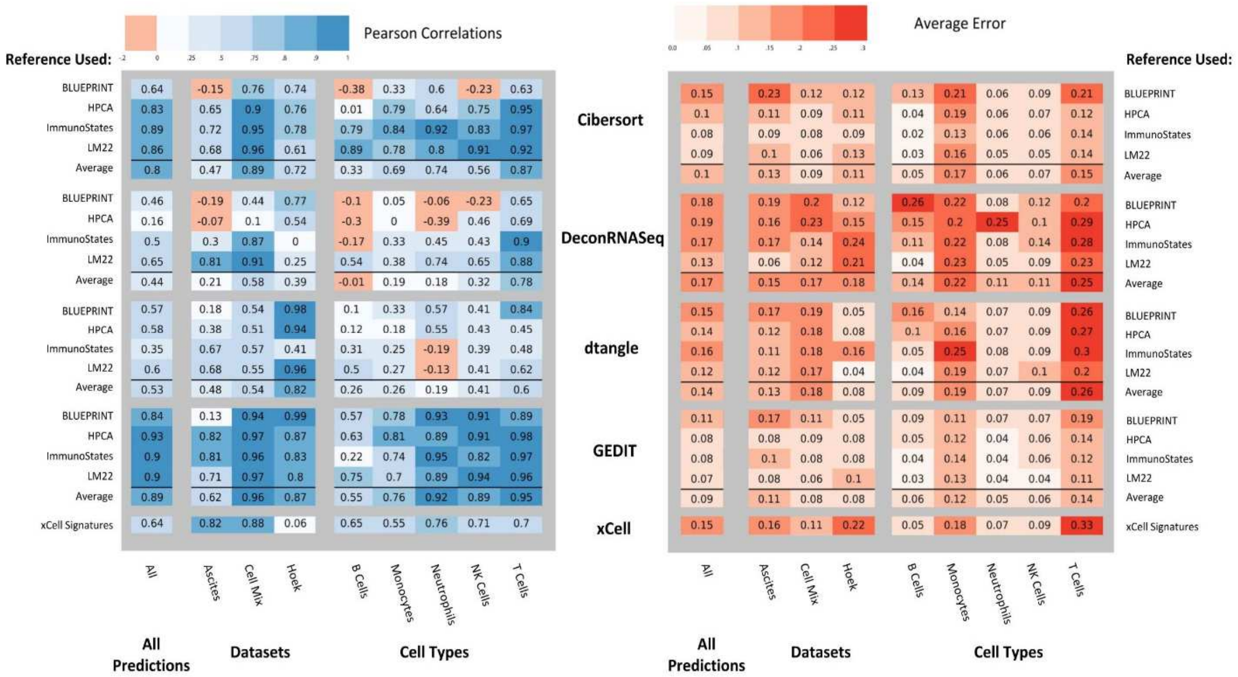
Table 3. High level characteristics of current cell type estimation tools. Some tools accept custom references, which allows the tool to estimate the abundance of cell types not present in the default reference. Tools listed here take one of two approaches: they either perform deconvolution (most commonly regression) or calculate a score based on intensity of marker gene expression. Depending on the tool, the output can be interpreted as fractions corresponding to the abundance of each cell type, or as scores for each cell type that cannot necessarily be compared in an inter-cellular manner.

To perform this benchmark, we utilize three datasets for which cell type fractions have been estimated using orthogonal methods. Two of these datasets were used in a recent benchmarking study

[29]. Both are profiled using RNA-seq, and represent samples collected either from human cancer ascites or human blood [30,31]. In both cases, cell type fractions have been evaluated by FACS sorting. The final dataset was prepared *in vitro* and consists of six cell types that were physically mixed together (in known proportions) to prepare 12 mixtures. These mixtures were then profiled using an Illumina HT12 BeadChip microarray. Adding to the previous benchmarking study, we also explore the effect of using four separate reference datasets: The Human Primary Cell Atlas, LM22, ImmunoStates, and a reference constructed from BLUEPRINT data. For each dataset, all tools (except xCell) were run four times, each time using a different reference matrix.

Compared to the other tools, GEDIT produces the most robust and consistently accurate results (Figure 4). For many tools, the quality of predictions varies greatly depending on the cell type, dataset, or choice of reference matrix. When results are averaged across the four possible reference choices, GEDIT produces the minimum error and maximum correlation for all three datasets. This result suggests that GEDIT is the best choice when researchers are using novel references matrices that have not been curated or tested.

Figure 4. Performance of five deconvolution tools when run on a set of 26 samples from three sources. Error and correlation between actual and predicted cell type fractions; calculated for each of the five cell types represented, for each of the three data sources analyzed, and for all predictions regardless of cell type or data source. Underlying cell type fractions are evaluated via FACS (ascites and blood) or by controlled mixing of purified cell types (CellMix).



Tool	Reference Used:	All	Ascites	Cell Mix	Hoek	B Cells	Monocytes	Neutrophils	NK Cells	T Cells
Cibersort	BLUEPRINT	0.64	-0.15	0.76	0.74	-0.38	0.33	0.6	-0.23	0.63
	HPCA	0.83	0.65	0.9	0.76	0.01	0.79	0.64	0.75	0.95
	ImmunoStates	0.89	0.72	0.95	0.78	0.79	0.84	0.92	0.83	0.97
	LM22	0.86	0.68	0.96	0.61	0.89	0.78	0.8	0.91	0.92
	Average	0.8	0.47	0.89	0.72	0.33	0.69	0.74	0.56	0.87
DeconRNASeq	BLUEPRINT	0.46	-0.19	0.44	0.77	-0.1	0.05	-0.06	-0.23	0.65
	HPCA	0.16	-0.07	0.1	0.54	-0.3	0	-0.39	0.46	0.69
	ImmunoStates	0.5	0.3	0.87	0	-0.17	0.33	0.45	0.43	0.9
	LM22	0.65	0.81	0.91	0.25	0.54	0.38	0.74	0.65	0.88
	Average	0.44	0.21	0.58	0.39	-0.01	0.19	0.18	0.32	0.78
dtangle	BLUEPRINT	0.57	0.18	0.54	0.98	0.1	0.33	0.57	0.41	0.84
	HPCA	0.58	0.38	0.51	0.94	0.12	0.18	0.55	0.43	0.45
	ImmunoStates	0.35	0.67	0.57	0.41	0.31	0.25	-0.19	0.39	0.48
	LM22	0.6	0.68	0.55	0.96	0.5	0.27	-0.13	0.41	0.62
	Average	0.53	0.48	0.54	0.82	0.26	0.26	0.19	0.41	0.6
GEDIT	BLUEPRINT	0.84	0.13	0.94	0.99	0.57	0.78	0.93	0.91	0.89
	HPCA	0.93	0.82	0.97	0.87	0.63	0.81	0.89	0.91	0.98
	ImmunoStates	0.9	0.81	0.96	0.83	0.22	0.74	0.95	0.82	0.97
	LM22	0.9	0.71	0.97	0.8	0.75	0.7	0.89	0.94	0.96
	Average	0.89	0.62	0.96	0.87	0.55	0.76	0.92	0.89	0.95
xCell	xCell Signatures	0.64	0.82	0.88	0.06	0.65	0.55	0.76	0.71	0.7

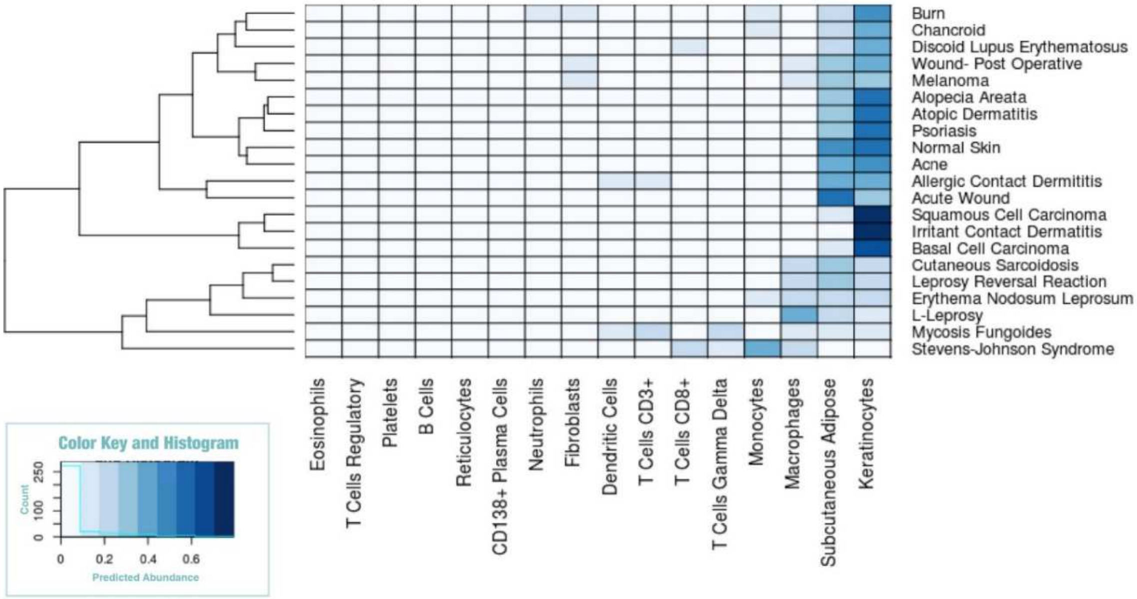
Tool	All	Ascites	Cell Mix	Hoek	B Cells	Monocytes	Neutrophils	NK Cells	T Cells	Reference Used:
Cibersort	0.15	0.23	0.12	0.12	0.13	0.21	0.06	0.09	0.21	BLUEPRINT
	0.1	0.11	0.09	0.11	0.04	0.19	0.06	0.07	0.12	HPCA
	0.08	0.09	0.08	0.09	0.02	0.13	0.06	0.06	0.14	ImmunoStates
	0.09	0.1	0.06	0.13	0.03	0.16	0.05	0.05	0.14	LM22
	0.1	0.13	0.09	0.11	0.05	0.17	0.06	0.07	0.15	Average
DeconRNASeq	0.18	0.19	0.2	0.12	0.26	0.22	0.08	0.12	0.2	BLUEPRINT
	0.19	0.16	0.23	0.15	0.15	0.2	0.25	0.1	0.29	HPCA
	0.17	0.17	0.14	0.24	0.11	0.22	0.08	0.14	0.28	ImmunoStates
	0.13	0.06	0.12	0.21	0.04	0.23	0.05	0.09	0.23	LM22
	0.17	0.15	0.17	0.18	0.14	0.22	0.11	0.11	0.25	Average
dtangle	0.15	0.17	0.19	0.05	0.16	0.14	0.07	0.09	0.26	BLUEPRINT
	0.14	0.12	0.18	0.08	0.1	0.16	0.07	0.09	0.27	HPCA
	0.16	0.11	0.18	0.16	0.05	0.25	0.08	0.09	0.3	ImmunoStates
	0.12	0.12	0.17	0.04	0.04	0.19	0.07	0.1	0.2	LM22
	0.14	0.13	0.18	0.08	0.09	0.19	0.07	0.09	0.26	Average
GEDIT	0.11	0.17	0.11	0.05	0.09	0.11	0.07	0.07	0.19	BLUEPRINT
	0.08	0.08	0.09	0.08	0.05	0.12	0.04	0.06	0.14	HPCA
	0.08	0.1	0.08	0.08	0.04	0.14	0.04	0.06	0.12	ImmunoStates
	0.07	0.08	0.06	0.1	0.03	0.13	0.04	0.04	0.11	LM22
	0.09	0.11	0.08	0.08	0.06	0.12	0.05	0.06	0.14	Average
xCell	0.15	0.16	0.11	0.22	0.05	0.18	0.07	0.09	0.33	xCell Signatures

The optimal choice of reference matrix varies greatly depending on the exact combination of tool, dataset, and cell type. While using LM22 often produces the most accurate results, there are many exceptions. For instance, DeconRNASeq and GEDIT produce their best results for the blood dataset when using the BLUEPRINT reference. For the ascites data, several tools prefer ImmunoStates as the optimal reference choice. The best choice of reference is highly dependent on the nature of the input data and on the tool being used. In practice, researchers may wish to perform deconvolution multiple times--in each case using a separate reference matrix--and compare results for consistency.

Skin Expression Data

We further validate GEDIT by using it to deconvolute a set of skin biopsies from humans with a variety of skin diseases [13]. The exact cell type composition of these samples is unknown, but we have reasonable expectations based on skin and disease biology. For example, macrophages are known to be abundant in granulomas of leprosy legions, and Steven-Johnson Syndrome produces blisters that fill with large numbers of monocytes [32,33]. We find that, in all cases, predictions made by GEDIT conform well with these biological expectations. Keratinocytes are highly predicted in most cases, as one would expect with skin samples (Figure 5). Deviations from this pattern correspond with disease biology. Monocytes are highly predicted in Stevens-Johnson syndrome, as are macrophages in the three leprosy samples, and T cells in the Mycosis Fungoides (T cell lymphoma) sample.

Figure 5. GEDIT predictions for 21 samples of various skin diseases. GEDIT correctly identifies keratinocytes and subcutaneous adipose as the most common cell. Deviations from this pattern correspond to disease biology. SJS represents blister fluid from Steven Johnson Syndrome, and is predominantly immune cells. LL and RR represent two forms of leprosy, which result in large numbers of macrophages. MF is a T Cell Lymphoma.

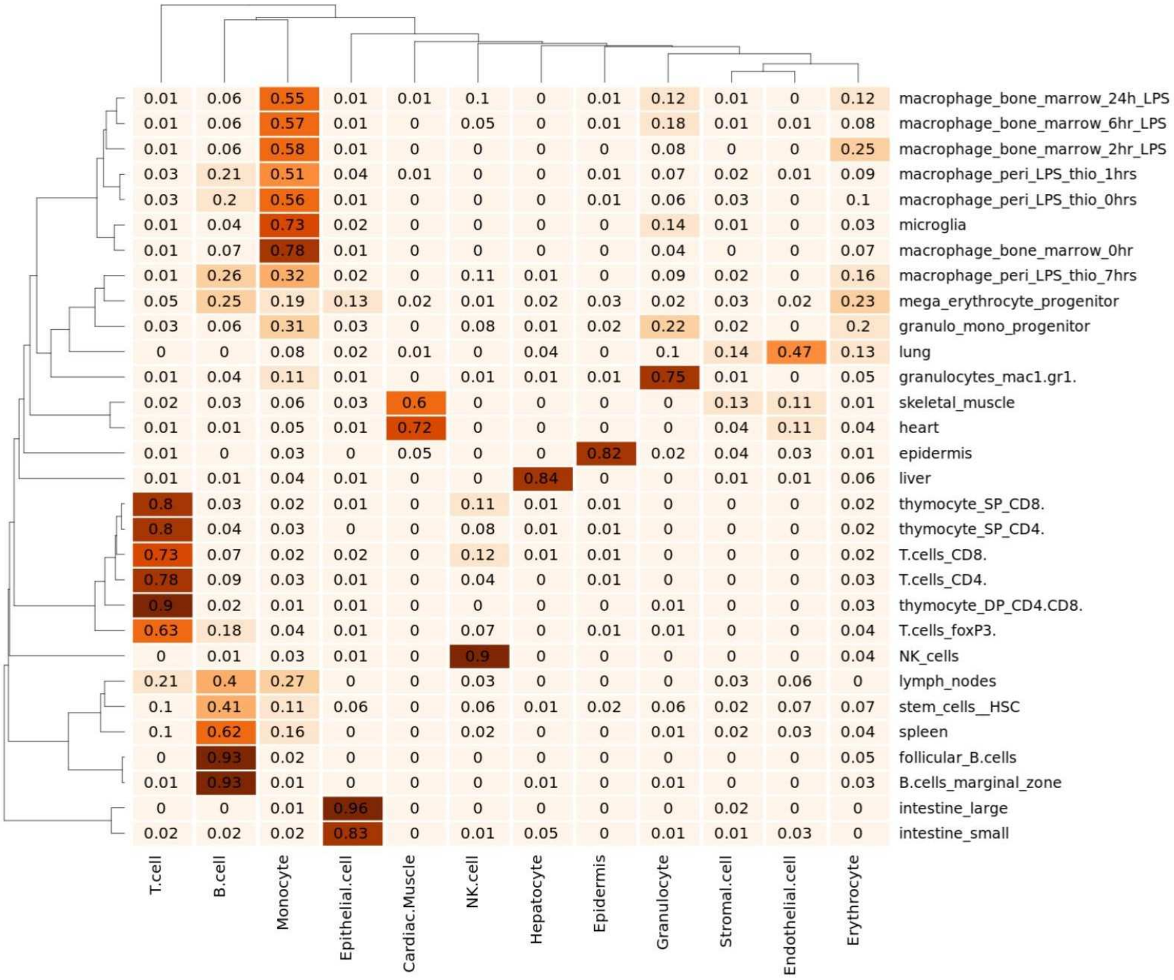


Application of GEDIT to Mouse Data

Unlike tools specifically designed for human data, GEDIT can be used to decompose data from any organism for which reference data is available. Here, we demonstrate the efficacy of GEDIT when applied to the Mouse Body Atlas, a collection of tissue and cell type samples collected from mice [22]. As reference data, we assembled a matrix of 12 cell types using single cell data from the Tabula Muris [20].

GEDIT correctly infers the identity of purified cell types, including six samples that consist of either pure NK cells, B cells, T cells, or granulocytes. An entry for macrophages is not available in the reference used, but most macrophage samples are identified as monocytes, which is the most similar cell type present in the reference matrix. For more complex tissues, GEDIT predicts cell type fractions that correspond to the biology of the samples. Hepatocytes are predicted to be highly prevalent in the liver sample (84%) and are not predicted in any other sample (less than 5% in all cases). Similar patterns hold for keratinocytes in the epidermis, epithelial cells in two intestinal samples and cardiac muscle cells in heart and muscle samples.

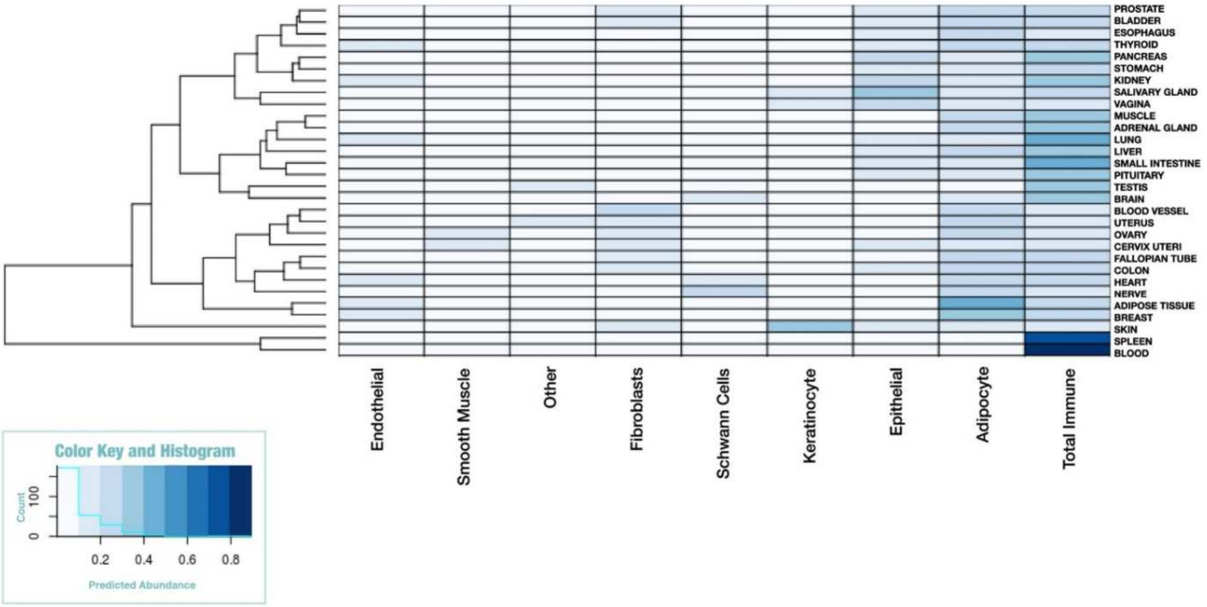
Figure 6. GEDIT predictions on 30 samples collected from various mouse tissues and cell types (mouse body atlas [22]). Predictions largely conform with tissue and cell biology.



Deconvolution of GTEx Database

To assess the use of GEDIT across very large datasets, we applied the tool to 17,382 GTEx RNA-seq samples collected from various tissues. However, no single reference contained all relevant cell types. For example, none of the available references contain both myocytes and adipocytes (Supplementary Figure 1). Therefore, we predicted proportions three times using three separate references (BlueCode, Human Primary Cell Atlas, Skin Signatures). We then combined these outputs by taking their median value. This allowed us to produce predictions spanning a larger number of cell types than are present in any one reference matrix (Figure 6).

Figure 7. GEDIT cell type predictions when applied to 17,382 samples from the GTEx database. Here, predictions have been averaged for each tissue of origin.



These predictions largely conform to biological expectations. For example, immune cells are predicted to have high abundance in blood and spleen, adipocytes in adipose tissue, Shwann cells in nerve and heart, and keratinocytes in skin. Each of these patterns matches expectations of which cell types should be present in these tissues. Neither cardiac myocytes nor smooth muscle are highly abundant in GTEx muscle samples. This is likely because the GTEx samples are collected from skeletal muscle, which is known to have an expression profile that is distinct from that of cardiac and smooth muscle.

GEDIT Availability

GEDIT can be run online at http://webtools.mcdb.ucla.edu/. Source code, associated data, and relevant files are available on GitHub at https://github.com/BNadel/GEDIT. We provide access to the tool, a set of varied reference data, and two sample mixture matrices. The website automatically produces a heatmap of predicted proportions for the user, as well as a .tsv file. The user also has access to the parameter choices of GEDIT (signature gene selection method, number of signature genes, row scaling).

Methods

GEDIT Algorithm

Signature Gene Selection

During signature gene selection, we automatically exclude genes with zero detected expression in half or more of cell types. Further, we treat all remaining expression values of zero as the lowest observed non-zero value in the matrix. Implementing this change has minimal effect on most genes but helps to reduce the scores of very lowly expressed genes. Such lowly expressed genes are highly susceptible to experimental noise and are generally poor signature genes. Moreover, including zeros can result in mathematical errors (e.g. dividing by zero, taking the log of zero). We consider this transformation valid, since values of zero generally do not mean zero expression, but rather an expression level below the detection limit of the technology used.

For any given gene, a scoring method takes as input the vector of the expression values across all reference cell types, and outputs a score. A gene is considered a potential signature gene in cell type X if it is expressed more highly in X than any other cell type. For each cell type, we keep only the N genes with the highest scores, where N is the NumSigs parameter.

Information entropy (H) is calculated using the following formula:

$$H = -\sum_{1}^{i}[p_i * \log_2(p_i)] \qquad (1)$$

where p_i is the probability of the i^{th} observation. To apply this to expression values, we convert the vector of expression values into a vector of probabilities by dividing by its sum. In an equal mixture of each cell type, the i^{th} probability can be interpreted as the fraction of transcripts originating from the i^{th} cell type.

Row Scaling

During this step, we apply a transformation on the expression values for each gene. Each gene has measured expression in N purified cell types and M samples. Each of these values, X_{old}, is transformed according to the following formula:

$$X_{new} = (X_{old} - Min)/(Max - Min) * Max^{p} \quad (2)$$

Where Min is the minimum of all M + N original values, Max is the maximum of those values, and p is a tunable parameter with natural range p $\in$[0.0,1.0]. This procedure produces values between the range of 0.0 and Max^{p}.

Reference Data

BLUEPRINT Reference Dataset

35 gene counts files were downloaded from the BLUEPRINT database, all collected from venous blood [17]. This included entries for CD14-positive, CD16-negative classical monocytes (5 samples), CD38-negative naive B cells (1), CD4-positive, alpha-beta T cell (8), central memory CD4-positive, alpha-beta T cell (2), cytotoxic CD56-dim natural killer cell (2), macrophage (4), mature neutrophil (10), and memory B Cell (1). When two or more transcripts appeared for a single gene, the transcript with the highest average expression was selected and others were excluded. Genes with no detected expression in

any sample were also excluded, and then each sample was quantile normalized. Samples generally clustered by cell type, but we excluded one CD4-positive alpha-beta T cell. Replicates for each cell type were then collapsed into a single entry by taking the median value for each gene.

ENCODE Reference Dataset

106 transcript quantification files were downloaded from the ENCODE database [18]. These included all RNA-seq experiments collected from adult primary cells, excluding four with warnings. Warnings indicated that three samples suffered from low replicate concordance and one sample from low read depth, and these samples were excluded. All samples were processed by the Gingeras Lab at Cold Spring Harbor and mapped to GRCH38.

The samples were quantile normalized and clustered. In cases where multiple transcripts were measured for a single gene, the expression of that gene was calculated as the sum of all transcripts. At this time, 18 additional samples were excluded as they did not cluster with their replicates. Based on sample descriptions and data clustering, we found that the remaining 88 samples represented 28 unique cell types. We produced an expression profile for each cell type by merging all samples of that cell type via median average. For example, a cluster of 19 samples were labelled as endothelial cells (collected from various body locations) and were merged into a single entry termed canonical endothelial cells. This dataset spans a wide range of stromal cell types (e.g. smooth muscle, fibroblast, epithelial), but contains only a single entry for blood cells, which are labelled mononuclear cells.

We also combined the ENCODE and BLUEPRINT reference matrices into a single reference matrix, which we call BlueCode. We combined, then quantile normalized, the columns of both matrices. Possible batch effects in this combined matrix have not been fully evaluated.

10x Reference Dataset

We obtained single cell expression data for nine varieties of immune cells from the 10x website [19]. This included at least 2446 cells for each cell type, and at least 7566 cells for all cells other than

CD14 monocytes. For each cell type, expression values for all cells were mean averaged to form an expression profile.

Tabula Muris Reference Dataset

We downloaded from the Tabula Muris single cell data for 12 clusters of mouse cell types. For each cluster, we averaged all cells of that cluster to produce a reference profile for the corresponding cell type.

Other Reference Datasets

Other datasets used in this project were obtained from their corresponding publications or GEO repositories. This includes a reference matrix of human skin signatures, the Human Body Atlas, the Human Primary Cell Atlas, LM22, ImmunoStates, the Mouse Body Atlas, and ImmGen [13–16,20,22,23].

Skin Diseases Data

We obtained expression data from 21 skin biopsies, collected from human patients with a variety of skin diseases. These data originally came from a wide range of sources and platforms, and were compiled into a single dataset by previous work [34]**.**

GTEx Data

GTExX data for 17,382 samples were obtained from the GTExX database (https://gtexportal.org/). We ran GEDIT on all samples three times, each time using a different reference matrix (BlueCode, the Human Primary Cell Atlas, and Skin Signatures). For each cell type, we calculated our initial estimate as the median estimate across the three sets of predictions (or fewer, if that cell type is missing from one to two of the reference matrices). Lastly, for each sample we divided the vector of predictions by its sum, such that the final predictions sum to 100%.

Multi-Tool Performance Evaluation

In Vitro Immune Cell Mixture

Combinations of six immune cells (Neutrophils, Monocytes, Natural Killer Cells, B cells, and CD4 and CD8 T Cells) were mixed together and sequenced using an affymetrix array. Whole blood from healthy human donors was supplied with informed consent through a sample sharing agreement with the UCLA/CFAR Virology Core Lab (grant number 5P30 AI028697). CD4+ T cells, CD8+ T cells, B cells, and NK cells were isolated using Stem Cell Technologies (Vancouver, BC, Canada) RosetteSep negative selection. Neutrophils were positively selected through the EasySep approach, according to the manufacturer's specifications. Cells were then counted by hemocytometer and added at defined percentages to a total cell count of two million cells to create six different mixtures. Subsequently cells were processed for RNA isolation by AllPrep DNA/RNA. Illumina HT12 BeadChip microarray was performed by the UCLA Neuroscience Genomics Core. Data was normalized by quantile normalization through R 'normalize.quantiles' function (R Core Team, 2013).

RNA-seq Benchmarking Mixtures

We also obtained two datasets used in a recent benchmarking study [29]. The first dataset is composed of three RNA-seq samples, each with two technical replicates that represent biopsies of ovarian cancer ascites [31]. The second dataset is composed of RNA-seq collected from the blood of healthy individuals, some of whom recently received an influenza vaccine [30]. These data were downloaded from the GitHub for the benchmarking paper, which also contained FACS estimates for six cell types for the ascites data (B cells, dendritic cells, NK cells, T cells, macrophages, neutrophils) and five cell types for the blood data (B cells, dendritic cells, T cells, monocytes, natural killer cells). However, since dendritic cells were never present at more than 3.5% abundance, we did not evaluate performance for this cell type.

Tools

We installed and ran GEDIT, CIBERSORT, DeconRNASeq and dtangle on the hoffman2 computational cluster at UCLA. xCell was run using the online interface at https://xcell.ucsf.edu/**.** The default choice for genes signatures (xCell =64) was used. The RNA-seq option was selected for the 2 RNA-seq datasets (blood and ascites), but not for the *in vitro* dataset, which was sequenced on microarray.

xCell produces 67 output scores, seven of which were used in this study. These were the entries labelled "B-Cells", "Macrophages", "Monocytes", "NK cells", "Neutrophils", "CD4+ T cells" and "CD8+ T Cells". As suggested by the xCell authors, the outputs for CD4 and CD8 T cell subtypes were summed to produce a final output for total T cells.

Reference Data

We evaluated the performance of the four reference-based tools (GEDIT, CIBERSORT, DeconRNASeq and dtangle) using each of four choices of reference matrix (LM22, ImmunoStates, BLUEPRINT, and the Human Primary Cell Atlas).The BLUEPRINT and Human Primary Cell Atlas reference matrices differ from ImmunoStates and LM22 in that they contain tens of thousands of genes, many of which should not be considered signature genes. This contrasts to ImmunoStates and LM22; each reference matrix contains fewer than 600 genes, which have been specifically identified as signature genes by previous work [13,20]. We include both forms of reference matrices in order to evaluate the input requirements of the tools studied.

Depending on the choice of reference matrix, reference-based tools often produce multiple outputs for some cell types, each representing a cell sub-type. This includes B cells (naïve and memory), Monocytes (CD14 and CD16), NK cells (resting and active) and T cells (many subtypes including varieties of CD4 and CD8). In each case, the outputs for each sub-type were summed in order to produce a total score for each greater cell type.

Conclusion:

GEDIT is an expression-based cell type quantification tool that offers unprecedented flexibility and accuracy in a wide variety of contexts. Using both simulated and experimental data, we demonstrate that GEDIT produces high-quality predictions for multiple platforms, species, and a diverse range of cell types, outperforming other tools in many cases. We include in the software package a comprehensive library of reference data, which facilitates application of GEDIT to a wide range of tissue types in both human and mouse. GEDIT can also accept reference data supplied by the user, which can be derived from bulk RNA-seq, scRNA-seq, or microarray experiments. GEDIT represents a competitive addition to the suite of existing tissue decomposition tools while maintaining for users flexibility and performance robustness.

We perform a benchmarking study as part of this project, in which we compare the performance of several deconvolution tools using multiple metrics. Unlike previous benchmarking studies, we explore the effect of reference choice by running tools multiple times with reference data from different sources. We find that, while choice of optimal reference is a complicated problem, the performance of GEDIT is highly robust to choice of reference, relative to other tools.

When extensively applied to several large public datasets, GEDIT produces predicted cell type fractions that conform with biological expectations. When used to decompose skin biopsies, keratinocytes are found to be the most abundant cell type and variations in the abundance of other cell types conform to expected immune responses across diseases. Similarly, cell type predictions of GTEx samples are concordant with our expectations of the dominant cell types across tissues. Schwann cells, keratinocytes, adipose cells, and immune cells are found to be most abundant in nerve, skin, adipose tissue, and blood, respectively.

Single cell RNA-seq is an emerging approach to study the composition of cell types within a sample. Due to biases associated with the capture of different cell types, these methods are not always capable of accurately quantifying cell type populations [8]. However, the pure reference profiles produced by existing methods can be used by GEDIT to generate accurate estimates of cell type populations. Thus, GEDIT circumvents some of the biases associated with the preparation of samples for both scRNA-seq and FACS. GEDIT is freely available, and therefore an extremely economical option to researchers, particularly those who profile expression data for other purposes.

GEDIT produces accurate results when tested on mixtures of human immune cells. Compared to other tools, GEDIT produces the lowest error in majority of scenarios in the studied mixtures. GEDIT provides increased flexibility over previously developed tools, as we provide a set of reference matrices for varied cell types for both mouse and human datasets.

GEDIT provides unique advantages compared to other tools, especially in terms of cell type, species and platform flexibility, and constitutes a useful addition to the existing set of tools for tissue decomposition. Our efficient decomposition methodology has been extensively optimized and we find that it performs robustly across a broad range of tissues in both mouse and human datasets. Our future work will extend reference matrices to facilitate application of GEDIT on varied bulk gene expression datasets.

Availability of Source Code and Requirements

- Project name: GEDIT
- Project Home Page: https://github.com/BNadel/GEDIT
- Programming Languages: Python 2.0, R
- Other requirements: numpy, glmnet
- Operating Systems: Linux
- License: MIT

Availability of Data and Materials

All data used in this paper are freely available on GitHub (https://github.com/purebrawn/GEDIT), as well as their original sources. Code for DeconRNASeq was obtained as an R package from the CRAN repository. Code for CIBERSORT was obtained by requesting it via the web portal (https://cibersort.stanford.edu/download.php), and code for dtangle from the project's GitHub page (https://github.com/BNadel/GEDIT).

Reference data is also available from their original sources. Most datasets can be found on project website pages or from public databases. These include BLUEPRINT (http://www.blueprint-epigenome.eu/), ENCODE (https://www.encodeproject.org), the Human Primary Cell Atlas (http://biogps.org/dataset/BDS_00013/primary-cell-atlas/), LM22 (http://cibersort.stanford.edu/ or GEO:GSE65136), 10x Genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets), Tabula Muris (https://tabula-muris.ds.czbiohub.org/), the Mouse Body Atlas (GEO:GSE10246), and ImmGen (http://www.immgen.org/Databrowser19/DatabrowserPage.html). Some reference matrices were obtained as supplementary files from the publications listed in Table 1.

Expression values for the blood and ascites RNA-seq datasets were obtained from the GitHub repository https://github.com/grst/immune_deconvolution_benchmark, and are also available at at https://figshare.com/s/711d3fb2bd3288c8483a and GEO: GSE64655). The *in vitro* mixture of immune cells was prepared by our lab, and available on our GitHub page.

Acknowledgments

We acknowledge the Biomedical Big Data Grant (5T32LM012424-03) for supporting Brian Nadel during the course of this research. We also acknowledge the Bruins-in-Genomics Summer Undergraduate Research Program for supporting Hannah Waddel and Misha Khan during the summer of 2017, when they contributed to this work. We also thank Lana Martin for her help with editing and proofreading the manuscript.

References

1. Bolen CR, Uduman M, Kleinstein SH. Cell subset prediction for blood genomic studies. BMC Bioinformatics. 2011;12: 258.

2. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21: 938–945.

3. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17: 174.

4. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12: 298–306.

5. Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biol. 2016;17: 231.

6. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017. pp. 395–398. doi:10.1038/nmeth.4179

7. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161: 1202–1214.

8. Hines WC, Su Y, Kuhn I, Polyak K, Bissell MJ. Sorting out the FACS: a devil in the details. Cell Rep. 2014;6: 779–781.

9. Frishberg A, Peshes-Yaloz N, Cohn O, Rosentul D, Steuerman Y, Valadarsky L, et al. Cell composition analysis of bulk genomics using single-cell data. Nat Methods. 2019;16: 327–332.

10. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10: 380.

11. Frishberg A, Brodt A, Steuerman Y, Gat-Viks I. ImmQuant: a user-friendly tool for inferring immune cell-type composition from gene-expression data. Bioinformatics. 2016;32: 3842–3843.

12. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18: 220.

13. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457.

14. Swindell WR, Johnston A, Voorhees JJ, Elder JT, Gudjonsson JE. Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients. BMC Genomics. 2013;14: 527.

15. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004;101: 6062–6067.

16. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013;14: 632.

17. Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica. 2013;98: 1487–1489.

18. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306: 636–640.

19. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.

20. Vallania F, Tam A, Lofgren S, Schaffert S, Azad TD, Bongen E, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. Nat Commun. 2018;9: 4735.

21. Consortium TTM, The Tabula Muris Consortium, Coordination O, Coordination L, Organ collection and processing, Library preparation and sequencing, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018. pp. 367–372. doi:10.1038/s41586-018-0590-4

22. Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res. 2008;4: 5.

23. Heng TSP, Painter MW, Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol. 2008;9: 1091–1094.

24. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29: 1083–1085.

25. Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. dtangle: accurate and robust cell type deconvolution. Bioinformatics. 2018 [cited 17 Jan 2019]. doi:10.1093/bioinformatics/bty926

26. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17: 218.

27. Lopez D, Montoya D, Ambrose M, Lam L, Briscoe L, Adams C, et al. SaVanT: a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. BMC Genomics. 2017;18: 824.

28. Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol. 2014;10: 720.

29. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35: i436–i445.

30. Hoek KL, Samir P, Howard LM, Niu X, Prasad N, Galassie A, et al. A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. PLOS ONE. 2015. p. e0118528. doi:10.1371/journal.pone.0118528

31. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8: 2032.

32. de Sousa JR, Lucena Neto FD, Sotto MN, Quaresma JAS. Immunohistochemical characterization of the M4 macrophage population in leprosy skin lesions. BMC Infect Dis. 2018;18: 576.

33. Lin C-C, Chen C-B, Wang C-W, Hung S-I, Chung W-H. Stevens-Johnson syndrome and toxic epidermal necrolysis: risk factors, causality assessment and potential prevention strategies. Expert Rev Clin Immunol. 2020;16: 373–387.

34. Inkeles MS, Scumpia PO, Swindell WR, Lopez D, Teles RMB, Graeber TG, et al. Comparison of molecular signatures from multiple skin diseases identifies mechanisms of immunopathogenesis. J Invest Dermatol. 2015;135: 151–159.

Supplementary Materials for the manuscript:

“The Gene Expression Deconvolution Interactive Tool (GEDIT):

Accurate Cell Type Quantification from Gene Expression Data”

Synthetic Mixture Generation

The deconvolution of synthetic mixtures using only a single matrix (to both generate the mixtures, and serve as a reference) is a trivial problem. In this context, the linear regression will always return the exact (or nearly exact) input proportions. Moreover, this is a poor simulation of real world data, as in reality the expression profile of any given cell type will vary to some extent between experiments. The mixtures submitted by the user will often be from different platforms than the reference data, and, in particular, cross-platform effects cannot be simulated using a single matrix, Therefore, in order to more meaningfully evaluate the performance of deconvolution, we used a separate matrix to produce mixtures from the one used as a reference.

Using distinct reference and mixture-generating matrices requires that we match cell types between the two matrices. Matching cell types across references is a non-trivial problem, as equivalent cell types may be labelled differently, and identically labelled cell types may not be equivalent. To address this problem, we defined the following procedure for identifying pairs of equivalent cell types between two reference matrices:

1. Joint quantile normalize the matrices, then log transform them
2. Calculate the Pearson correlations between each cell in the first matrix and each cell in the second matrix
3. Pair cell types that are more highly correlated with each other than with any other cell type in the reference
4. Manually exclude cell pairings with mismatching descriptions

Using this procedure, we identified 5 pairings of reference matrices that can be used for the generation of synthetic mixtures (Table 2). Since simulations can be done in both directions for each pair, this represents 10 possible choices of a mixture generating matrix and a reference matrix.

Matrix1	Matrix2	Number of Cell Types	Platforms
BluePrint	Human Primary Cell Atlas	5	RNASeq to Affymetrix U133 Microarray
BluePrint	10x Single Cell	4	Bulk RNASeq to SC RNASeq
BluePrint	Skin Signatures	6	RNASeq to Affymetrix/Illumina HT-12 Microarray
Human Primary Cell Atlas	Skin Signatures	10	Affymetrix U133 Microarray to Affymetrix/Illumina HT-12 Microarray
10x Single Cell	Skin Signatures	4	SC RNASeq to Affymetrix/Illumina HT-12 Microarray

Supplementary Table 1. Pairs of reference matrices used to generate synthetic mixtures.

For each of these 10 pairs of matrices, 1,000 cell type proportions were generated randomly. Specifically, a cell type was selected at random and assigned a weight between 0 and 1.0 (randomly sampled from the uniform distribution). Next, one of the remaining cell types is randomly selected and assigned a weight between 0.0 and the remaining weight (1.0 minus the sum of weights already assigned). This is repeated until the final cell type, which is assigned all remaining weight.

The final simulated expression profile is produced by summing the expression profiles of each cell type, multiplied by the simulated weight. We believe this procedure produces biologically reasonable mixtures, as they are composed primarily of a small number of cell types, with many other cell types present at low levels.

Signature Gene Selection

We have tested a total of 6 signature gene scoring algorithms; Entropy, fsDiff, fsRatio, meanDiff, meanRatio, and Zscore. For a given gene, these algorithms take as input the vector of expression values across all cell types, and return a score. Each gene is a candidate signature gene for the cell type in which it is most highly expressed, and only genes with the highest signature scores are accepted. The number of genes selected is determined by the NumSigs parameter, which is by default set to 50.

One scoring approach is to compare the highest observed expression value to the mean of all other expression values. This comparison can be performed by division or subtraction (MeanDiff and MeanRat). Alternately, these same comparisons can be made between the highest observed expression value, and the second highest observed value (fsDiff and fsRat). The Zscore method is calculated the same way as MeanDiff, except that it is divided by the variance of the expression vector.

When run on 10,000 simulated mixtures, selecting genes by entropy produced the lowest maximum, mean, and upper quartile error (Figure 2A). We therefore use entropy as the default setting, but allow the user to select any of the other 5 scoring methods. Using entropy has the potential to select genes that are highly expressed in 2 or more cell types, and lowly expressed in the rest. While these genes are not unique to a single cell type, they can still offer valuable information for deconvolution.

Number of Signature Genes (NumSigs, MinSigs)

GEDIT’s second parameter is the number of signature genes that are selected per cell type. On simulated data, any number of signature genes between 40 and 200 produce near-optimal results (Figure 2B).

We provide an option that allows more signature genes for some cell types than others. In this scheme, both an average and a minimum number of signature genes are specified by the user (NumSigs and MinSigs, respectively). For each of N cell types present in the reference, MinSigs genes are selected that are maximally expressed in that cell type. However, a total of N*NumSigs genes are selected, and the remaining N*(NumSigs-MinSigs) genes are simply those with the highest score, regardless of the cell type in which they are maximally expressed.

On simulated data, we found that adjusting the MinSigs parameter had minimal effect on predictions (Figure 2D), and by default GEDIT sets MinSigs equal to NumSigs.

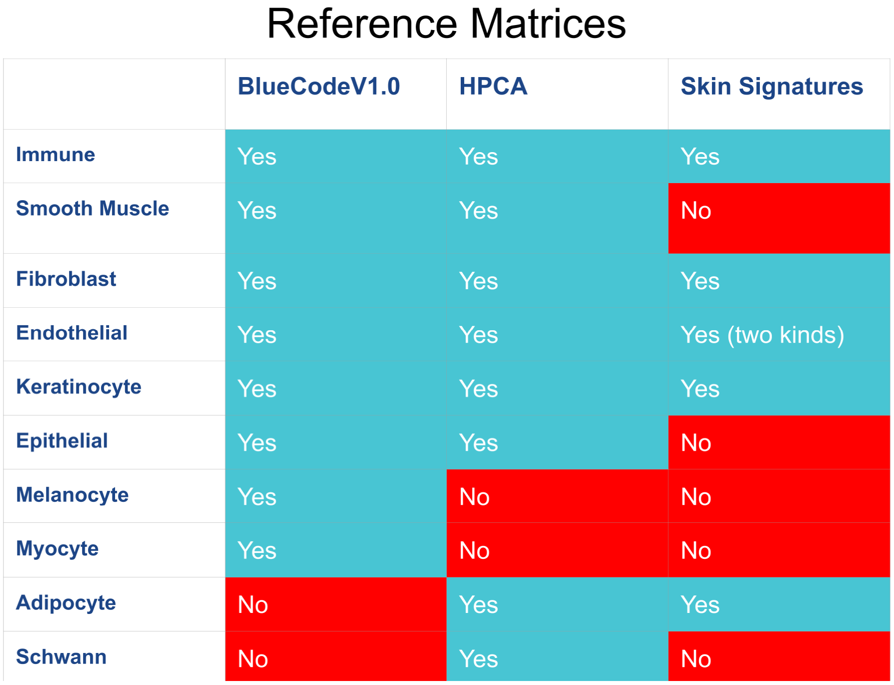
Row Scaling

The extent of row scaling is controlled by the row scaling parameter, with allowed values between 0.0 and 1.0. At 1.0 a gene with 10x higher expression will have 10x the influence (same as if no row scaling were performed). At a value of 0.0, all genes have equal influence. In simulated experiments, a row scaling value of 0.0 produced the lowest mean error, substantially improving accuracy (Figure 2C). Values outside the natural range of 0.0 to 1.0 produce high error, as well (data not shown).

Deconvolution of GTEX Database

We used GEDIT to estimate the cell type proportions of 10 cell types for 17,382 samples in the GTEX database. Since no single reference matrix contained all 10 cell type, we took an approach utilizing several reference matrices (Supplementary Figure 1). First, we combined the BLUEPRINT reference matrix (which contained only immune cells) with the ENCODE reference matrix (containing mostly stromal cells). We did this by concatenating the columns then quantile normalizing, and refer to the resulting matrix as “BlueCodeV1.0”. We then ran GEDIT deconvolution on the entire GTEX database three times; once using BlueCodeV1.0 as the reference, once using the Human Primary Cell Atlas, and once using the Skin Signatures matrix. For each predicted fraction in each sample, we took the median value of the 1-3 predictions produced. Lastly, we divided the predictions of each sample by their sum, such that predictions summed to 1.0. These values were used as final estimates for the fraction of each cell type in each sample.

Supplementary Figure 1. Cell types present in the 3 reference matrices used to predict cell type fractions of GTEX samples

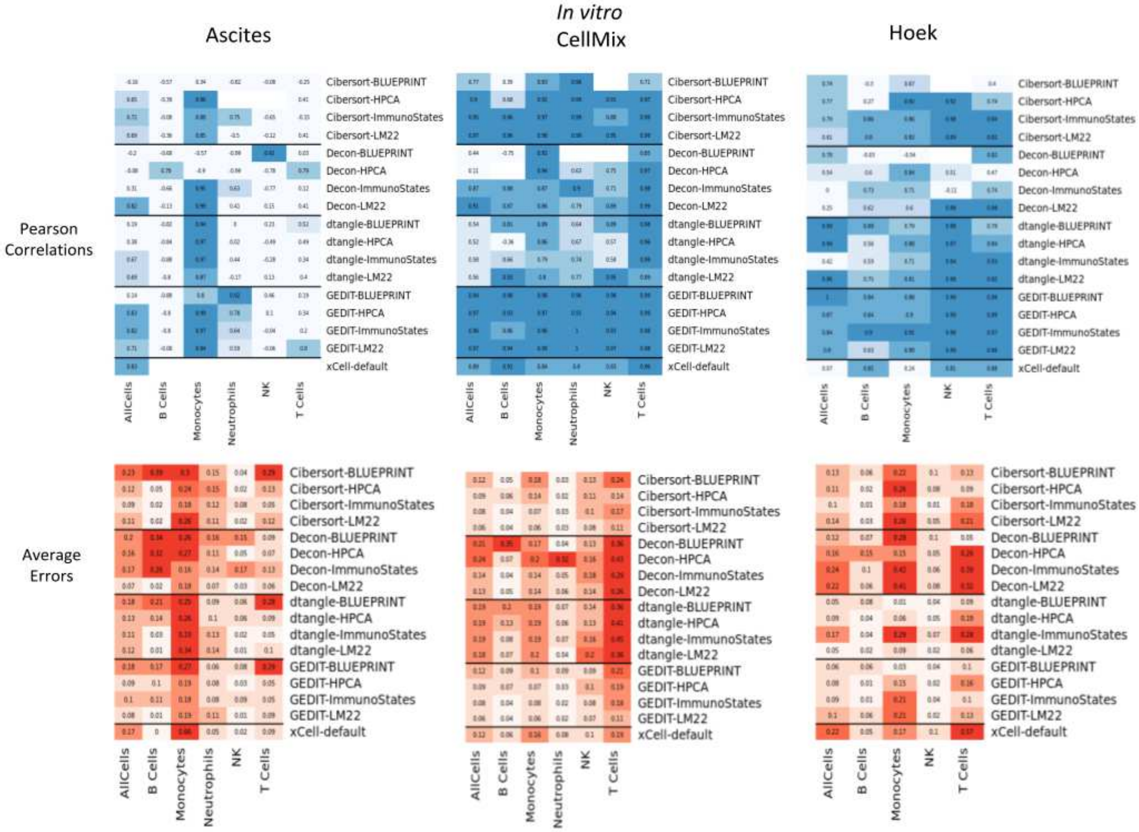


Reference Matrices

	BlueCodeV1.0	HPCA	Skin Signatures
Immune	Yes	Yes	Yes
Smooth Muscle	Yes	Yes	No
Fibroblast	Yes	Yes	Yes
Endothelial	Yes	Yes	Yes (two kinds)
Keratinocyte	Yes	Yes	Yes
Epithelial	Yes	Yes	No
Melanocyte	Yes	No	No
Myocyte	Yes	No	No
Adipocyte	No	Yes	Yes
Schwann	No	Yes	No

Multi-Tool Comparison

Here, we include figures presenting the pearson errors and correlations for each mixture used (and each cell type in those mixtures). This contrasts Figure 4, where values are considered for either each cell types (regardless of dataset), or each dataset (regardless of cell type).


Ascites
In vitro CellMix
Hoek
Pearson Correlations
Average Errors
Cibersort-BLUEPRINT
Cibersort-HPCA
Cibersort-ImmunoStates
Cibersort-LM22
Decon-BLUEPRINT
Decon-HPCA
Decon-ImmunoStates
Decon-LM22
dtangle-BLUEPRINT
dtangle-HPCA
dtangle-ImmunoStates
dtangle-LM22
GEDIT-BLUEPRINT
GEDIT-HPCA
GEDIT-ImmunoStates
GEDIT-LM22
xCell-default
AllCells
B Cells
Monocytes
Neutrophils
NK
T Cells

Chapter 3

Systematic Evaluation of Current Deconvolution Tools and Reference Datasets

Brian Nadel, Alice Mouton, Benjamin Shou, Feiyang Ma, Dennis J. Montoya, Matteo Pellegrini, Serghei Mangul

Abstract

Estimating cell type composition of blood or tissue samples is an important biological problem relevant to both laboratory studies and clinical care. Many computational tools have been published that estimate cell type abundance using gene expression data. These tools take a variety of approaches, all of which leverage either reference data or signature genes from purified cell types in order to evaluate the abundance of each cell type. In this study, we have compiled a set of 8 tools, and comprehensively evaluate their performance in a variety of contexts. Specifically, we have run these tools on 5364 mixtures of known proportions, spanning both immune cell types and stromal cell types. 12 mixtures represent *in vitro* synthetic mixtures, 300 represent *in silico* synthetic mixtures prepared using single cell data, and 5052 represent clinical samples with cell populations quantified by automated cell counting. Moreover, we have evaluated the performance of each tool on 11 versions of 6 reference matrices, and report the optimal choice of reference matrix for each tool.

Introduction

The cell type composition of heterogeneous tissues is fundamental to the biology and function of those tissues. In clinical settings, knowledge of cell type populations can glean insight into the nature of a wide range of diseases, as well as inform treatment. In cancer, for instance, the abundance of certain T cells correlates strongly with survivability, as well as the efficacy of immunotherapy treatment. In laboratory settings, researchers frequently observe gene expression changes that are difficult to interpret. Such patterns can result either from changes in cell type abundances or from modulation of the expression of one or more cell types. Researchers rely on cell type quantification methods to distinguish between these two cases and lend greater power to their experiments.

Several approaches exist to quantify cell type populations, but all suffer from some form of limitation or bias. Cell flow cytometry via FACS sorting is often considered an accurate method, but is extremely slow, costly, and difficult to apply to large studies. Moreover, this technology struggles to quantify cell types with unusual morphologies, such as neurons, myocytes, and adipocytes. More recently, single cell methods such as Drop-seq have become available. However, these methods suffer from the same cost and cell type limitations as FACS sorting. In addition, both methods have potential to introduce bias. Subtle differences in the way samples are prepared can drastically change the numbers of each cell type successfully captured by these technologies. As a result, pure samples of each cell type can be obtained, but the numbers of cells observed frequently do not reflect the biology of the original sample.

In-silico deconvolution using gene expression data has been developed as a possible alternative. Dozens of tools have been published taking this approach, but little attention has been given to evaluating the accuracy of these tools in varying contexts. Many tools are designed to be applied to particular cell types or platforms, and it is currently unclear how robust predictions are when applied to novel contexts.

In addition, most deconvolution tools require access to reference data. These data quantify the expression profiles of isolated cell types, and are necessary for tools to model combinations thereof. Studies have shown that the impact of choice of reference can have a large impact on accuracy of results, perhaps even greater than choice of tool (Vallania *et. al*, 2018). While invaluable work has been done assembling high-quality reference data (Newman et al 2015, Vallania et al 2018), rigorous evaluation of reference choice has not been performed for many tools. This is, in part, because authors can test their tool only on references available at the time of publication, and newer reference data is constantly becoming available.

In this study, we perform a benchmarking study that represents the most comprehensive evaluation of deconvolution methods to date. In total, we have performed over 300,000 deconvolution tasks, exhaustively searching 90 pairwise combinations of method and reference, in over 4,000 samples. Previous deconvolution studies rely largely on simulated data, which fails to capture the true complexity of tissue samples in living organisms. Here, however, we use over 4,500 clinical samples to evaluate the problem in a more powerful, complete and unbiased manner. The cell type composition of the clinical blood samples has been evaluated via impedance-based electronic cell counter, a gold standard for high-throughput cell type quantification in blood. In addition, we have thoroughly evaluated the effect of reference choice on the accuracy of deconvolution prediction.

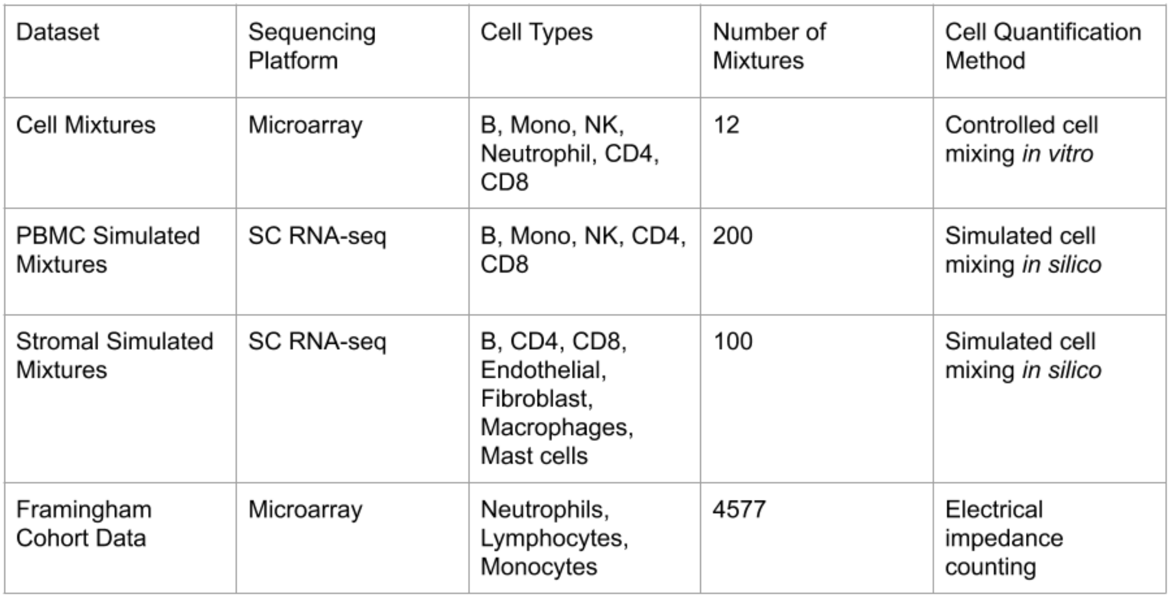
Results

In this study, we compare nine deconvolution tools and evaluate their performances using several different datasets with known proportions. The tools included in the study are Cibersort (normal and absolute mode), the DCQ algorithm, DeconRNASeq, dtangle, EPIC, GEDIT, MCP-Counter, SaVanT, and xCell.

Tool	Publication	Reference Input Type	Algorithm	Language	Output Type
Cibersort	Newman *et. al.*, 2015	Expression matrix	Support Vector Regression	R	Predicted Fractions
Cibersort (Absolute Mode)	Newman *et. al.*, 2015	Expression matrix	Support Vector Regression	R	Scores
DCQ (ImmQuant)	Altboum *et. al.*, 2014, Frishberg *et. al.*, 2016	Expression matrix	DCQ	R	Scores
DeconRNASeq	Gong *et. al.*, 2013	Expression matrix	Linear Regression	R	Predicted Fractions
dtangle	Hunt *et. al.*, 2018	Expression matrix	Regression in log-space	R	Predicted Fractions
EPIC	Racle *et. al.*, 2017	Expression matrix and signature gene list	Linear Regression	R	Predicted Fractions
GEDIT	Unpublished	Expression Matrix	Linear Regression	Python and R	Predicted Fractions
MCP-Counter	Becht *et. al.*, 2016	Signature gene list	Marker Gene Expression	R	Scores
SaVanT	Lopez *et. al.*, 2017	Signature gene list	Marker Gene Expression	Python	Scores
xCell	Aran *et. al.*, 2017	built-in	Marker Gene Expression	R	Predicted Fractions

Table 1. Deconvolution tools evaluated by this benchmark.

We use four datasets to perform this benchmark, which include 300 synthetic mixtures prepared *in silico* using single cell data, 14 mixtures prepared *in vitro* and sequenced using microarray, and 3,728 clinical samples

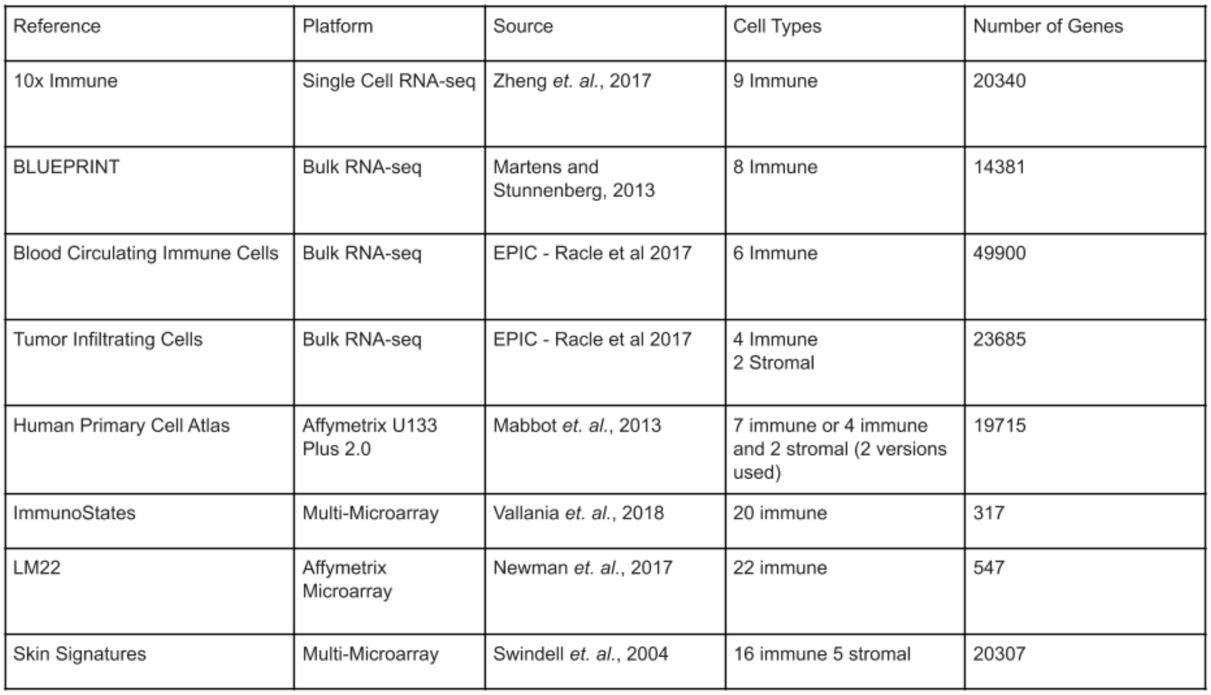


Dataset	Sequencing Platform	Cell Types	Number of Mixtures	Cell Quantification Method
Cell Mixtures	Microarray	B, Mono, NK, Neutrophil, CD4, CD8	12	Controlled cell mixing *in vitro*
PBMC Simulated Mixtures	SC RNA-seq	B, Mono, NK, CD4, CD8	200	Simulated cell mixing *in silico*
Stromal Simulated Mixtures	SC RNA-seq	B, CD4, CD8, Endothelial, Fibroblast, Macrophages, Mast cells	100	Simulated cell mixing *in silico*
Framingham Cohort Data	Microarray	Neutrophils, Lymphocytes, Monocytes	4577	Electrical impedance counting

Table 2. Mixture data used in this benchmark. These come from 3 independent sources, and represent a combination of *in silico* simulation, *in vitro* experiments, and clinical samples. Cell Mixtures was prepared by mixing 6 immune cell types together in known proportions, then sequencing via

microarray. Simulated data was prepared using single cell RNA-seq data, where random cells were selected, their expression values summed, and the cell type ratios noted. The Framingham Cohort data is collected from the blood of healthy individuals, and cell types quantified using electrical impedance.

Reference data is an essential requirement for most tools included in this benchmark. Depending on the tool, this can either take the form of an expression matrix, a list of signature genes for each cell type, or both. It has been shown that choice of reference can have an enormous impact on the quality of results, and we explore that relationship here. Specifically, for each combination of tool and mixture, we test several possible choices of reference matrix. We have identified 7 reference matrices that contain the necessary cell types to be used with our mixtures. These come from a variety of sources and platforms, including LM22, immunoStates, 10x Genomics, EPIC, BLUEPRINT, and the Human Primary Cell Atlas.



Reference	Platform	Source	Cell Types	Number of Genes
10x Immune	Single Cell RNA-seq	Zheng *et. al.*, 2017	9 Immune	20340
BLUEPRINT	Bulk RNA-seq	Martens and Stunnenberg, 2013	8 Immune	14381
Blood Circulating Immune Cells	Bulk RNA-seq	EPIC - Racle et al 2017	6 Immune	49900
Tumor Infiltrating Cells	Bulk RNA-seq	EPIC - Racle et al 2017	4 Immune 2 Stromal	23685
Human Primary Cell Atlas	Affymetrix U133 Plus 2.0	Mabbot *et. al.*, 2013	7 immune or 4 immune and 2 stromal (2 versions used)	19715
ImmunoStates	Multi-Microarray	Vallania *et. al.*, 2018	20 immune	317
LM22	Affymetrix Microarray	Newman *et. al.*, 2017	22 immune	547
Skin Signatures	Multi-Microarray	Swindell *et. al.*, 2004	16 immune 5 stromal	20307

Table 3. The set of reference matrices used in this study. They span several platforms and some contain stromal cells, as well as immune.

First, we evaluate the effect of choice of reference for each tool, and determine the best reference for each application. Our metric for quantifying performance is correlation between predicted cell type fractions, and actual cell type fractions (as measured by orthogonal means). We find that choice of reference does indeed, have a large impact on quality of results. Moreover, there is no single choice of reference that performs best for all tools or for all mixtures.

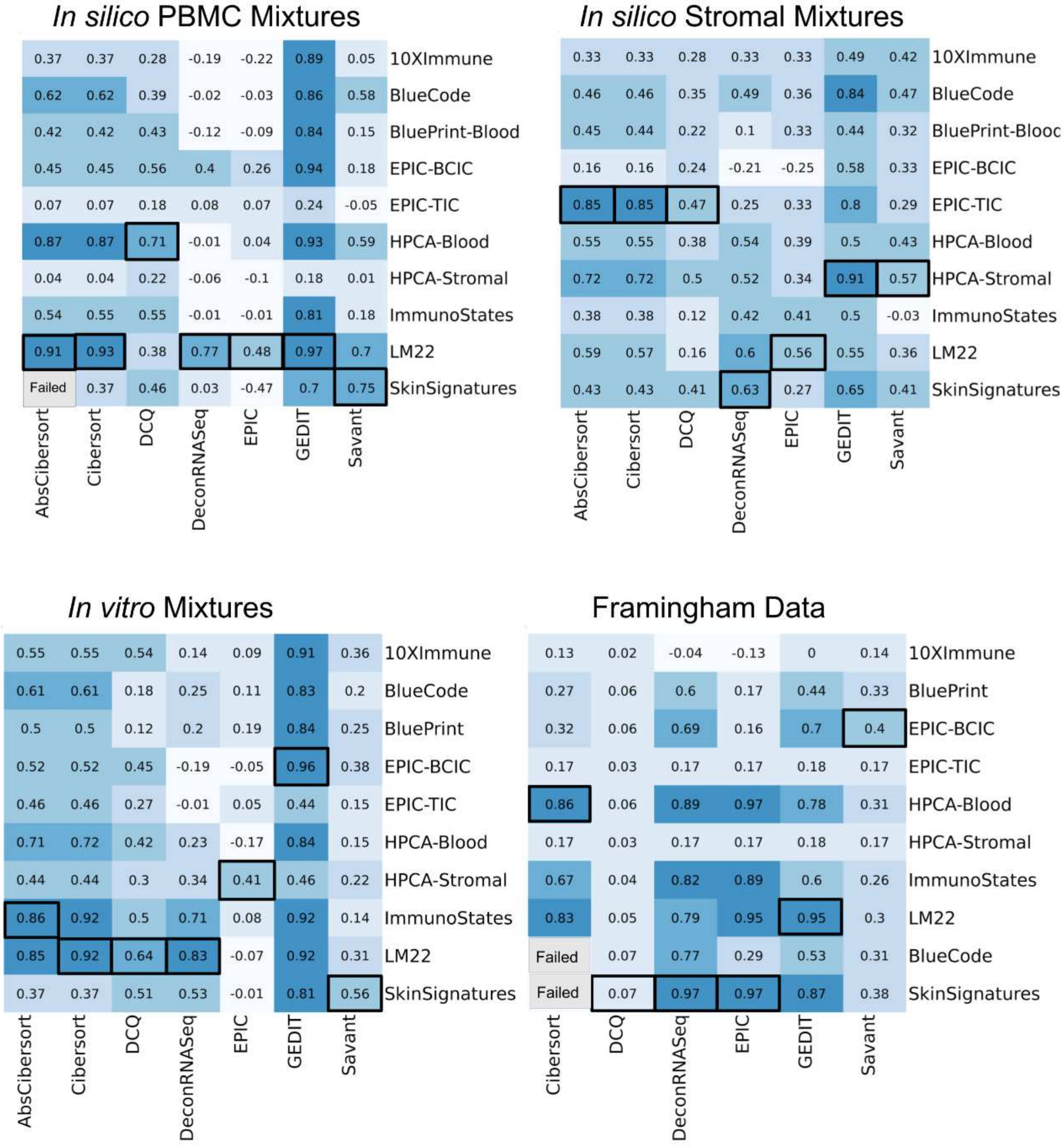
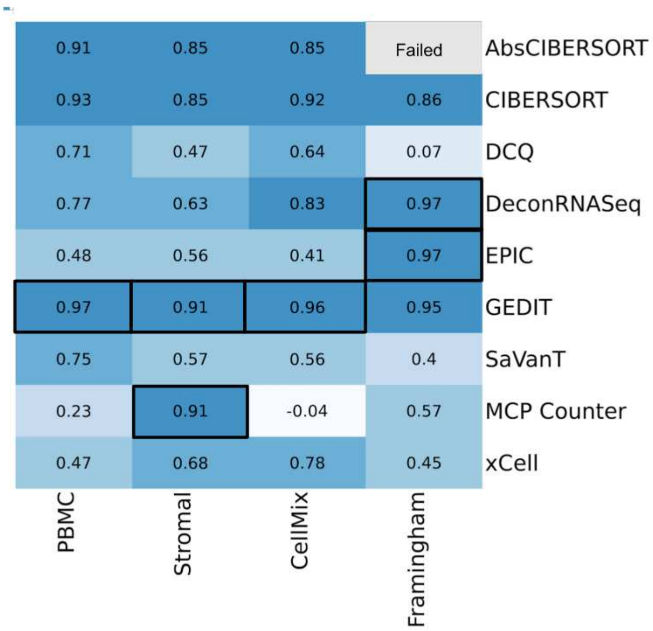


Figure 1. Pearson correlations between predicted and actual cell fractions for each combination of tool and reference matrix. Tools that do not accept custom references are not shown (MCP-Counter, xCell). Absolute Cibersort failed to run on the Framingham data due to high resource usage.

Next, we compare performance between tools when the optimal reference is used in each case. We use two metrics to compare the accuracy of predicted fractions compared to actual fractions: correlation and accuracy. In the case of correlation, this can be simply computed for every tool. By this metric, the best tool to use depends on the mixture being deconvoluted. CIBERSORT and GEDIT produce the most reliably accurate results, though each are outperformed by DeconRNASeq and EPIC when applied to the Framingham dataset. For each mixture tested, multiple tools produce highly accurate results (correlation greater than .9), though no single tool does this reliably for all mixtures.



	PBMC	Stromal	CellMix	Framingham
AbsCIBERSORT	0.91	0.85	0.85	Failed
CIBERSORT	0.93	0.85	0.92	0.86
DCQ	0.71	0.47	0.64	0.07
DeconRNASeq	0.77	0.63	0.83	0.97
EPIC	0.48	0.56	0.41	0.97
GEDIT	0.97	0.91	0.96	0.95
SaVanT	0.75	0.57	0.56	0.4
MCP Counter	0.23	0.91	-0.04	0.57
xCell	0.47	0.68	0.78	0.45

Figure 2. Correlations between predicted and actual fractions for each mixture and tool. For tools that accept custom reference data, the reference data that resulted in the highest correlation is shown here (see figure 2).

Some tools do not explicitly predict fractions, and evaluating error for these tools requires modification of their outputs. SaVaNT, MCP-Counter, DCQ, and CIBERSORT (absolute mode) do not predict cell type fractions. Instead, they produce scores that are meant to be compared between samples, rather than between cells. In order to explore whether inter-cellular comparisons are feasible or valid for these tools, we perform a simple transformation to convert the output into predicted fractions. For each sample, we divide each cell-type score by the sum of scores for all cell types. Therefore, the resulting sum to 1.0, and can be treated as fractions. It should be noted that this is not how these tools were designed to be used, but in many cases, they produce accurate results.

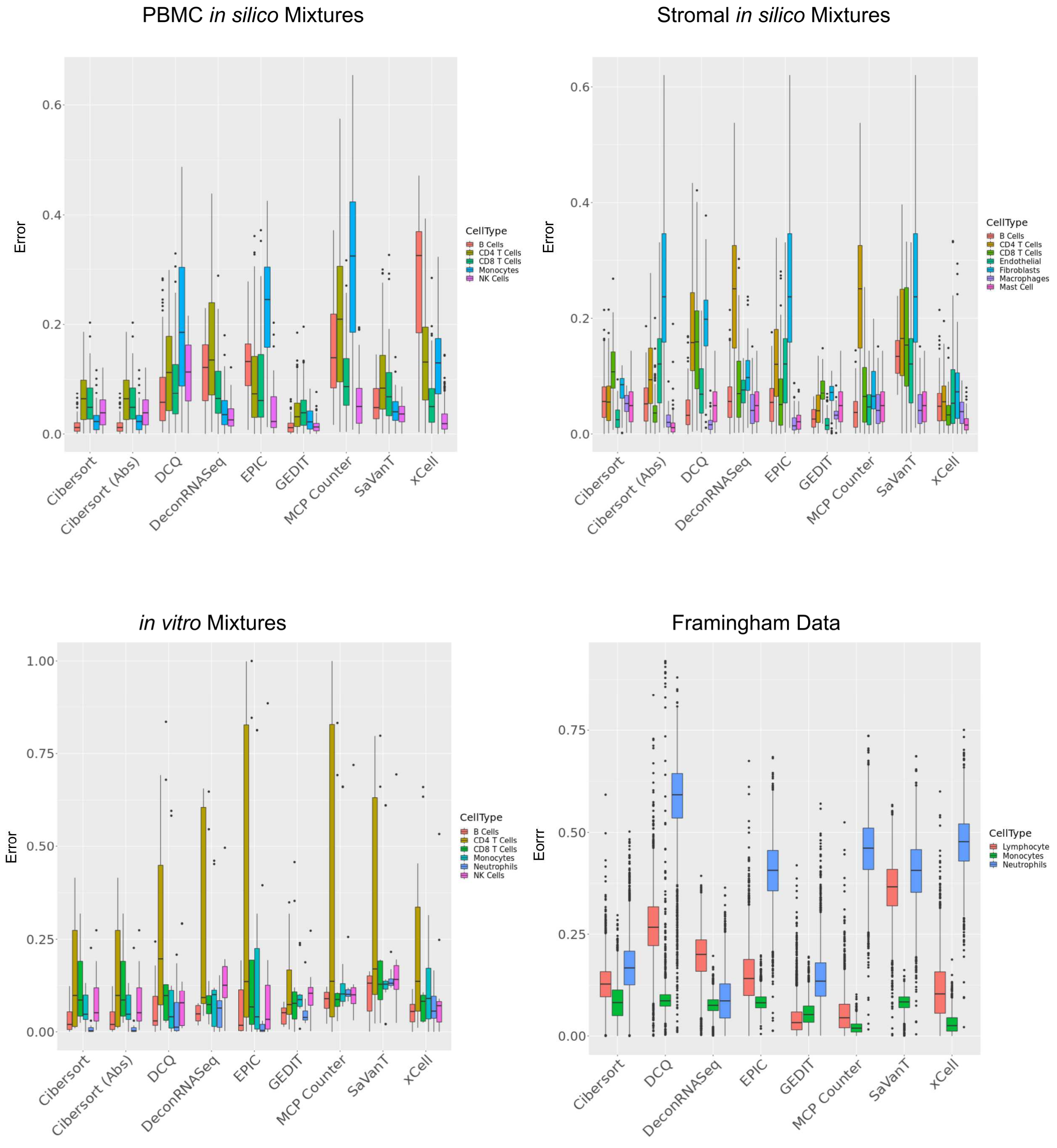


Figure 3. Distribution of errors when each tool is used to predict fractions for each mixture.

Error of predicted fractions varies greatly depending on the exact combination of tool, cell type, and mixture. CIBERSORT and GEDIT generally perform well in most cases. When using the absolute mode of CIBERSORT, compared to the default version, either has minimal or negative effect. This is unsurprising, since the default CIBERSORT is designed to predict fractions, whereas the absolute version is not. xCell performs best on the *in silico* simulated mixtures, but produces high error for some cell types in the *in vitro* mixture (e.g. B Cells).

For all tools, we observe some of the highest errors when applied to the Framingham data. This is likely due to complexities of living tissues, including varying sub-states for many cell types, that are not adequately reproduced in the simulated data. In particular, neutrophils prove the most difficult to predict for most tools.

Conclusion

Expression based cell type deconvolution is an increasingly popular means of interpreting biological data. However, current approaches and requisite data are numerous, and it is important that users have a clear way of identifying the best choices for their needs. Here, we perform the most comprehensive benchmarking project to date, in which we compare many popular tools, and explore the intricacies of reference choices and other factors.

We explore the effect of choice of reference matrix, and demonstrate its importance. Some tools, like for particular tools, such as and the extreme sensitivity of some tools with respect to choice of reference. Moreover, there is no universally optimal reference, even for particular mixtures. Different tools appear to have preferences for certain references, perhaps because these tools were developed with particular platforms or cell types in mind.

We find that most tools can perform competitive, accurate results when run in the correct context. However, the performance of several tools deteriorates dramatically when run on particular mixtures or using particular reference data.

Methods

Selected tools

We have selected available deconvolution tools able to infer the relative abundances of immune cell types based on the gene expression profiles. In total, we have identified 9 deconvolution tools, all which estimate cell type abundance

CIBERSORT [1]. We have used CIBERSORT version 1.04 installed on UCLA Hoffman2 cluster (R version 3.4.0); the full R package can be downloaded from https://cibersort.stanford.edu/download.php (account required). CIBERSORT includes a default reference ("LM22.txt" downloadable from above link) that consists of 22 distinct immune cell types. Bulk expression data was run with both LM22 and HPCA signatures (link GitHub). All files were run with 500 permutations (author recommends >100). All other parameters are set to their default status. Due to long runtime when CIBERSORT was interactively on Hoffman2, files were fed into CIBERSORT using a wrapper, and the program was run using a submission script (link GitHub).

EPIC. We use Version 1.1, ran online (https://gfellerlab.shinyapps.io/EPIC_1-1/) with both pre-built reference (tumor infiltrating and blood circulating immune); both of these references can be downloaded for cluster use from https://github.com/smangul1/deconvolution-benchmarking/tree/master/ReferenceDatasets. Results are displayed as cell fractions.

xCell. Using online tool found at http://xcell.ucsf.edu/. The bulk expression data is submitted under "upload gene expression data" and the default gene signatures were used (xCell, n=64). RNASeq option was selected for PBMC1, PBMC2, and Stromal datasets but not for CellMixtures (microarray platform).

SaVant was obtained from it's authors and run using 50 signature genes per cell type.

Microenvironment Cell Populations-counter (MCPcounter). Is a R package that we downloaded from the github repository (https://github.com/ebecht/MCPcounter). The package was installed on UCLA Hoffman2 cluster (R version 3.2.0) with its dependency ("devtools", "curl") but it can be run as well on personal computer due to its quick runtime (average 10 secondes). The R package takes a gene expression matrix (with features in rows and samples in columns) as input and give an abundance score for eight immune and two stromal cell populations. The abundance scores correspond to the mean expression of markers that are specific of some cell populations. MCP infer T cells and CD8 T cells but does not infer CD4 T cells. The feature type in the input can be in affymetrix, HUGO or Entrez id. Although the package allows for the use of an external reference, we used the package in default mode with their own reference due to the high complexity of formatting. The default reference was used for this study, as the use of custom references would require direct modification of the MCP-counter code.

Mixture Data

Cell mixture microarray

Whole blood from healthy human donors was supplied with informed consent through a sample sharing agreement with the UCLA/CFAR Virology Core Lab (grant number 5P30 AI028697). CD4+ T cells, CD8+ T cells, B cells, and NK cells were isolated using Stem Cell Technologies (Vancouver, BC, Canada) RosetteSep negative selection, while neutrophils were positively selected through EasySep approach, according to manufacturer's specifications. Cells were then counted by hemocytometer and added at defined percentages to a total cell count of two million cells to create six different mixtures. Subsequently cells were processed for RNA isolation by AllPrep DNA/RNA. Illumina HT12 BeadChip microarray was performed by the UCLA Neuroscience Genomics Core. Data was normalized by quantile normalization through R 'normalize.quantiles' function

PBMC and Stromal Single Cell Mixtures

We obtained 2 datasets (PBMC1 and PBMC2) from 10x Genomics (https://www.10xgenomics.com/resources/datasets/) and 1 from GEO (Puram et al).
For PBMC1, we used 1000 cells for each sorted cell types. For each cell type, we randomly selected 1-1000 cells, then we sum the expression of all the selected cells to create a synthetic mixture. The process was repeated 100 times, thus 100 mixture was created. For PBMC2, we firsted clustered the cells and identified the cell types for the dataset. Then we used the same five cell types in PBMC2 and created the 100 mixtures the same way as we did from PBMC1. For stromal cells, we created 100 mixtures the same way as we did for PBMC2 except for that we included some stromal cell types. For each mixture, true fraction for each cell type is calculated as the number of cells of that cell type selected, divided by the total number of cells across all cell types.

Framingham Data

The "gold standard" is a) cell counts and cell percent. The cell counting was performed on a Beckman Coulter HmX hematology analyzer. The following metrics from whole blood were obtained - HbA1c, basophil count and percent, eosinophil count and percent, hematocrit, hemoglobin, lymphocyte count and percent, MCH, MCHC, MCV, monocyte count and percent, MPV, neutrophil count and percent, platelet count, RBC, RDW, and WBC. Find information for the variables here.

Reference Data

Blueprint

35 gene counts files were downloaded from the BLUEPRINT database, all collected from venous blood. This included entries for CD14-positive, CD16-negative classical monocytes (5 samples), CD38 negative naive B cells (1), CD4-positive, alpha-beta T cell (8), central memory CD4-positive, alpha-beta T cell (2), cytotoxic CD56-dim natural killer cell (2), macrophage (4), mature neutrophil (10), and memory B Cell (1). When two or more transcripts appeared for a single gene, the transcript with the highest average expression was selected, and others excluded. Genes with no detected expression in any sample were also excluded, and then each sample was quantile normalized. Samples generally clustered by cell type, though one sample of CD4-positive, alpha-beta T cells did not, and was excluded. Replicates for each cell type were then collapsed into a single entry by taking the median value for each gene.

LM22

The LM22 reference matrix was assembled as part of the original Cibersort publication, and contains 547 selected to distinguish 22 human hematopoetic cell types

ImmunoStates

The ImmunoStates reference matrix was recently published, and contains 318 genes selected to distinguish between 22 immune cell types (Vallania *et. al*, 2018).

References

Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol. 2014;10: 720. doi:10.1002/msb.134947

Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18: 220. doi:10.1186/s13059-017-1349-1

Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17: 218. doi:10.1186/s13059-016-1070-5

Frishberg A, Brodt A, Steuerman Y, Gat-Viks I. ImmQuant: a user-friendly tool for inferring immune cell-type composition from gene-expression data. Bioinformatics. 2016;32: 3842–3843. doi:10.1093/bioinformatics/btw535

Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29: 1083–1085. doi:10.1093/bioinformatics/btt090

Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. dtangle: accurate and robust cell type deconvolution. Bioinformatics. 2018 [cited 17 Jan 2019]. doi:10.1093/bioinformatics/bty926

Lopez D, Montoya D, Ambrose M, Lam L, Briscoe L, Adams C, et al. SaVanT: a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. BMC Genomics. 2017;18: 824. doi:10.1186/s12864-017-4167-7

Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013;14: 632. doi:10.1186/1471-2164-14-632

Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica. 2013;98: 1487–1489. doi:10.3324/haematol.2013.094243

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., … Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5), 453–457. https://doi.org/10.1038/nmeth.3337

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., & Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *ELife*, 6. https://doi.org/10.7554/eLife.26476

Swindell WR, Johnston A, Voorhees JJ, Elder JT, Gudjonsson JE. Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients. BMC Genomics. 2013;14: 527. doi:10.1186/1471-2164-14-527

Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T. D., Bongen, E., … Khatri, P. (2018). Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature Communications*, 9(1), 4735. https://doi.org/10.1038/s41467-018-07242-6

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049. doi:10.1038/ncomms14049

Chapter 4
Manuscript for:
Cell type specific genetic regulation of gene expression across human tissues

Authors: Sarah Kim-Hellmuth[1,2,3†*], François Aguet[4†], Meritxell Oliva[5,6†], Manuel Muñoz-Aguirre[7,8], Valentin Wucher[7], Silva Kasela[2,3], Stephane E. Castel[2,3], Andrew Hamel[4,9], Ana Viñuela[10,11,12,13], Amy Roberts[10], Serghei Mangul[14,15], Xiaoquan Wen[16], Gao Wang[17], Alvaro Barbeira[5], Diego Garrido-Martín[7], Brian Nadel[18], Yuxin Zou[19], Jie Quan[20], Andrew Brown[11,21], Angel Martinez-Perez[22], José Manuel Soria[22], GTEx Consortium, Gad Getz[4,23], Emmanouil T Dermitzakis[11,12,13], Kerrin Small[10], Matthew (*1*)alin S. Xi[24], Hae Kyung Im[5], Roderic Guigó[7,25], Ayellet Segrè[4,9], Barba (*2*)[5.26], Kristin G. Ardlie[4], Tuuli Lappalainen[2,3*]

Affiliations:

[1] Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany

[2] New York Genome Cente (*3*)Y, USA

[3] Department of Systems Biology, Columbia University, New York,

(*4*)The Broad Institute o (*5*), Cambridge, MA, USA

[5] Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA

[6] Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA

[7] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain

[8] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain

[9] Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA

[10] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

[11] Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

[12] Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switz

(*6*)s Institute of Bioinformatics, Geneva, Switzerland

(*2*)of Computer Science, University of Califo (*7*)es, Los Angeles, CA, USA

[15] Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, USA

[16] Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[17] Department of Human Genetics, University of Chicago, Chicago, IL, USA

[18] Department of Computer Science, University of California, Los Angeles, CA, USA

[19] Department of Statistics, University of Chicago, Chicago, IL, USA

[20] Inflammation & Immunology, Pfizer, Cambridge, MA, USA.

[21] Population Health and Genomics, University of Dundee, Dundee, Scotland, UK

[22] Unit of Genomic of Complex Diseases, Institut d'Investigació Biomèdica Sant Pau (IIB-Sant Pau), Barcelona. Spain

[23] Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

[24] Foundational Neuroscience Center, AbbVie, Cambridge, MA, USA

[25] Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

[26] Center for Genetic Medicine, Department of Pharmacology, Northwestern University, Feinberg School of Medicine, Chicago, IL, USA

†These authors contributed equally to this work.

*Correspondence to: skimhellmuth@gmail.com, tlappalainen@nygenome.org

Abstract

The Genotype-Tissue Expression (GTEx) project has identified expression and splicing quantitative trait loci (*cis*-QTLs) for the majority of genes across a wide range of human tissues. However, the interpretation of these QTLs has been limited by the heterogeneous cellular composition of GTEx tissue samples. Here, we map interactions between computational estimates of cell type abundance and genotype to identify cell type interaction QTLs for seven cell types. We show that cell type interaction eQTLs contribute to the tissue specificity and allelic heterogeneity of *cis*-eQTLs. Using colocalization analyses with 87 complex traits, we demonstrate that in addition to pinpointing the cellular origin of known *cis*-QTLs, cell type interaction QTLs enable the discovery of hundreds of loci that are masked in bulk tissue.

One Sentence Summary

Estimated cell type abundances from bulk RNA-seq across tissues reveal the cellular specificity of quantitative trait loci.

Main text

The Genotype-Tissue Expression (GTEx) project (*1*) and other studies (*17*)(*16*)(*15*)(*14*)(*13*)(*12*)(*11*)(*10*)(*9*)(*8*)(*2-5*) have shown that genetic regulation of the transcriptome is widespread. GTEx in particular has built an extensive catalog of expression and splicing quantitative trait loci in *cis* (*cis*-eQTLs and *cis*-sQTLs) across tissues, showing that QTLs are generally either highly tissue-specific or widely shared, even across dissimilar tissues and organs (*1*, *6*). However, the vast majority of these studies have been performed using heterogeneous bulk tissue samples comprising diverse cell types. This limits the power, interpretation, and downstream applications of QTL studies. Genetic effects that are active only in rare cell types may be left undetected, mechanistic interpretation of QTL sharing across tissues and other contexts is complicated without understanding differences in cell type composition, and inference of downstream molecular effects of regulatory variants without the specific cell type context is challenging. Efforts to map eQTLs in individual cell types have been largely restricted to blood, using purified cell types (*7-10*) or single cell sequencing (*11*). Cell type specific eQTLs can also be computationally inferred from bulk tissue measurements, using the estimated proportion or enrichment of relevant cell types to test for an interaction with genotype, but such approaches have also been largely limited to blood cell types (*12*, *13*) and adipocytes (*14*). These studies identified thousands of cell type interactions in eQTLs discovered in whole blood samples from large cohorts [5,683 samples (*12*); 2,116 samples, (*13*)], indicating that large numbers of interactions are likely to be identified by expanding this type of analysis to other tissues and cell types.

In this study, we applied cell type deconvolution to characterize the cell type specificity of *cis*-eQTLs and *cis*-sQTLs for seven cell types across the majority of GTEx tissues (Fig. 1A). Estimating the cell type composition of a tissue biospecimen from RNA-seq remains a challenging problem {Cobos:2018ga} and multiple approaches for inferring cell type proportions have been proposed (*15*). We performed extensive benchmarking for multiple cell types across several expression datasets (fig. S1). The xCell method (*16*), which estimates the enrichment of 64 cell types using reference profiles, was most robust based on correlation with cell counts in blood (fig. S1A), *in silico* simulations (fig. S1B), and correlation with expression of marker genes for each cell type (fig. S1C). Furthermore, the inferred abundances reflected differences in histology and tissue pathologies (fig. S1D,E). For each cell type, we selected tissues where the cell type was highly enriched to map cell type interacting eQTLs in *cis* (fig. S2A). The xCell scores for these tissue-cell type pairs were also highly correlated with the PEER factors used to correct for unobserved confounders in the expression data for QTL mapping (*1*) (fig. S2B). We used a linear regression model for gene expression that included an interaction between cell type enrichment and genotype, thus enabling identification of eQTLs where the effect size is correlated with the enrichment of the cell type (Fig. 1B). Since QTLs identified this way are not necessarily specific to the estimated cell type but may reflect another (anti)correlated cell type, we refer to these eQTLs as cell type interacting eQTLs, or cell type ieQTLs. We applied an analogous approach to map cell type interacting splicing QTLs (isQTLs), using intron excision ratios that reflect alternative isoform usage, quantified by LeafCutter {Li:2017cy} (Fig. 1B). Across cell types and tissues, we detected 3347 protein coding and lincRNA genes with an ieQTL (ieGenes) and 987 genes with an isQTL (isGenes) at 5% FDR per cell type-tissue combination (Fig. 2A, fig. S3A and Table S1). The QTL effect of ieQTLs and isQTLs can increase or decrease as a function of cell type enrichment (Fig. 1C). This correlation is usually positive (55%; median across cell type-tissue combinations); for example, a keratinocyte ieQTL for *CNTN1* in skin had a particularly strong effect in samples with high enrichment of keratinocytes. However, a significant number of ieQTLs the effect was negatively correlated (22%) or ambiguous (20%) (fig. S4A,B), with the interaction likely capturing a QTL that is active in another cell type. Notably, while 85% of ieQTLs corresponded to genes with at least one standard eQTL, 21% of these ieQTLs were not in LD ($R^2 < 0.2$) with any of the corresponding eGene's conditionally independent eQTLs (fig. S4C), indicating that ieQTL analysis often reveals genetic regulatory effects that are not detected by standard eQTL analysis. There was only a modest correlation between sample size and ieQTL/isQTL discovery (Spearman's $\rho = 0.53$ and 0.35, respectively; fig. S3B), which may be explained by inter-individual variance in cell type enrichments driven by tissue heterogeneity effects being a major determinant in discovery power. For example, breast and transverse colon both stratified into at least two distinct groups based on histology: epithelial vs. adipose tissue (breast) and mucosal vs. muscular tissue (colon) (fig. S1B). Downsampling analyses in whole blood and transverse colon revealed linear relationships between sample size and ieQTL discovery in these tissues, suggesting that significantly larger numbers of ieQTLs may be discovered with larger sample sizes (fig. S3C).

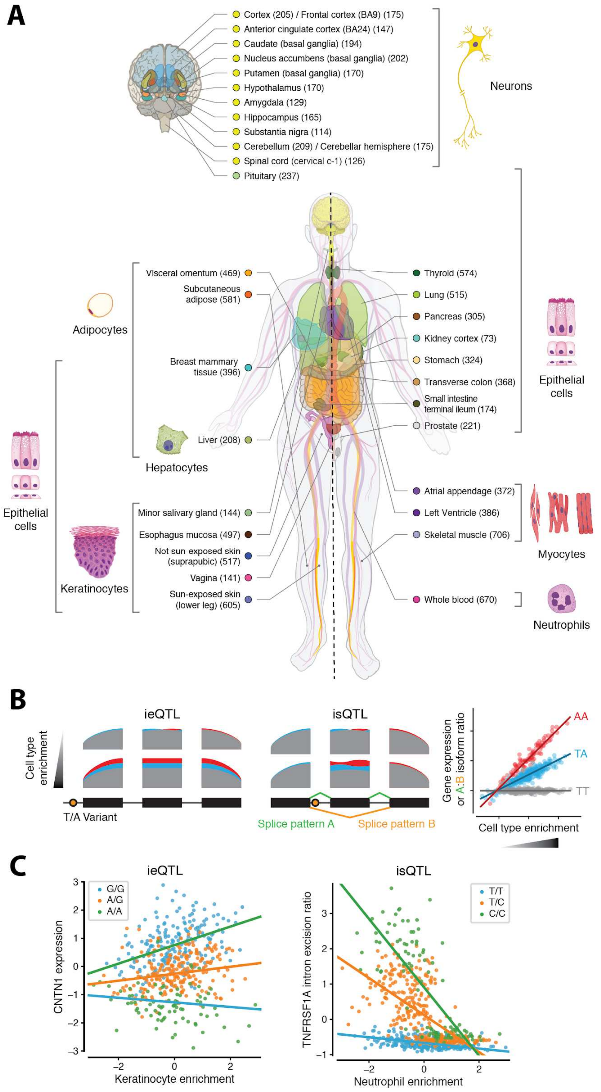


Fig. 1. Study design of mapping cell type ieQTLs and isQTLs in GTEx v8 project. (**A**) Illustration of 43 cell type-tissues pairs included in the GTEx v8 project. Cell types with median xCell enrichment score > 0.1 within a tissue were used (fig. S2). (**B**) Schematic representation of a cell type interacting eQTL and sQTL. (**C**) Example cell type ieQTL and isQTL. The *CNTN1* eQTL effect in not sun-exposed skin is associated with keratinocyte abundance (left panel). The *TNFRSF1A* sQTL effect in whole blood is associated with neutrophil abundance, but is only detected in samples with lower neutrophil abundances (right panel). Each data point represents an RNA-seq sample and is colored by the ieQTL and isQTL genotypes, respectively. The regression lines correspond to the coefficients of the interaction model.

Since external replication data sets are sparse, we used allele-specific expression (ASE) data of eQTL heterozygotes {Castel:QLGuYtcV} to correlate individual-level quantifications of the eQTL effect size (measured as allelic fold-change,aFC) with individual-level cell type enrichments. If the eQTL is active in the cell type of interest, we expect to see low aFC in individuals with low cell type abundance, while individuals with high cell type abundance are expected to have higher aFC (fig. S5A). Spearman correlation p-values can then be used to assess how many cell type ieQTLs show evidence of validation using this approach. The median proportion of ieQTLs with a significant aFC-cell type correlation ($P < 0.05$) was 0.63 (Fig. 2B). For 13 cell type-tissue combinations with > 20 significant ieQTLs (5% FDR), the corresponding $\pi 1$ statistic on the correlation p-values (*17*) confirmed the high validation rate (mean $\pi 1 = 0.76$, fig. S5B). While this approach does not constitute formal replication in an independent cohort, it is applicable to all tested cell type-tissue combinations, and corroborates that ieQTLs are not statistical artefacts of the interaction model. Next, we performed replication analyses in external cohorts, including whole blood from the GAIT2 study (*18*), purified neutrophils (*8*), adipose and skin tissues from the TwinsUK study for ieQTLs (*5*) and temporal cortex from the Mayo RNA-sequencing study for both ieQTLs and isQTLs (*19*). Overall replication was moderate to high ($\pi 1 = 0.32 - 0.67$) with the highest replication rates observed in purified neutrophils for whole blood (fig. S6A+E). The differences in replication rates likely reflect a combination of lower power to detect cell type ieQTLs/isQTLs compared to standard eQTLs/sQTLs, as well as differences in tissue heterogeneity across studies. Taken together, these results show that ieQTLs and isQTLs can be detected with reasonable robustness for diverse cell types and tissues.

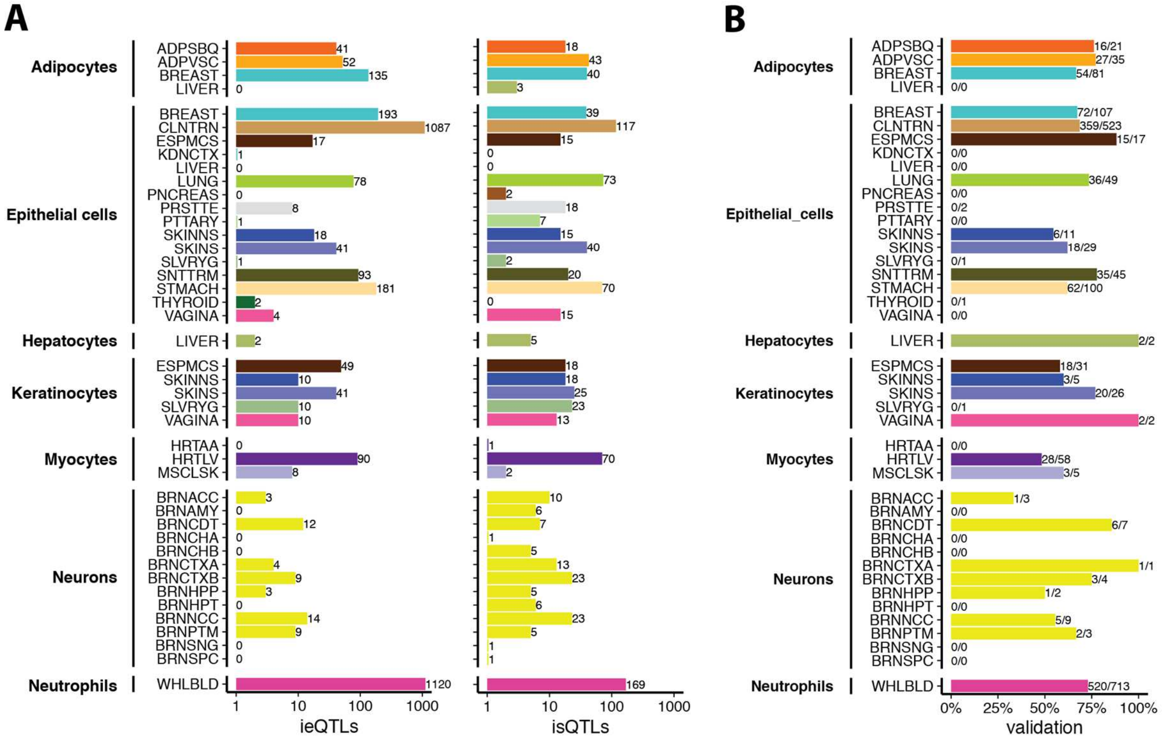


Fig. 2. Cell type ieQTL and isQTL discovery. (**A**) Number of cell type ieQTLs (left panel) and isQTLs (right panel) discovered in each cell type-tissue combination at FDR < 5%. Bar labels show the number of ieQTLs and isQTLs, respectively. See Fig. 1A for the legend of tissue colors. (**B**) Proportion of cell type ieQTLs that validated in ASE data. Validation was defined as ieQTLs for which the Spearman correlation between allelic fold-change (aFC) estimates from ASE and cell type estimates was significant ($p < 0.05$). Tissue abbreviations are provided in table. Bar labels indicate the number of ieQTLs with validation/number of ieQTLs tested].

Next, we sought to determine to what extent cell type ieQTLs contribute to the tissue specificity of *cis*-eQTLs. First, we analyzed ieQTL sharing across cell types, observing that ieQTLs for one cell type were generally not ieQTLs for other cell types (e.g., myocyte ieQTLs in muscle tissues were not hepatocyte ieQTLs in liver, etc.; fig. S7B). To determine if a significant cell type interaction effect correlates with the tissue-specificity of an eQTL, we tested whether cell type ieQTLs are predictors of tissue sharing. We annotated the top *cis*-eQTLs per gene (5% FDR) across tissues with their cell type ieQTL status (5% FDR) for the five cell types with at least 20 ieQTLs (adipocytes, epithelial cells, keratinocytes, myocytes, and neutrophils). This annotation was included as a predictor in a logistic regression model of eQTL tissue sharing based on eQTL properties including effect size, minor allele frequency, eGene expression correlation, genomic annotations, and chromatin state(*1*). In all five cell types, ieQTL status was a strong negative predictor of tissue-sharing, with the magnitude of the effect similar to that of enhancers, indicating that ieQTLs are an important mechanism for tissue-specific regulation of gene expression (Fig. 3A, fig. S7A). We corroborated this finding using multi-tissue eQTL mapping with mashr (*1*), testing whether eGenes that are tissue-specific (eQTLs discovered with LSFR < 0.05 only in the tissue/tissue type of interest) have a higher proportion of cell type ieQTLs compared to eGenes that are shared across tissues (LSFR < 0.05 in multiple tissues). Indeed, the proportion of cell type ieQTLs across all 43 cell type-tissue combinations was significantly higher in tissue-specific eGenes compared to tissue-shared eGenes (p = 1.9e-05, one-sided Wilcoxon rank sum test, Fig. 3B) further highlighting the contribution of cell type-specific genetic gene regulation to tissue specificity of eQTLs.

To examine the sharing patterns of cell type ieQTLs across tissues we used two cell types with ieQTLs mapped in >10 tissues (16 tissues for epithelial cells and 13 for neurons). We observed that while standard eQTLs were highly shared across the subsets of 16 and 13 tissues, cell type ieQTLs tended to be highly tissue specific, reflected by an average of four and five tissues with shared ieQTL effects compared to 11 and 12 for eQTLs in epithelial and brain tissues respectively (Fig. 3C+D, left panels). ~25% of neuron ieQTLs were shared between nine brain tissues, highlighting that tissues of the cerebrum (e.g., cortex, basal ganglia, limbic system) show particularly high levels of sharing compared to cerebellar tissues, the hypothalamus, and the spinal cord (Fig. 3D, left panel). This pattern was absent when analyzing standard eQTLs. Pairwise tissue sharing comparisons further confirmed that cell type ieQTLs showed greater tissue specificity and more diverse tissue sharing patterns than standard eQTLs, which were broadly shared across all tissues (Fig. 3C+D, middle and right panels). These results show that incorporating cell type composition is essential for characterizing the sharing of genetic regulatory effects across tissues.

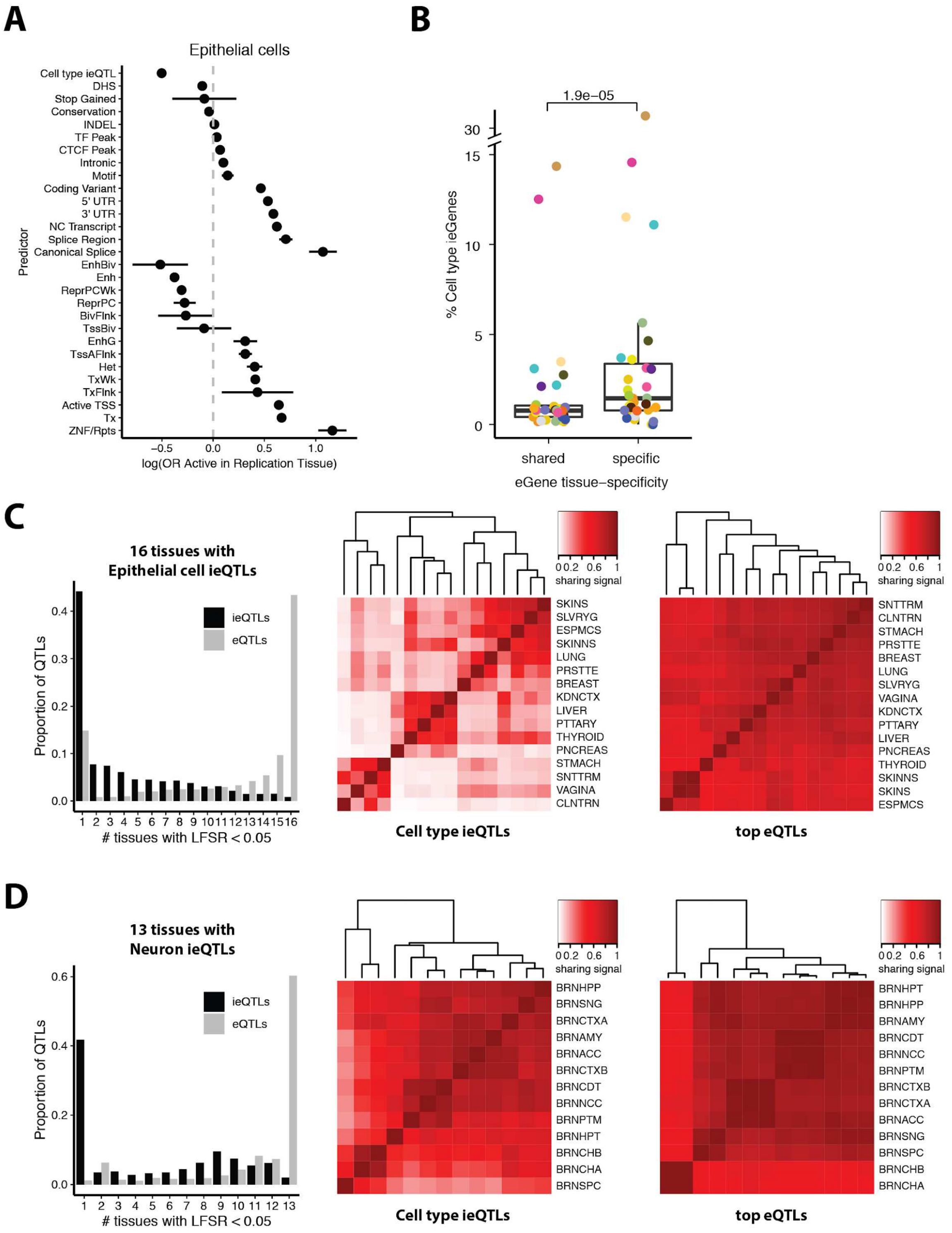


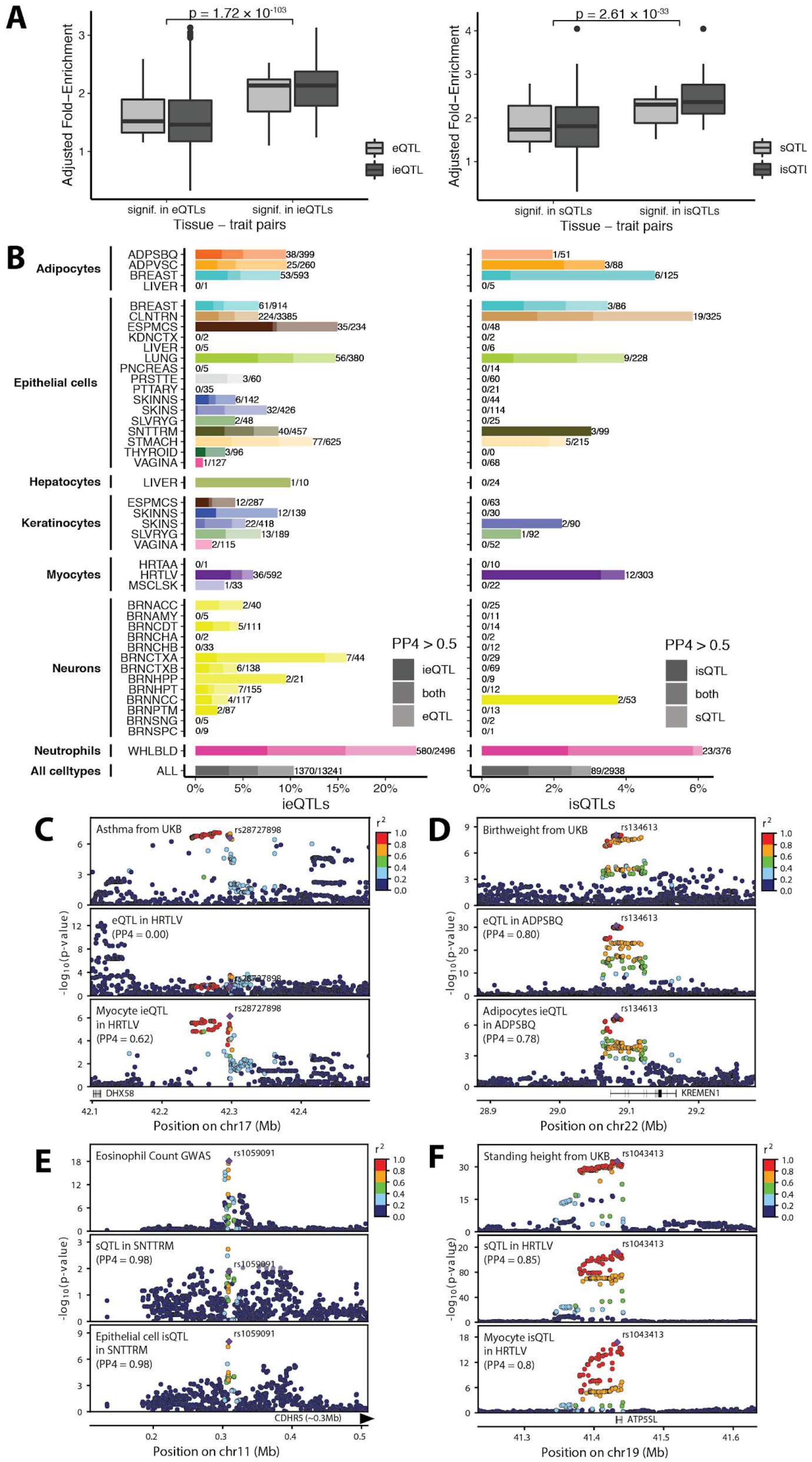
Fig. 3. Cell type ieQTLs contribute to *cis*-eQTL tissue specificity. (**A**) Coefficients from logistic regression models of *cis*-eQTL tissue sharing, using epithelial cell ieQTL status as a predictor. All significant (FDR < 0.05) top *cis*-eQTLs per tissue were annotated based on if they were also a significant (FDR < 0.05) ieQTL for a given cell type. The coefficients represent the log(odds ratio) that an eQTL is active in a replication tissue if it is an ieQTL. Chromatin states were defined using matched Epigenomics

Roadmap tissues and the 15-state ChromHMM (*20*). Genomic annotations, conservation, and overlaps with Ensembl regulatory build TF, CTCF, and DHS peaks are also included. Abbreviations of predictors are provided in table. Bars represent the 95% confidence interval. (**B**) Proportion of cell type ieQTL-genes (ieGenes) among tissue-specific and tissue-shared eGenes. An eGene is considered tissue-specific if its eQTL had a MASHR local false sign rate (LFSR, equivalent to FDR) < 0.05 only in the cell type ieQTL tissue (or tissue type) otherwise it is considered tissue-shared. Results of all 43 cell type-tissue combinations are shown. See Fig. 1A for the legend of tissue colors. (**C+D**) Tissue activity of cell type ieQTLs and eQTLs, where a cell type ieQTL and eQTL was considered active in a tissue if it had a LFSR < 0.05 (left panel). Pairwise tissue-sharing of ieQTLs (middle panel) or lead standard *cis*-eQTLs (right panel) respectively. The color-coded sharing signal is the proportion of significant QTLs (LFSR < 0.05) that are shared in magnitude (within a factor of 2) and sign between two tissues.

Given that a substantial fraction of cell type ieQTLs and isQTLs discovered were not detected as standard QTLs, we compared their role in complex traits relative to standard QTLs. We used QTLEnrich (*21*) to test 87 GWAS traits in 25 and 8 tissues with >100 ieQTLs or isQTLs, respectively (40% FDR). 314 tissue-trait pairs showed significant enrichment of GWAS variants among ieQTLs and 121 among isQTLs (Fig. 4A). We also computed GWAS enrichment for tissue-trait pairs with significant GWAS enrichment among standard QTLs (1106 trait-tissue pairs for eQTLs and 491 pairs for sQTLs) (*1*), and there was no significant difference between GWAS enrichment among cell type iQTLs and standard QTLs in matched tissue-trait pairs. However, GWAS enrichment was significantly higher in the subset of tissue-trait pairs significant in iQTLs compared to tissue-trait pairs significant in standard QTLs, which indicates that cell type specific regulatory effects on gene expression play an important role in mediating complex trait associations.

We next asked whether cell type iQTLs can be linked to loci discovered in genome-wide association studies (GWAS) as well as pinpoint the cellular specificity of these associations. To this end, we tested 13,702 ieGenes and 2,938 isGenes (40% FDR) for colocalization with 87 GWAS traits (*1*, *22*), using both the cell type ieQTL/isQTL and corresponding standard QTL. 1,370 (10.3%) cell type ieQTLs and 89 (3.7%) isQTLs colocalized with at least one GWAS trait (Fig. 4B). The larger number of colocalizations identified for neutrophil ieQTLs and isQTLs in whole blood relative to other cell type-tissue pairs likely reflects a combination of the larger number of ieQTLs and isQTLs and the abundance of significant GWAS loci for blood-related traits in our set of 87 GWASs. Our analysis revealed a substantial proportion of loci for which only the ieQTL/isQTL colocalizes with the trait (467/1370, 34%), or where the joint colocalization of the ieQTL/isQTL and corresponding standard eQTL indicates the cellular specificity of the trait as well as it's potential cellular origin (401/1370, 29%). For example, a colocalization between the *DHX58* gene in the left ventricle of the heart and an asthma GWAS was only identified through the corresponding myocyte ieQTL (PP4 = 0.64), but not the standard eQTL (PP4 = 0.00; Fig. 4B). Cardiac cells such as cardiomyocytes are not primarily viewed to affect the immune system. However, cardiomyocytes presence along pulmonary veins and their potential contribution to allergic airway disease have been previously described (*23*). An example where both the standard eQTL and the cell type ieQTL colocalize with the trait is given in Fig. 4C for *KREMEN1* in subcutaneous adipose tissue and a birth weight GWAS (PP4 ~0.8); *KREMEN1* has been linked to adipogenesis in mice (*24*). We highlight two analogous examples for isQTLs: the epithelial cell isQTL for *CDHR5* in small intestine colocalized with eosinophil counts whereas the standard sQTL did not (Fig. 4D), and conversely, both the standard sQTL and myocyte isQTL for *ATP5SL* in the left ventricle of the heart colocalized with standing height (Fig. 4E). Together, these results show that cell type interaction

QTLs are a powerful instrument for interpreting the genetic architecture and underlying cellular specificity and potential origins of complex traits.


A
p = 1.72 × 10^-103
Adjusted Fold−Enrichment
1
2
3
signif. in eQTLs
signif. in ieQTLs
Tissue − trait pairs
eQTL
ieQTL
p = 2.61 × 10^-33
Adjusted Fold−Enrichment
1
2
3
4
signif. in sQTLs
signif. in isQTLs
Tissue − trait pairs
sQTL
isQTl
B
Adipocytes
ADPSBQ 38/399 1/51
ADPVSC 25/260 3/88
BREAST 53/593 6/125
LIVER 0/1 0/5
Epithelial cells
BREAST 61/914 3/86
CLNTRN 224/3385 19/325
ESPMCS 35/234 0/48
KDNCTX 0/2 0/2
LIVER 0/5 0/6
LUNG 56/380 9/228
PNCREAS 0/5 0/14
PRSTTE 3/60 0/60
PTTARY 0/35 0/21
SKINNS 6/142 0/44
SKINS 32/426 0/114
SLVRYG 2/48 0/25
SNTTRM 40/457 3/99
STMACH 77/625 5/215
THYROID 3/96 0/0
VAGINA 1/127 0/68
Hepatocytes
LIVER 1/10 0/24
Keratinocytes
ESPMCS 12/287 0/63
SKINNS 12/139 0/30
SKINS 22/418 2/90
SLVRYG 13/189 1/92
VAGINA 2/115 0/52
Myocytes
HRTAA 0/1 0/10
HRTLV 36/592 12/303
MSCLSK 1/33 0/22
Neurons
BRNACC 2/40 0/25
BRNAMY 0/5 0/11
BRNCDT 5/111 0/14
BRNCHA 0/2 0/2
BRNCHB 0/33 0/12
BRNCTXA 7/44 0/29
BRNCTXB 6/138 0/69
BRNHPP 2/21 0/9
BRNHPT 7/155 0/12
BRNNCC 4/117 2/53
BRNPTM 2/87 0/13
BRNSNG 0/5 0/2
BRNSPC 0/9 0/1
Neutrophils
WHLBLD 580/2496 23/376
All celltypes
ALL 1370/13241 89/2938
0%
5%
10%
15%
20%
ieQTLs
0%
2%
4%
6%
isQTLs
PP4 > 0.5
ieQTL
both
eQTL
PP4 > 0.5
isQTL
both
sQTL
C
Asthma from UKB
rs28727898
eQTL in HRTLV
(PP4 = 0.00)
Myocyte ieQTL
in HRTLV
(PP4 = 0.62)
-log10(p-value)
DHX58
42.1
42.2
42.3
42.4
Position on chr17 (Mb)
r²
1.0
0.8
0.6
0.4
0.2
0.0
D
Birthweight from UKB
rs134613
eQTL in ADPSBQ
(PP4 = 0.80)
Adipocytes ieQTL
in ADPSBQ
(PP4 = 0.78)
-log10(p-value)
KREMEN1
28.9
29.0
29.1
29.2
Position on chr22 (Mb)
E
Eosinophil Count GWAS
rs1059091
sQTL in SNTTRM
(PP4 = 0.98)
Epithelial cell isQTL
in SNTTRM
(PP4 = 0.98)
-log10(p-value)
CDHR5 (~0.3Mb)
0.2
0.3
0.4
0.5
Position on chr11 (Mb)
F
Standing height from UKB
rs1043413
sQTL in HRTLV
(PP4 = 0.85)
Myocyte isQTL
in HRTLV
(PP4 = 0.8)
-log10(p-value)
ATP5SL
41.3
41.4
41.5
41.6
Position on chr19 (Mb)

]

Fig. 4. Cell type iQTLs improve GWAS-QTL matching. (A) GWAS enrichment based on GWAS summary statistics of the most significant iQTL or standard QTL per eGene/sGene with QTLEnrich. (**B**) Proportion of cell type ieQTLs (left panel) or isQTLs (right panel) with evidence of colocalization using COLOC posterior probabilities (PP4 > 0.5), for ieQTLs and isQTL at FDR < 0.4. Color saturation indicates if a trait colocalized with the cell type iQTL only (dark), the *cis*-QTL only (light) or both QTLs (medium). Bar labels indicate the number of cell type iQTLs with evidence of colocalization (either as iQTL or *cis*-QTL)/number of iQTLs tested. (**C**) Association p-values in the *DHX58* locus for a asthma GWAS (top panel), bulk heart left ventricle *cis*-eQTL (middle panel), and myocyte ieQTL (bottom panel). (**D**) Association p-values in the *KREMEN1* locus for a birth weight GWAS (top panel), bulk subcutaneous adipose *cis*-eQTL (middle panel), and adipocyte ieQTL (bottom panel). (**E**) Association p-values in the *CDHR5* locus for an eosinophil count GWAS (top panel), bulk small intestine terminal ileum *cis*-sQTL (middle panel), and epithelial cell isQTL (bottom panel). (**F**) Association p-values in the *ATP5SL* locus for a standing height GWAS (top panel), bulk heart left ventricle *cis*-sQTL (middle panel), and myocyte isQTL (bottom panel).

By mapping interaction effects between cell type enrichment and genotype on the transcriptome across GTEx tissues, we were able to identify thousands of eQTLs and sQTLs that are likely to be cell type specific. Notably, the ieQTLs and isQTLs we report here include immune and stromal cell types in tissues where cell type specific QTLs have not yet been characterized. Cell type ieQTLs are strongly enriched for tissue- and cellular specificity, and it is therefore likely that many more cell type ieQTLs remain to be discovered for cell types and tissues not considered in this study. Given that a large fraction of colocalizations with GWAS traits are only found with cell type ieQTLs, it will be essential to exhaustively characterize cell type specific QTLs to contribute towards a mechanistic understanding of these loci. However, the substantial allelic heterogeneity observed in standard eQTLs and limited power to deconvolve QTLs that are specific to rare cell types or with weak or opposing effects indicate that many more cell type specific eQTLs exist beyond those that can be computationally inferred from bulk tissue data. We therefore anticipate that single-cell QTL studies will be essential to complement the approaches presented here.

References:

1. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* (2019).

2. U. Võsa *et al.*, Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*, 1–57 (2018).

3. R. Joehanes *et al.*, Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18**, 16 (2017).

4. H. Kirsten *et al.*, Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci†. *Human Molecular Genetics*. **24**, 4746–4763 (2015).

5. A. Buil *et al.*, Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2014).

6. GTEx Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. **348**, 648–660 (2015).

7. B. P. Fairfax *et al.*, Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).

8. V. Naranbhai *et al.*, Genomic modulators of gene expression in human neutrophils. *Nat Commun*. **6**, 7545 (2015).

9. S. Kim-Hellmuth *et al.*, Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat Commun*. **8**, 266 (2017).

10. S. Kasela *et al.*, Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet*. **13**, e1006643–21 (2017).

11. M. G. P. van der Wijst *et al.*, Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

12. H.-J. Westra *et al.*, Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet*. **11**, e1005223–17 (2015).

13. D. V. Zhernakova *et al.*, Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2016).

14. C. A. Glastonbury, A. Couto Alves, J. S. El-Sayed Moustafa, K. S. Small, Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs. *Am. J. Hum. Genet.* **104**, 1013–1024 (2019).

15. G. Sturm *et al.*, Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*. **35**, i436–i445 (2019).

16. D. Aran, Z. Hu, A. J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).

17. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. **100**, 9440–9445 (2003).

18. L. Vila *et al.*, Heritability of thromboxane A2 and prostaglandin E2 biosynthetic machinery in a Spanish population. *Arterioscler. Thromb. Vasc. Biol.* **30**, 128–134 (2010).

19. M. Allen *et al.*, Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data*. **3**, 160089 (2016).

20. J. Ernst, M. Kellis, Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. **12**, 2478–2492 (2017).

21. E. R. Gamazon *et al.*, Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 1–18 (2018).

22. GTEx GWAS working group, Downstream consequences of genetic regulatory effects on complex human disease. *bioRxiv* (2019).

23. S. S. Folmsbee, C. J. Gottardi, Cardiomyocytes of the Heart and Pulmonary Veins: Novel Contributors to Asthma? *Am. J. Respir. Cell Mol. Biol.* **57**, 512–518 (2017).

24. C. Christodoulides *et al.*, The Wnt antagonist Dickkopf-1 and its receptors are coordinately regulated during early human adipogenesis. *J. Cell. Sci.* **119**, 2613–2620 (2006).

Acknowledgments:

We thank the donors and their families for their generous gifts of organ donation for transplantation, and tissue donations for the GTEx research project; Mariya Khan and Christopher Stolte for the illustrations in Figure 1.

Funding: The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH) and by the National Cancer Institute (NCI), the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), the National Institute on Drug Abuse (NIDA), the National Institute of Mental Health (NIMH) and the National Institute of Neurological Disorders and Stroke (NINDS).

This work was funded by following funding sources: Marie-Skłodowska Curie fellowship H2020 Grant 706636 (S.K-H.), HHSN268201000029C (F.A.,K.G.A.,A.S.) 5U41HG009494 (F.A.,K.G.A.), U01HG007593 U01HG007598 (M.O.,B.E.S.), 1K99HG009916-01 (S.E.C.), R01MH107666 (H.I.), P30DK020595 (H.I.), R01HG002585 (G.W.,M.S.), R01MH101814 (M.M-A.,V.W.,R.G.,D.G-M.,A.V.,E.T.D.), BIO2015-70777-P, Ministerio de Economia y Competitividad and FEDER funds (M.M-A.,V.W.,R.G.,D.G-M.), FPU15/03635, Ministerio de Educación, Cultura y Deporte (M.M-A.), R01MH109905 (A.Ba.), Searle Scholar Program (A.Ba.), EU IMI program (UE7-DIRECT-115317-1) (A.V.,E.T.D.), FNS funded project RNA1 (31003A_149984) (A.V.,E.T.D.), R01MH106842 (T.L.), R01HL142028 (T.L.,S.K.), R01GM122924 (T.L.,S.E.C.), UM1HG008901 (T.L.), R01GM124486 (T.L.).

Author contributions: S.K.-H., F.A. and T.L. conceived the study. S.K.-H. and F.A. led the writing, figure generation and editing of the manuscript and supplement. S.K.-H. coordinated analyses of all contributing authors; S.K.-H. and F.A. generated pipelines and performed iQTL mapping; S.K.-H., F.A., M.O., M.M.-A., V.W., D.G.-M., S.M., B.N., J.Q. performed cell type benchmarking analysis; S.K. performed ASE aFC validation with ASE data and validation pipeline generated by S.E.C; S.E.C. performed QTL tissue activity prediction analysis; F.A., A.V., A.R. performed replication analysis; A.H. performed QTLEnrich analysis; S.K.-H. performed tissue specificity, multi-tissue analysis and colocalization analysis; G.W. and Y.Z. provided software support for multi-tissue eQTL analysis; X.W. and H.I. provided advice on colocalization analysis; A.B., A.M.-P., J.M.-S. contributed to replication analysis; F.A. and K.G.A. generated and oversaw GTEx v8 data generation, LDACC & pipelines; A.Ba. and R.B. generated GWAS data; K.S., M.S., H.S.X., G.G., E.T.D., H.I., R.G., A.S., B.E.S., K.G.A., T.L. supervised the work of trainees in his/her lab; M.O. and T.L. contributed to editing of the manuscript; All authors read and approved the final manuscript.

Competing interests: F.A. is an inventor on a patent application related to TensorQTL; S.E.C. is a co-founder, chief technology officer and stock owner at Variant Bio; D.G.M. is co-founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme; H.I. has received speaker honoraria from GSK and AbbVie; E.T.D. is chairman and member of the board of Hybridstat LTD; G.G. receives research funds from IBM and Pharmacyclics, and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, POLYSOLVER and TensorQTL; T.L. is a scientific advisory board member of Variant Bio with equity, and Goldfinch Bio.

Data and materials availability:

All GTEx open-access data, including QTL summary statistics and visualizations, are available on the GTEx Portal (https://gtexportal.org). All GTEx protected data are available via dbGaP (accession phs000424.v8). Access to the raw sequence data is now provided through the AnVIL platform (https://gtexportal.org/home/protectedDataAccess). The QTL mapping pipeline is available at https://github.com/broadinstitute/gtex-pipeline, tensorQTL is available at https://github.com/broadinstitute/tensorqtl. Residual GTEx biospecimens have been banked, and remain available as a resource for further studies (access can be requested on the GTEx Portal, at https://www.gtexportal.org/home/samplesPage).

Material and Methods

GTEx data

The GTEx V8 data contains a total of 17,382 RNA-seq samples from 948 post-mortem donors, with 838 donors having genotype data from whole genome sequencing available in a phased analysis freeze VCF. QTL analyses were based on tissues with at least 70 RNA-seq samples from genotypes donors, corresponding to a total of 15,201 samples, and subsets of these data were used as detailed in the following sections.

Benchmarking of xCell cell type estimates

For neutrophils and neurons we compared xCell enrichment scores with estimates from CIBERSORT (*1*) using the built-in LM22 signature matrix or a custom signature matrix (*2*). For adipocytes, myocytes and keratinocytes, we compared xCell enrichment scores with estimates from the Gene Expression Deconvolution Interactive Tool (GEDIT) (*3*), using default settings and reference data from the Human Body Atlas (*4*) and Swindell et al. (*5*), which are both available on the GEDIT website (http://webtools.mcdb.ucla.edu). For epithelial cells, we compared xCell enrichment scores with estimates. In brief, we employed constrained linear models (lsqlincon function from the pracma R package) to perform cell type deconvolution from gene expression of 368 transverse colon samples. For hepatocytes we were not able to find suitable reference data to compare our xCell estimates against. Estimated cell type abundances were compared between methods using Spearman correlation.

For *in silico* simulations, we prepared synthetic cell type mixtures for colon, liver, adipose, skin, muscle, brain and blood tissues. The gene expression of cell types from colon, liver, adipose, skin, muscle was obtained from cell lines from the Human Protein Atlas (*6*), brain cell types were obtained from Zhang et al. (*2*) and blood cell types from Monaco et al. (*7*). xCell was then used to estimate the enrichment of these cell types in the *in silico* mixtures. We varied the proportion of cell types in each of the mixtures using linear combinations of gene expression profiles of each of the cell types. In total, we generated 100 mixtures for each of the tissues and compared the simulated ground truth to xCell enrichment scores using Spearman correlation.

Histology images were obtained from the GTEx Portal (https://gtexportal.org), and visually inspected for features matching estimated cell type abundances (e.g., presence of mucosal or muscular layers in transverse colon)

To further assess how well the xCell enrichment scores capture the proportion of specific cell types, we computed the Spearman correlation between the xCell scores and a list of cell type specific marker genes curated from the literature (figs. S1C and S1D). The following markers were used for each cell type: Adipocytes: FASN (https://www.ncbi.nlm.nih.gov/pubmed/11728175); Epithelial cells: CDH1, CLDN7 (https://www.ncbi.nlm.nih.gov/pubmed/27942595, https://www.ncbi.nlm.nih.gov/pubmed/14502431); Hepatocytes: AFP (https://www.ncbi.nlm.nih.gov/pubmed/16965562); Keratinocytes: KRT10 (https://www.ncbi.nlm.nih.gov/pubmed/27641957); Myocytes: MYH7, TNNI1 (https://www.ncbi.nlm.nih.gov/pubmed/30122443, https://www.ncbi.nlm.nih.gov/pubmed/25358788); Neurons: GAD1 (https://www.ncbi.nlm.nih.gov/pubmed/28846088); Neutrophils: STX3 (Zhernakova et al.) The xCell scores also captured differences in histology, as shown for the samples corresponding to the 5th and 95th percentiles of xCell scores in fig. S1C. Histology images and pathology notes were obtained from the GTEx portal (https://gtexportal.org).

xCell enrichment scores of 43 cell type-tissue combinations were tested for association with 37 histological phenotypes that were available for those tissues and curated applying language model on pathology reports of v8 GTEx samples (available via the GTEx portal). Phenotypes needed to have at least three annotated cases in the tissue of interest to be tested. Difference in cell type enrichment scores between cases and controls were assessed using two-sided Wilcoxon Rank-sum test.

xCell cell type enrichment in GTEx

Cell type enrichment scores were computed by running xCell on the full TPM gene expression matrix (from RNA-SeQC) of 17,382 RNA-seq samples from the GTEx V8 release, using the xCellAnalysis function from the R package.

Identification of cell type interacting eQTLs and sQTLs

Cell type interaction QTLs were mapped using used a linear regression model with an interaction term accounting for interactions between genotype and cell type enrichment:

$p \sim g + i + g \circ i + C$

where p is the phenotype vector (e.g., gene expression or intron excision ratio), g is the genotype vector, i is the inverse normal transformed xCell enrichment score, and the interaction term $g \circ i$ corresponds to point-wise multiplication of genotypes and cell type enrichment scores.

C is a matrix of covariates that were also used in regular QTL mapping. These covariates include genotype principal components to correct for population structure and PEER factors. Interaction QTLs were identified by testing for the significance of the interaction term, and mapping was performed using tensorQTL (*8*), which computes regression coefficients and p-values for all terms in the model, enabling comparisons of interaction and main effects. Variants within ±1Mb of the TSS of each gene were tested, as for regular QTL mapping. To avoid potential regression outlier effects, we restricted ieQTL mapping to variants with $\text{MAF} \geq 0.05$ in the samples belonging to each of the top and bottom halves of the enrichment score distribution, for each tissue-cell type combination (using the --maf_threshold_interaction 0.05 option in tensorQTL). For isQTL mapping, this threshold was set to $\text{MAF} \geq 0.1$. The same filtered and normalized gene expression and splicing phenotype matrices used for regular QTL mapping were used for interaction QTL mapping. To identify genes with at least one significant ieQTL or isQTL (ieGenes or isGenes, respectively), the top nominal p-values for each gene or phenotype was corrected for multiple testing at the gene level using eigenMT (*9*). Significance across genes was computed by adjusting the eigenMT-corrected p-values using Benjamini-Hochberg, and applying a 0.05 FDR threshold. For isQTLs, the p-value corresponding to the top splicing phenotype was selected for each gene-variant pair, and corrected by the number of phenotypes tested ($\tilde{p} = \min(n * p, 1)$, where n is the number of splicing phenotypes for the gene) prior to running eigenMT. QTL mapping and FDR correction were performed using expression and splicing phenotypes for all biotypes in the GENCODE v26 annotation, but downstream analyses are based on protein coding and lincRNA genes only.

Cell type ieQTL validation using aFC of allele-specific expression data

We used allele-specific expression (ASE) data of eQTL heterozygotes to correlate individual-level allelic fold-change (aFC) of an eQTL with individual-level cell type enrichments. For this analysis, we used the phASER haplotype-based ASE data for genes with ≥10 eQTL heterozygous individuals with ≥8 reads of ASE data per gene and nominally significant ASE aFC. The exact number of ieQTLs tested for ASE aFC validation after these filtering steps is indicated as bar labels in Fig. 2B. Since $\pi 1$ statistic For 13 out of 43 cell type-tissue combinations > 20 ieQTL at 5% FDR were available. Spearman correlation p-values were used to assess how many cell type ieQTL show evidence of validation. We report the proportion of ieQTLs with a significant aFC-cell type correlation ($P < 0.05$) for all tested cell type-tissue combinations (Fig. 2B). For 13 cell type-tissue pairs with > 20 ieQTLs at 5% FDR we also calculated the corresponding $\pi 1$ statistic on the correlation p-values (Fig. S5B).

Replication in external data sets

To assess replication of cell type ieQTLs and isQTLs, we examined p-values for matched variant-gene pairs in external cohorts where xCell cell type enrichment analysis and ieQTL and isQTL mapping was performed. Adipocyte and Keratinocyte ieQTLs were tested in TwinsUK adipose and skin tissues respectively (*10*). Neutrophil ieQTLs were tested for replication in whole blood from the GAIT2 study (*11*) and from purified neutrophils (*12*). Neuron ieQTLs and isQTLs were tested in temporal cortex from the Mayo RNA-sequencing study (*13*).

Multi-tissue ieQTL analysis

MashR (*14*) was used to estimate ieQTL activity across 16 epithelial tissues and 13 brain tissues respectively. We assessed the significance of the top ieQTL-SNP per gene (the SNP with the largest univariate |Z|-statistic across tissues) where ieQTL effect size and standard error was available in all tested tissues. MashR was run in the exchangeable Z (EZ) mode, and 250,000 randomly selected SNP*GENE pairs that were tested across all tissues were used to fit the mash model. Effect size estimates and local false sign rate (LFSR) outputted by MashR were used for QTL magnitude and activity respectively. For pairwise tissue-sharing analysis of ieQTL we considered only iQTLs that were significant (LFSR < 0.05) in at least one of the two tissues. We defined an ieQTL to be shared if they had the same sign and similar magnitude (effect within a factor of 2 of one another), which is implemented in the mashr function `get_pairwise_sharing`.

Tissue specificity analysis of ieQTLs

Modeling Determinants of eQTL Tissue Specificity – A logistic regression model of eQTL tissue activity was built to predict whether a eQTL identified in a given discovery tissue is active in a given replication tissue given a set of predictors derived from genomic annotations, and tissue specific gene expression, and chromatin states. eQTL activity was defined as MashR LFSR < 0.05 in a replication tissue. Basic QC on the eQTL data used to build the model was performed as follows: expression level > 0 in both discovery and replication tissues, eQTL MAF > 0.005 in both discovery and replication tissues, difference in expression level > quantile(0.005) and < quantile (0.995) to exclude the most extreme cases of expression difference. R v3.5.1 was used with speedglm v0.3-2. When plotting model predictor coefficients they were standardized using the standardize R package v0.2.1 so that they could be plotted on the same scale. When reporting AUCs for the model including different sets of features it was trained on eQTLs spanning chromosomes 1-20 and tested on eQTLS from chromosomes 21 and 22. Otherwise, tissue level AUCs were generated by holding out holding out individual tissues and predicting the activity of eQTLs found in the other 21 tissues in the held-out tissue. In total, 22 tissues were used for the analyses, which were chosen based on having appropriately paired epigenomic state predictions from the ROADMAP Epigenomics Project (*15*). In cases where there were two extremely similar tissues (defined by pairwise tissue gene expression clustering), the tissue with the higher sample size was used. Chromatin state sharing was defined as either shared on not shared based on if the ROADMAP Chromatin state prediction was the same (shared) or different (not shared) between the pairwise tissues. The ROADMAP core 15-state model was used (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html).

To test cell type ieQTLs as predictors of tissue specificity we only included cell types with > 20 ieQTLs (FDR > 0.05) in at least one of the tested tissues. Neurons and Hepatocytes had max. 14 and 2 ieQTLs (FDR > 0.05) respectively and were excluded from the analysis. Following predictors were also used in the model: distance between variant and TSS, variant MAF in GTEx, effect size in discovery tissue (aFC), global gene expression correlation between discovery and replication tissue, variant effect prediction, linsight conservation score (Huang, et al., 2017), variant is INDEL, Roadmap state and sharing between discovery and replication tissue, variant overlaps Ensembl Regulatory Build TF binding site (in any Ensembl tissue), variant overlaps Ensembl Regulatory Build CTCF binding site (in any Ensembl tissue), variant overlaps Ensembl Regulatory Build DHS site (in any Ensembl tissue), variant overlaps Ensembl Regulatory Build predicted motif site. The Ensembl Regulatory Build annotations were from Zerbino et al., 2015.

GWAS enrichment analysis

To test whether ieQTLs or isQTLs were enriched for GWAS hits, we applied an updated version of QTLEnrich (available at https://github.com/segrelabgenomics/eQTLEnrich). QTLEnrich assesses the enrichment of top ranked trait associations (GWAS p-value<0.05 used here) amongst a set of significant QTL-variants in a given tissue, accounting for potential confounding factors such as allele frequency, distance to the transcription start site and local level of LD (number of LD proxy variants; r2 ≥ 0.5). Briefly, an enrichment p-value was computed for each GWAS-tissue pair tested, as the fraction of 100,000 randomly sampled sets of null variants (of equal size to that of the QTL-variant set). Fold-enrichment was computed as the number of QTL-variants or null variants with a GWAS p-value below 0.05 divided by 5% of the variant set size. An adjusted fold-enrichment was computed for each QTL-variant set, as the fold-enrichment of the QTL-variant set divided by the median fold-enrichment of 1,000 randomly sampled sets of confounder-matched null variants. QTLEnrich was applied to 87 GWAS and GTEx tissues with at least 100 ieQTLs or isQTLs at 40% FDR testing 25 and 8 tissues respectively. GWAS enrichments among iQTLs at 40% and 10% FDR were comparable (Fig. Sx). To compare GWAS enrichment among iQTLs vs standard QTLs we matched tissue-trait pairs for standard QTLs at 5% FDR. Bonferroni correction was used to determine significant trait-tissue pairs, and the adjusted fold-enrichment was used as the test statistic to rank significant tissues based on their enrichment, as it corrects for enrichment of trait associations amongst matched null variants.

Cell type iQTL-GWAS colocalization analysis

Colocalization analysis was conducted using the coloc R package (*16*) Coloc uses summary statistics from QTL and GWAS studies in a Bayesian framework to identify GWAS signals that colocalize with QTLs. To maximize our discovery power, we ran coloc for all cell type ieGenes at FDR < 0.4 and 87 GWAS traits. All variants of the cis-QTL region (+/- 1 MB of the TSS of an ieGene) that were available for both the QTL and the GWAS trait were used in the function `coloc.abf()` with either cis-ieQTL or corresponding cis-eQTL p-values and GWAS effect size estimates and their variances. Given the high sensitivity of colocalization results to the choice of priors, we use model-based priors computed with enloc (*17*). The same model-based priors used for cis-eQTLs were used for cis-ieQTLs assuming that regular cis-eQTLs reflect the average signal of all ieQTLs for a particular gene. The corresponding prior can thus be interpreted as the average prior of all ieQTLs for that gene and can be used as an approximated prior for each individual cell type ieQTL. We defined an ieGene or eGene as having evidence of colocalization when posterior probability of colocalization (PP4) was higher than 0.5. All coloc results with PP4 ≥ 0.50 are reported in Supplementary Table XXX.

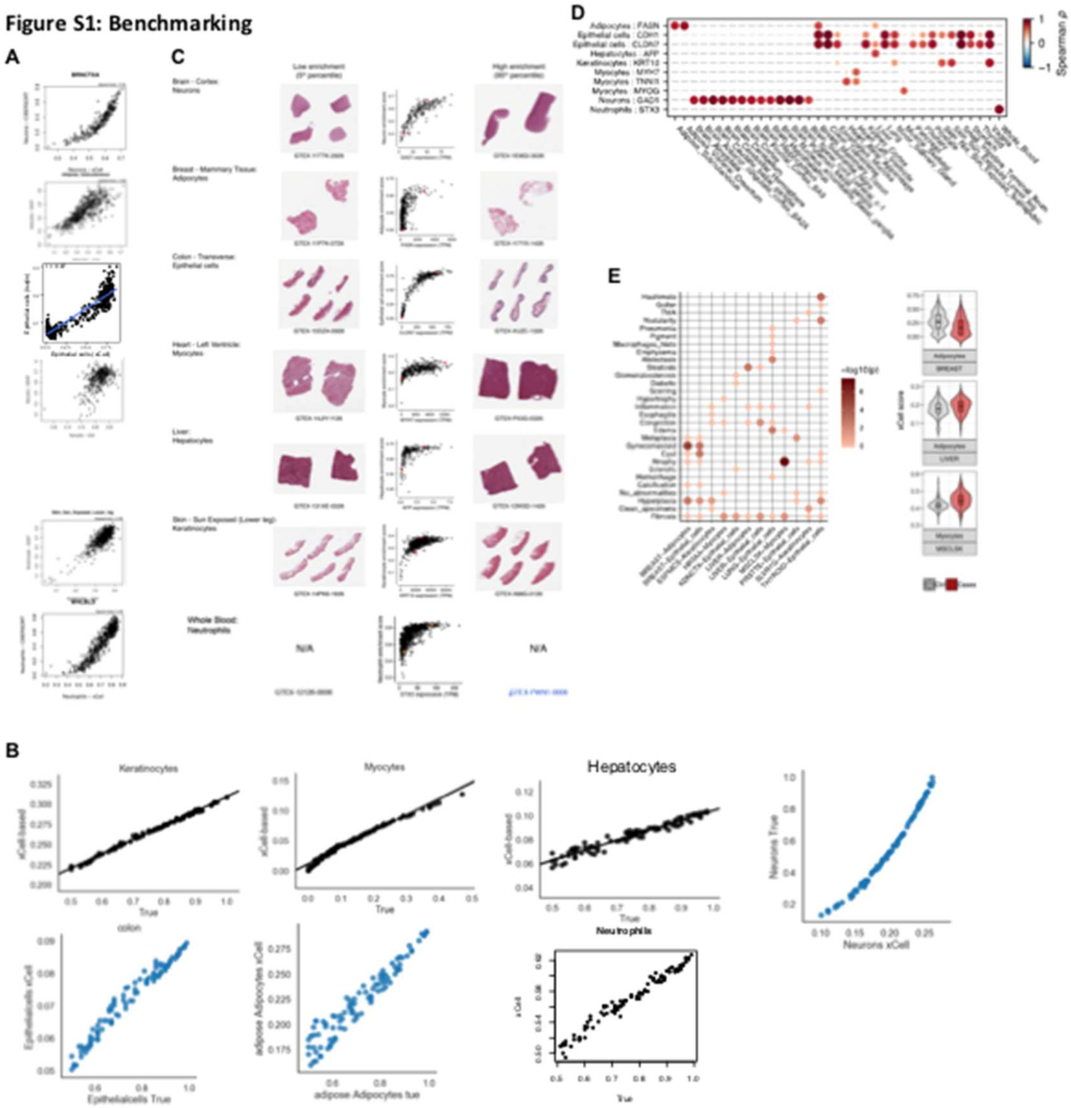


Fig. S1. Benchmarking of xCell cell type estimates. (**A**) Concordance rates between xCell and other available methods. Besides xCell estimates, neuron and neutrophil estimates were generated using Cibersort; adipocyte, keratinocyte and myocyte estimates were generated using GEDIT and epithelial cell estimates were generated using lsqlin. The corresponding tissue where the cell types were estimated are indicated above each plot. Sample sizes for each tissue are shown in the top left corner [will be added] and Spearman correlation coefficients are shown in the top right corner. (**B**) in silico simulation using synthetic cell type mixtures for colon, liver, adipose, skin, muscle, brain and blood tissues. (**C**) Histological correlates of samples with xCell estimates in the 5thand 95thpercentile. Scatterplot show correlation of xCell estimates and cell type marker genes. Red dots indicate the samples that were taken from the 5thand 95th percentile to illustrate histological correlates. (**D**) Spearman correlation between expression of cell type-specific marker genes (Supplementary Table X) and xCell enrichment scores for combinations of cell types and tissues used in this paper (other combinations are represented by gray lines). Normalized expression data for mapping QTLs was used (see Methods). (**E**) Cell type estimates of 42 cell type-tissue pairs were tested for association with 37 histological phenotypes that were available for those tissues and curated applying language model on pathology reports of v8 GTEx samples. Phenotypes needed to have at least 3 annotated cases in the tissue of interest to be tested. In rows, only phenotypes that were significantly associated in at least one cell type-tissue pair are shown. In columns, only cell type-pairs that were significantly associated with at least

one phenotype are shown. Color and size of each dot resembles Wilcoxon Rank-sum p-values. Three examples are shown in the right-hand side panel. Adipocytes in Breast tissue are depleted in gynecomastia samples compared to controls. Adipocytes in liver cases with steatosis are elevated compared to control. Myocytes in atrophic skeletal muscle are elevated.

Figure S2: xCell enrichment scores in GTEx

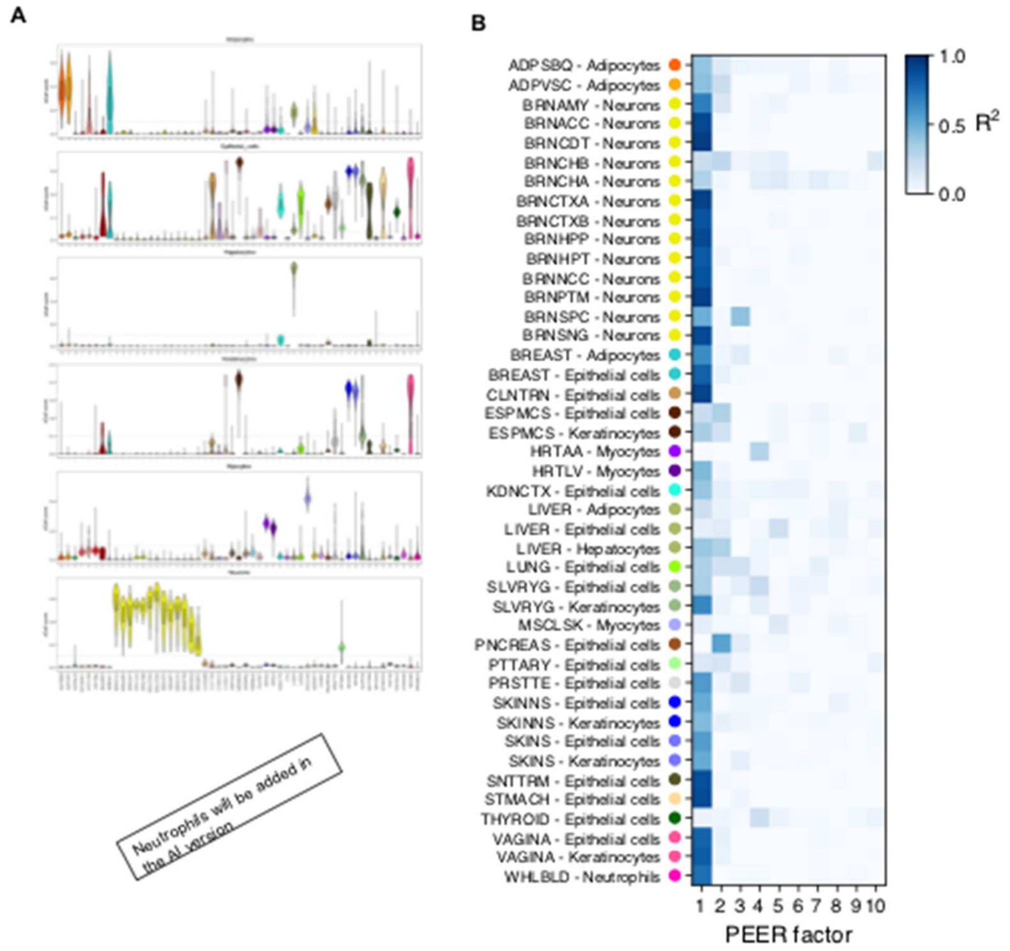


Fig. S2. xCell enrichment scores across 49 tissues. (A) Violinplots of xCell estimates of seven cell types across 49 tissues. For each cell type, interaction eQTL analysis was performed only in tissues where the cell type had a median xCell score > 0.1 (dashed horizontal line). (**B**) Correlation between xCell enrichment scores for the cell types indicated on the y-axis and the first ten PEER factors computed for each tissue, showing that the top PEER factors capture cell type heterogeneity.

Figure S3: ieQTL and isQTL discovery

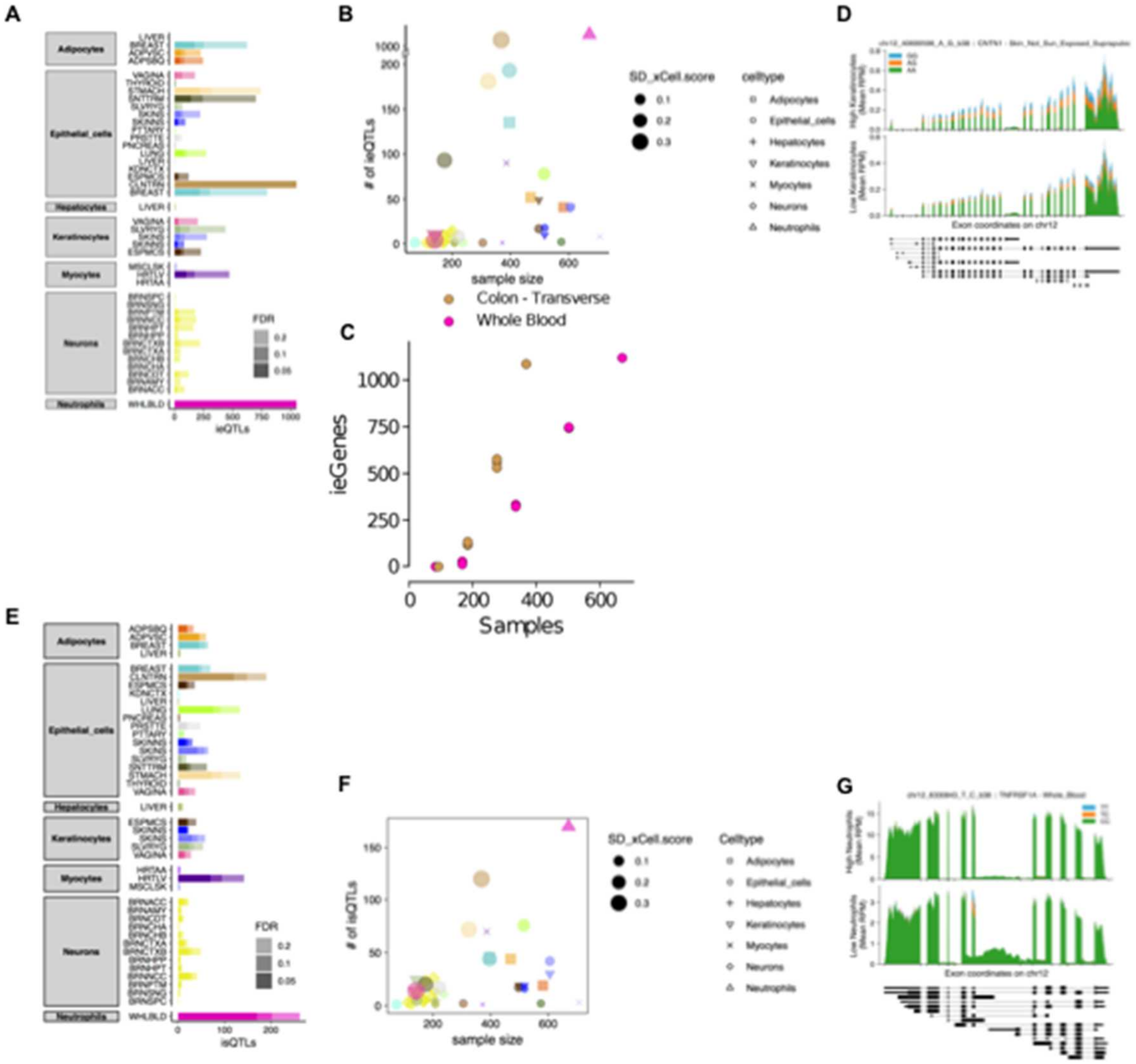


Fig. S3. Cell type ieQTL discovery. (**A**) Number of cell type ieQTL discovered in each cell type-tissue pair at indicated FDR thresholds. x-axis is truncated at 1000 for ease of display. Number of neutrophil-ieQTLs in Whole Blood were 1120/ 1380/ 1781 (FDR 0.05/ 0.1/ 0.2). Number of epithelial cell-ieQTLs in Colon Transverse were 1087/ 1513/ 2169 (FDR 0.05/ 0.1/ 0.2). (**B**) The number of cell type ieQTLs per tissue, as a function of sample size and variance of cell type estimates. The size reflects the standard deviation of cell type estimates in the corresponding tissue. Cell types are depicted as symbols. (**C**) Downsampling analyses in whole blood and transverse colon. (**D**) ieQTL pileup for CNTN 1 in Skin not sun exposed tissues. (**E**) Number of cell type isQTLs discovered in each cell type-tissue pair at indicated FDR thresholds. (**F**) The number of cell type isQTLs per tissue, as a function of sample size and variance of cell type estimates. (**G**) isQTL pileup for TNFRSF1A in Whole blood.

Figure S4: Correlation of ieQTLs and cell type estimates

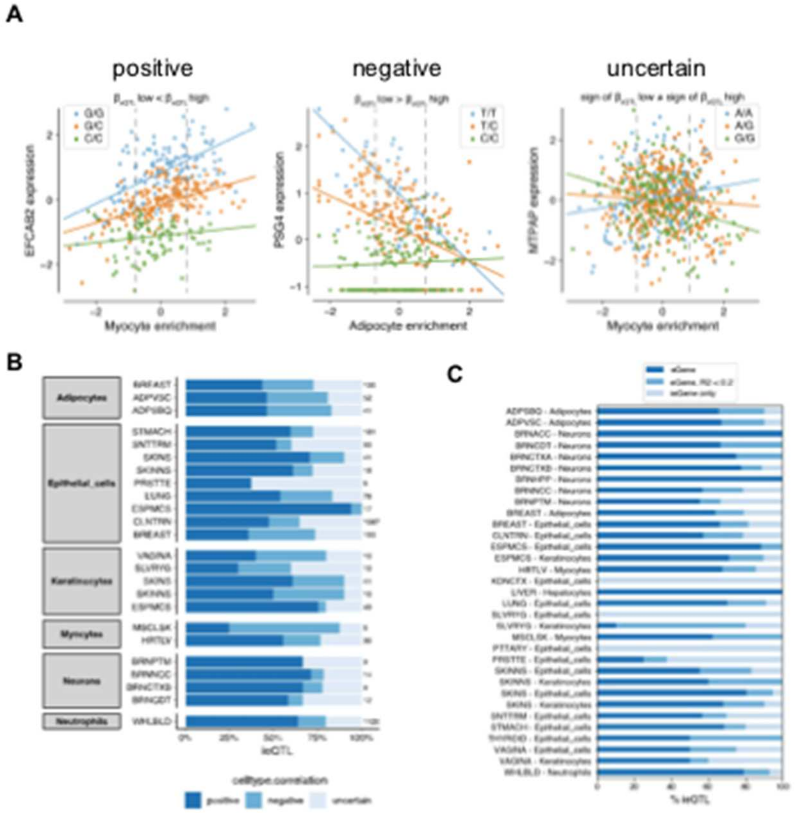


Fig. S4. Correlation of ieQTLs and cell type estimates. (**A**) ieQTL examples that are either positively (left) or negatively (middle) correlated with cell type estimates or where cell type correlation is uncertain (right). To categorize ieQTLs into these groups genotype main effects at low (25th percentile) vs high (75th percentile) cell type enrichment were compared. ieQTLs with "positive" cell type correlation show an increase of the genotype main effect from low to high cell enrichments. ieQTLs with "negative" cell type correlation show a decrease and the "uncertain" group contains ieQTLs where the sign flips between low and high cell type enrichments. (**B**) Stacked bar plots of proportion of cell type ieQTLs (min. 5 ieQTLs with FDR < 0.05) that show positive, negative or uncertain cell type correlation. Number at the end of each stacked bar plot indicate the total number of cell type ieQTLs per tissue at FDR 0.05. (**C**) Proportion of cell type ieQTLs discovered in 43 tissue-cell type pairs, with shading indicating whether the ieQTL was discovered by cis-eQTL analysis in bulk tissue.

Figure S5: replication using ASE aFC

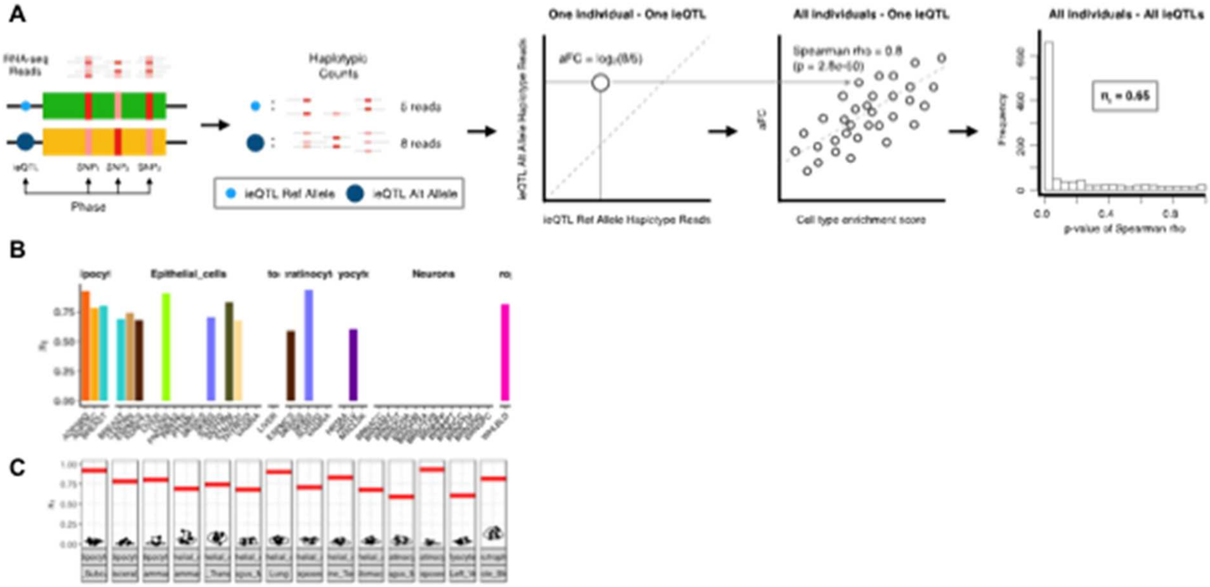


Fig. S5. Validation of cell type ieQTLs using correlation of ASE aFC and cell type abundance estimates. (**A**) Schematic representation of validation pipeline. (**B**) pi1 replication rate for cell type-tissue pairs with > 20 ieQTLs. (**C**) pi1 null distributions for 1000 randomly sampled ieQTLs.

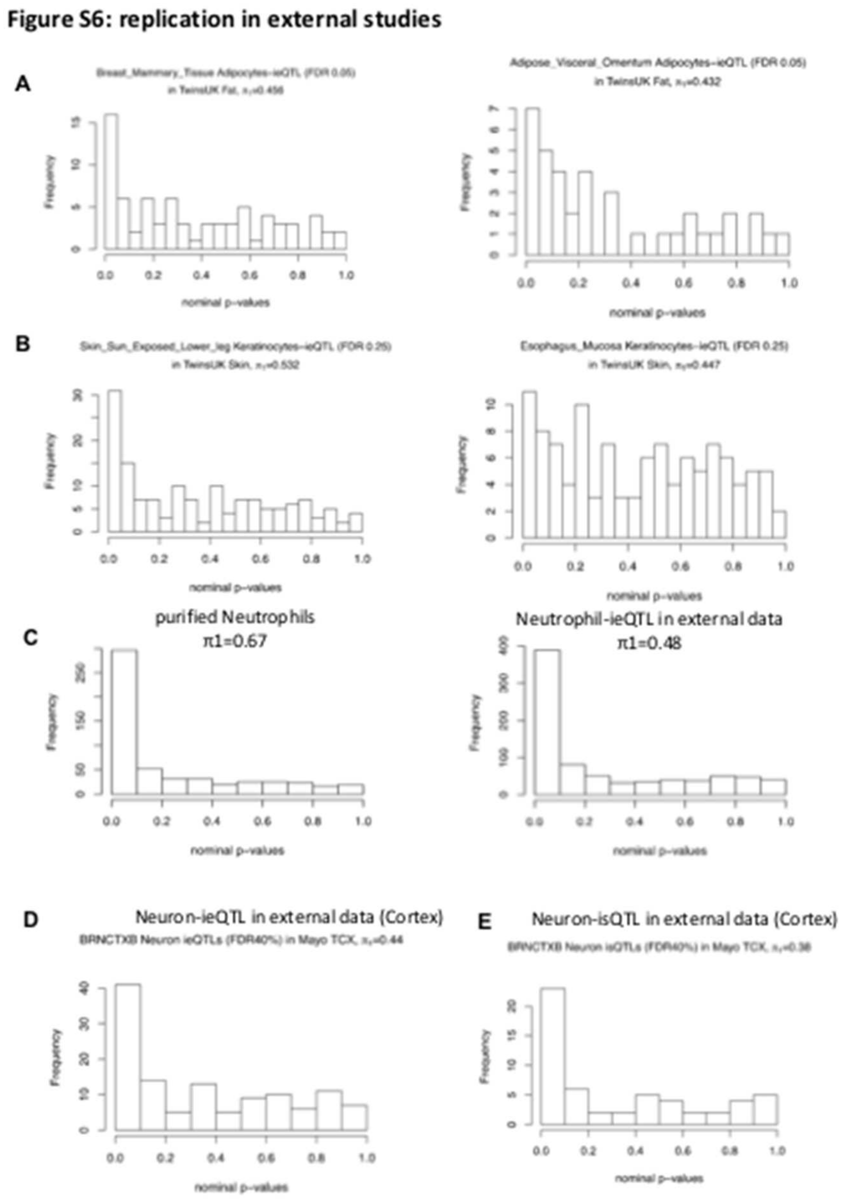


Fig. S6. Replication of ieQTLs and isQTLs in external studies. (**A**) [will be added]

Figure S7: Mechanisms of eQTL tissue-specificity

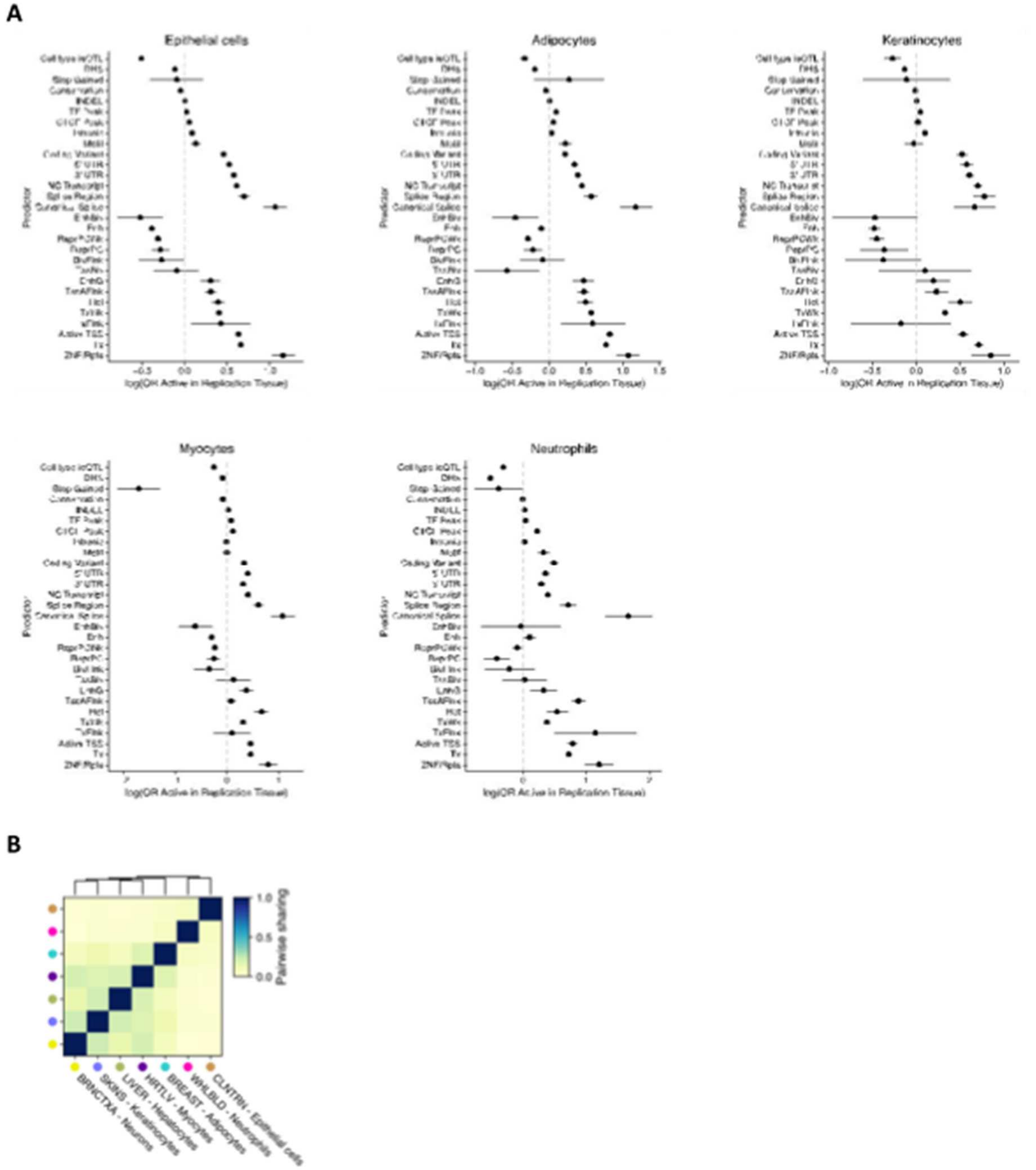


Fig. S7. Mechanism of eQTL tissue-specificity. (**A**) Coefficients from logistic regression models of cis-eQTL tissue sharing incorporating cell type ieQTL annotations. Models were built per cell type that was tested for ieQTLs. All top significant (FDR < 5%) cis-eQTLs per tissue were annotated based on if they were also a significant (FDR < 5%) ieQTL for a given cell type. The coefficients represent the log(OR) that an eQTL is active in a replication tissue if it is an ieQTL. Bars represent the 95% confidence interval. (**B**)Pairwise sharing by magnitude and sign of ieQTLs across seven different tissue-cell type pairs.

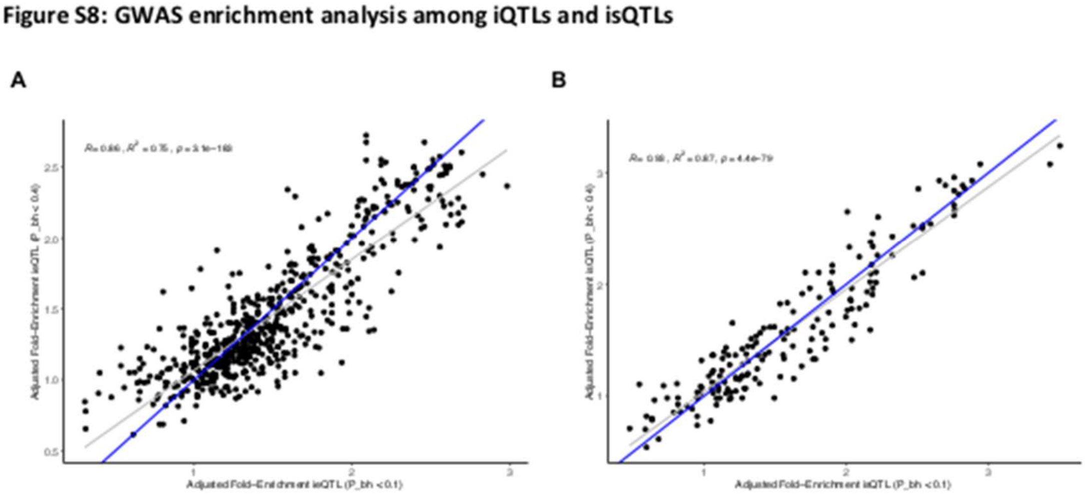


Fig. S8. QTLEnrich analysis of iQTLs. Comparison of GWAS adjusted fold-enrichment among ieQTLs (**A**) and isQTLs (**B**) at 40% FDR vs 10% FDR. Grey line represent the fit, blue line indicate the diagonal.

References

1. A. M. Newman *et al.*, Robust enumeration of cell subsets from tissue expression profiles. *Nat Meth*. **12**, 453–457 (2015).

2. Y. Zhang *et al.*, Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron*. **89**, 37–53 (2016).

3. B. Nadel *et al.*, The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data. *bioRxiv*. **14**, 395–30 (2019).

4. A. I. Su *et al.*, A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*. **101**, 6062–6067 (2004).

5. W. R. Swindell, A. Johnston, J. J. Voorhees, J. T. Elder, J. E. Gudjonsson, Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients. *BMC Genomics*. **14**, 527–20 (2013).

6. M. Uhlén *et al.*, Proteomics. Tissue-based map of the human proteome. *Science*. **347**, 1260419–1260419 (2015).

7. G. Monaco *et al.*, RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *CellReports*. **26**, 1627–1640.e7 (2019).

8. A. Taylor-weiner *et al.*, Scaling computational genomics to millions of individuals with GPUs. *bioRxiv*, 1–6 (2018).

9. J. R. Davis *et al.*, An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am. J. Hum. Genet.* **98**, 216–224 (2015).

10. A. Buil *et al.*, Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2014).

11. L. Vila *et al.*, Heritability of thromboxane A2 and prostaglandin E2 biosynthetic machinery in a Spanish population. *Arterioscler. Thromb. Vasc. Biol.* **30**, 128–134 (2010).

12. V. Naranbhai *et al.*, Genomic modulators of gene expression in human neutrophils. *Nat Commun*. **6**, 7545 (2015).

13. M. Allen *et al.*, Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data*. **3**, 160089 (2016).

14. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **5**, 1–15 (2018).

15. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* (2019).

16. C. Giambartolomei *et al.*, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. **10**, e1004383 (2014).

17. X. Wen, R. Pique-Regi, F. Luca, Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet*. **13**, e1006646–25 (2017).