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

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

[1]Bioinformatics Interdepartmental Degree Program, University of California Los Angeles, Los Angeles, CA

[2]Molecular Biology Institute, Department of Molecular Cellular and Developmental Biology, and Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, CA

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

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

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

*Corresponding Author

E-mail: brian.nadel@gmail.com

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 are also limited to particular platforms, whether RNA-Seq or specific microarray models. To overcome these limitations, we present the Gene Expression Deconvolution Interactive Tool, or GEDIT. Using simulated and experimental data, we demonstrate that GEDIT produces accurate results for both stromal and hematopoietic cell types. Moreover, GEDIT is capable of producing inputs using RNA-Seq data, microarray data, or a combination of the two. Finally, we provide reference data from 7 sources spanning a wide variety of stromal and hematopoietic types. GEDIT also accepts user submitted reference data, thus allowing deconvolution of any cell type, provided that accurate reference data is available.

Author Summary

The Gene Expression Deconvolution Interactive Tool (GEDIT) is a software tool that uses gene expression data to estimate cell type abundances. The tool accepts expression data collected from blood or tissue samples and sequenced using either RNA-Seq or microarray technology. GEDIT also requires reference data describing the expression profile of purified cell types. Several reference matrices are provided with this publication and on the tool's website (webtools.mcdb.ucla.edu), and the user also has the option to supply their own. The tool then applies a linear regression to predict which cell types are present in the tissue sample, and in what proportions. GEDIT applies several novel techniques and outperforms other tools on test data.

Introduction

Cell type composition is an important variable in both biology and medicine. In laboratory experiments, cell sample heterogeneity can act as a confounding variable. Observed changes in gene expression may be the result of 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 can correlate highly with prognosis [2, 3, 4]. Moreover, it has been shown that 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, these methods suffer from significant limitations. FACS sorting is extremely slow, with some samples requiring hours of highly skilled labor. Similarly, single cell RNA-Seq methods remain expensive for studies with large sample sizes. Moreover, some cell types, such as neurons, myocytes, and adipocytes, are difficult for these technologies to capture because of cell size and morphology.

Moreover, both FACS sorting and single cell methods can introduce biases, as these technologies require that samples be dissociated into single cell suspensions. Many stromal cell types are tightly connected to one another in extracellular matrices, and the procedures necessary to separate those cells can damage some, while others remain in clumps that are not sequenced. Consequently, subtle differences in sample preparation can produce dramatically different results (8). While it is possible to obtain pure samples of each cell type in this way, the observed cell counts may no longer represent the biology of the original sample. The recent Cell Population Mapping tool utilizes single cell reference data to perform bulk deconvolution, but requires that single cell data be available,

which is not always the case [9].

In recent years, digital means of cell type quantification, termed cell type deconvolution, have become popular. However, they remain approximate and are often limited to use on particular cell types or platforms. ImmQuant can estimate cell type fractions for immune cells, if supplied with the proper signature gene lists [10]. xCell can produce estimates for the 64 cell types supported by the tool, but does not allow the inclusion of additional cell types [11]. CIBERSORT is designed specifically for hematopoietic cell types sequenced using the HGU133A platform, and is not recommended for application to RNA-Seq or stromal data [12]. CIBERSORTX provides greater versatility but is slow to run compared to other tools [13].

To overcome some of the limitations of existing 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 7 distinct sources and used these data to generate thousands of synthetic mixtures. We then used these synthetic mixtures to test and refine the approaches and parameters used by GEDIT, in order to produce optimal results. Next, we run both GEDIT and competing tools on an *in vitro* mixture of immune cells and compare performance. Lastly, we use GEDIT to deconvolute two examples of human tissue samples: 21 skin samples from patients with skin diseases, and 17,382 samples of varied tissues from the GTEX database. The GEDIT tool is included in Supplementary File 1, and also available online at http://webtools.mcdb.ucla.edu/.

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 the current study, we have assembled or downloaded a set of 8 reference matrices, each containing the expression of 8-29 human cell types (Table 1). These data sources span multiple platforms, including bulk RNA-Seq, microarray, and single-cell RNA-Seq. Several of these matrices are novelly assembled from public sources as part of this study, and are included in Supplementary File 2. Complete details on the sources and assembly of these matrices are described in the methods [12,14–20].

Matrix	Platform	# of Cell Types	Cell Types Included
Human Skin Signatures [14]	Affymetrix Genome Plus 2.0/ Illumina Human HT-12 V3.0	21	Immune
Human Body Atlas [15]	Affymetrix U133A/GNF1H	13	Immune
Human Primary Cell Atlas [16]	Affymetrix U133 Plus 2.0	26	Immune and stromal
BLUEPRINT* [17]	RNA-Seq	8	Immune
ENCODE* [18]	RNA-Seq	29	Stromal, limited immune
BlueCode [17,18]*	RNA-Seq	35	Stromal and immune
LM22 [12]	Affymetrix Microarray	22	Immune, with subtypes
10x Single Cell Dataset* [19]	RNA-Seq	9	Immune
ImmunoStates [20]	Multi-Microarray	20	Immune, with subtypes

Table 1. Library of Reference Data assembled or obtained for GEDIT. Asterisk denotes matrices

117 novelly assembled as part of this publication. BlueCode represents the combined ENCODE and
118 BLUEPRINT matrices.
119
120 **Synthetic Mixture Generation**
121
122 To evaluate the accuracy of GEDIT, we first apply the tool to synthetic mixtures for which the
123 true cell type fractions are known. In order to meaningfully evaluate the performance of deconvolution,
124 we used one matrix to produce mixtures and another to serve as the reference. The deconvolution of
125 synthetic mixtures using only a single matrix (to both generate the mixtures and serve as a reference)
126 is a trivial problem. In this context, the linear regression will always return the exact (or nearly exact)
127 input proportions. Moreover, this is a poor simulation of real-world data, as in reality the expression
128 profile of any given cell type will vary to some extent between experiments. Moreover, cross-platform
129 effects cannot be simulated using a single matrix.
130 Using distinct reference and mixture-generating matrices requires that we match cell types
131 between the two matrices. Matching cell types across references is a non-trivial problem, as
132 equivalent cell types may be labelled differently, and identically labelled cell types may represent cells
133 in substantially different states or contexts. To address this problem, we defined the following
134 procedure for identifying pairs of equivalent cell types between two reference matrices:
135
136 1. Joint quantile normalize the matrices, then log transform them
137 2. Calculate the Pearson correlations between each cell in the first matrix and each cell in the
138 second matrix
139 3. Pair cell types that are more highly correlated with each other than with any other cell type in
140 the reference
141 4. Manually exclude cell pairings with mismatching descriptions
142
143 Using this procedure, we identified 5 pairings of reference matrices that can be used for the

144 generation of synthetic mixtures (Table 2). Since simulations can be performed in both directions, this
145 represents 10 possible choices of a mixture generating matrix and a reference matrix. The exact
146 matrices used to generate synthetic mixtures are available in Supplementary File 3.
147

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

148

149 Table 2. Pairs of reference matrices used to generate synthetic mixtures. For each pair of reference
150 matrices, 4-10 cell types were considered equivalent for the purpose of synthetic testing. For example,
151 “mature neutrophils” in BLUEPRINT and “neutrophils” in the Human Primary Cell Atlas were
152 considered equivalent.

153 For each of these 10 pairs of matrices, 1,000 cell type proportions were generated randomly.
154 Specifically, a cell type was selected at random and assigned a weight between 0 and 1.0 (randomly
155 sampled from the uniform distribution). Next, one of the remaining cell types is randomly selected and
156 assigned a weight between 0.0 and the remaining weight (1.0 minus the sum of weights already
157 assigned). This is repeated until the final cell type, which is assigned all remaining weight. The final
158 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 accurate mixtures, as they are composed primarily of a small number of cell types, with many other cell types present at low levels.

GEDIT Pipeline

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

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

In order to assess the effects of GEDIT's 4 parameter settings, which are described in detail below, we generated thousands of synthetic mixtures *in silico*. We then ran GEDIT on the simulated data described above while varying our 4 parameter settings (Table 3).

Input	Description	Allowed Values	Default Value
RefMat	Matrix of purified cell types	N by M matrix; One row per gene, one column per cell types	NA

MixMat	Matrix of mixtures to be deconvoluted	N by P matrix; One row per gene, one column per sample	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]	0.0

Table 3. GEDIT inputs include two matrices and four parameter settings.

Preprocessing and Quantile Normalization

The first step in the GEDIT pipeline is to render the two matrices comparable. This is done by including only genes present in both matrices and discarding all others. Genes with zero detected expression in all cell types as they contain no useful information for deconvolution. Each column of both matrices are then quantile normalized, such that they follow the same distribution; the target distribution is the starting distribution of the entire reference matrix.

Signature Gene Selection (SigMeth)

Starting with the normalized reference matrix, GEDIT identifies signature genes. Gene expression experiments can measure tens of thousands of genes simultaneously, but many of these genes are not informative 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 (and lowly expressed in the rest) are more informative. By using only such 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. scRNA-Seq cell type assignments).

We have implemented and tested a total of 6 signature gene scoring algorithms. For a each gene, these algorithms produce a signature score, using as input the vector of expression values across all cell types. Each gene is a candidate signature gene for the cell type in which it is most highly expressed, and for each cell type the NumSigs genes with the highest scores are accepted as signature genes. NumSigs is a tunable parameter with a default value of 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.

A final scoring method is the calculation of information entropy. Information entropy quantifies the amount of information in a probability distribution, with highly uniform distributions having the highest entropy. Entropy is minimized when expression is detected only in a single cell type and maximized when equal expression values are measured across all cell types. Thus, by ranking genes by negative entropy, genes with expression specific to a small subset of cell types will have high scores.

When run on 10,000 simulated mixtures, the entropy produced the lowest maximum, mean, and upper quartile error (Fig 2A). We therefore use entropy as the default setting but allow the user to select others. Unlike the other 5 selection methods tested, 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.

Fig 2. Effect of GEDIT parameter choices on accuracy of predictions in simulated experiments. 10,000 simulations 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 reference matrix. When not otherwise noted, parameters used were: signature selection method = entropy; number of signatures = 50, row scaling = 0.0; number of fixed genes = number of signatures.

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 (Fig 2B).

We provide an option that allows more signature genes to be selected 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, across all cell types a total of N*NumSigs genes are selected, and the remaining N*(NumSigs-MinSigs) genes are 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 (Fig 2C), and by default GEDIT sets MinSigs equal to NumSigs.

Row Scaling (RowScale)

One complication in the application of linear regression to gene expression data is the drastically different scale at which some genes are expressed. Take, for example, the two genes CD14 and THEMIS (Table 4). These have both been identified as strong signature genes: CD14 for monocytes and THEMIS for CD4+ T cells. However, CD14 is expressed at much higher levels in most

cell types. This is problematic because genes like CD14 will have a much larger impact on the estimation of cell type composition, compared to genes like THEMIS. That is, the possible penalty resulting from a poor fit of CD14 is much larger than the penalty from a poor fit of THEMIS.

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

Table 4. Example of two signature genes with drastically different magnitudes of expression. CD14 is a signature gene for monocytes, and THEMIS for CD4+ T cells. The row scaling transformation applied by GEDIT serves to lessen the dominating effect of highly expressed genes.

In order to equalize the effect of each signature gene on the linear regression, we implement a transformation we term 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 (Fig 2D). Values outside the natural range of 0.0 to 1.0 produce high error, as well (data not shown).

Comparison to Other Tools

To evaluate the performance of GEDIT on real data, and compare its results to those of other tools, we generated expression data from 12 *in vitro* mixtures of 6 immune cells using an Affymetrix array. We then selected 6 contemporary deconvolution tools (Table 4), ran the tools on our 12 mixtures and quantified their error. This represents an independent method of evaluating GEDIT and the other tools in the study.

Tool	Compatible with Outside Reference	Supported Platforms	Local Version Available	Supported Cell Types
GEDIT	Yes	RNA-Seq, microarray	Yes	Immune and Stromal
CIBERSORT	Yes	Microarray	Upon request	Immune
xCell	No	RNA-Seq, microarray	No	Immune and Stromal
ImmQuant	Yes	RNA-Seq, microarray	Yes	Immune
dtangle	Yes	RNA-Seq, microarray	Yes	Immune and Stromal
CIBERSORTX	Yes	RNA-Seq, microarray	Upon request	Immune and Stromal

Table 4. High level characteristics of 6 current deconvolution tools. All tools were used to estimate cell type fractions of 12 mixtures of immune cells, and their accuracy compared.

All tools in the study, except for xCell, require that the user submit a reference matrix. We ran each of these tools 4 times using 4 choices of reference matrix: The Human Primary Cell Atlas, LM22, ImmunoStates, and a reference constructed from BLUEPRINT data.

Unlike the other 4 tools, the outputs of xCell do not necessarily sum to 1.0 (in these cases, the total instead ranges from .9 to 2.8). Thus, for each sample we normalized the output vector by dividing each output by the sum of all outputs, such that predictions do sum to 1.0. Both the default output and normalized output were evaluated, with the renormalized output having notably lower error.

For each of the 4 tools that require a reference matrix, using the LM22 reference matrix yielded the most accurate overall results. Using BLUEPRINT, the only RNA-Seq reference, generally yielded high error. This may be due to issues associated with cross-platform analysis. Dtangle failed to run on BLUEPRINT altogether, producing either all zero values or a combination of zero and non-numeric

values, depending on whether the input data was normalized before using the tool.

In terms of average error across all cell types, the most accurate predictions were produced by GEDIT, using the LM22 reference matrix (Fig 3). These predictions were also the most accurate for 4 of the 6 cell types in the study, with the exceptions being CD4 and CD8 T Cells. For all cell types, there is a strong positive correlation between the GEDIT predicted proportion and the true proportion (Fig 4)

Fig 3. Average absolute error between true fractions and predicted fractions of an *in vitro* mixture of immune cells. We report results for each combination of tool, cell type, and reference matrix (xCell, which does not use a reference). For xCell the default output was taken, as well as the normalized output, where predictions were divided by the sum across all 6 cell types, such that predictions sum to 1.0. Dtangle failed to run when using BLUEPRINT as the reference.

Fig 4. Predicted vs actual proportions of current deconvolution tools when run on an *in vitro* mixture of 6 immune cells sequenced on an Illumina HT12 BeadChip microarray. For each tool, we use the reference matrix that minimizes total error on this dataset.

Dtangle produced inaccurate estimates for many cell types in the study. When using the LM22 matrix, the tool did not detect B cells, NK cells, or neutrophils in any samples, despite these being present in proportions of up to .194, .89, and .395 respectively. However, when using the LM22 matrix, dtangle produced by far the best estimates for CD8 T cells. The accuracy of these CD8 T cell predictions was highly dependent on the reference used, with the quality of predictions sharply declining when using either the Human Primary Cell Atlas or ImmunoStates.

All tools performed well on monocytes relative to the other cell types in the study. Error was also low for B Cells and Neutrophils, but these cell types were not as rigorously tested by this study, as each mixture contained no more than 20% or 40% of these cells, respectively. By contrast, most tools struggled to correctly predict CD4 and CD8 T Cells. This demonstrates the difficulty of

distinguishing highly similar cell types.

Skin Expression Data

We also used GEDIT to analyze a set of skin biopsies from patients with various skin diseases. The predicted cell types are consistent with skin biology; in most samples, keratinocytes are the most highly predicted followed by subcutaneous adipose (Fig 5). Deviations from this pattern correspond to disease biology. Monocytes are highly predicted in Stevens-Johnson syndrome, a sample collected from blister fluid. Macrophages are known to be abundant in granulomas of leprosy lesions and are predicted to be abundant in the 3 leprosy samples. T cells are most abundant in the T Cell Lymphoma sample.

Fig 5. GEDIT predictions when run on 21 samples of various skin diseases. GEDIT identifies keratinocytes and subcutaneous adipose as the most common cell types. Deviations from this pattern correspond to disease biology. The Steven Johnson Syndrome sample was collected from blister fluid and is predominantly immune cells. L-Leprosy and Leprosy Reversal Reaction are known to result in large numbers of macrophages, and macrophages are predicted to be highly abundant in these samples. Mycosis Fungoides is a T Cell Lymphoma and thus the high numbers of predicted T Cells conform to biological expectation.

GTEX

We also applied GEDIT to 17,382 GTEX RNA-Seq samples collected from various tissues. However, no single reference contained all cell types we wished to predict. For example, none of the available references contains both myocytes and adipocytes, (Supplementary Fig 1). Therefore, we took a novel approach in which we predicted proportions 3 times using 3 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 present in any one reference matrix (Fig 6).

Fig 6. GEDIT cell type predictions when applied to 17,382 samples from the GTEX database. Here, predictions have been averaged for each tissue of origin.

Supplementary Fig 1. Cell types present in the 3 reference matrices used to predict cell type fractions of GTEX samples. The Skin Signatures matrix contains entries for both lymphatic and vascular endothelial cells.

These predictions conform to biological expectations. For example, immune cells are predicted at high abundance in blood and spleen, adipocytes are highly predicted in adipose tissue, Schwann cells in nerve and heart, and keratinocytes in skin. All these patterns match expectations of which cell types are be present in these tissues. Neither cardiac myocytes nor smooth muscle are highly predicted 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.

Online Tool

GEDIT is available online at http://webtools.mcdb.ucla.edu/. We provide access to the tool, as well as an array of 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 4 parameters of GEDIT, and may adjust them as desired (signature gene selection method, number of signature genes, row scaling).

Methods

Reference Data Assembly

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 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.

106 transcript quantification files were downloaded from the ENCODE database [18]. These included all total RNA-Seq experiments collected from adult primary cells, excluding 4 with warnings (3 low replicate concordance, 1 low read depth). All samples were processed by the Thomas Gingeras lab at Cold Spring Harbor and mapped to GRCH38. The samples were quantile normalized, then clustered, and 18 samples were excluded as they did not cluster with their replicates. The remaining 88 samples were merged (via median) in accordance with clustering and sample descriptions, resulting in reference profiles for 28 cell types. For example, 19 samples labelled as endothelial cells, collected from various body locations, formed a cluster and were merged into a single entry we termed canonical endothelial cells. Where multiple transcripts were measured for a single gene, the expression of that gene was calculated as the sum of those transcripts. This dataset spans a wide range of stromal cell types (smooth muscle, fibroblast, epithelial, etc..), but contains only a single entry for blood cells, labelled mononuclear cells.

We also combined the ENCODE and BLUEPRINT reference matrices into a single reference matrix, which we call BlueCode. This was done by combining the columns of both matrices, then quantile normalizing them. This combined reference spans both blood cell types and a wide range of stromal cell types. Possible batch effects in this combined matrix have not been fully evaluated.

We obtained single cell expression data for 9 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.

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 noise, and generally poor signature genes. Moreover, including zeros can result in unusually high signatures or in mathematical errors, such as dividing by zero or 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:

$$H = -\sum[p_i * log_2(p_i)] \qquad (1)$$

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 a mixture consisting of equal fractions of each cell type, p_i can be interpreted as the probability an observed read came 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 the 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}.

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.

***In Vitro* Immune Cell Mixture**

Combinations of 6 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, 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

448 performed by the UCLA Neuroscience Genomics Core. Data was normalized by quantile
449 normalization through R 'normalize.quantiles' function (R Core Team, 2013).
450
451 **Comparison to Other Deconvolution Tools**
452
453 We deconvolved our *in vitro* mixture of immune cells using 5 tools (GEDIT, CIBERSORT ,
454 CIBERSORTX, xCell, and dtangle) [2,11,12,21] and 4 reference matrices (BLUEPRINT, Human
455 Primary Cell Atlas, LM22, ImmunoStates; [12,16,17,20]. Cell types other than the 6 present in the
456 mixture were excluded from the reference matrices before used as input to each tool. The LM22
457 reference contains 2 types of B cells (naive and memory), 2 types of Natural Killer Cells (active,
458 resting) and 4 types of CD4 T Cells. After running each tool, the predictions for these cell types were
459 summed to produce a final prediction for B cells, NK cells, and CD4 T cells, respectively. Similarly, the
460 ImmunoStates reference contains 2 types of NK cells (bright and dim) and 2 types of B Cells (memory
461 and naive); predictions for these cell subtypes were summed to produce final predictions for NK cells
462 and B cells, respectively.
463
464 xCell produces 67 output scores, 13 of which were used in this study. These were the entries
465 labelled "B-Cells", "Monocytes", "NK cells", "Neutrophils", 5 subtypes of CD4 T cells, and 4 subtypes of
466 CD8 T Cells. The outputs for CD4 and CD8 subtypes were summed to produce a final output. These
467 outputs did not sum to 1.0, with that sum instead ranging from 0.9 to 2.8. Thus, we normalized each
468 sample, dividing each output by the sum of all outputs, such that predictions do sum to 1.0. When
469 comparing to other tools, both the default output and normalized output were evaluated.
470
471 **GTEX Data**
472
473 GTEX data for 17,382 samples were obtained from the GTEX database
474 (https://gtexportal.org/). We ran GEDIT on all samples 3 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 3 sets of predictions (or fewer, if that cell type is missing from 1-2 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%.

Conclusion:

The Gene Expression Deconvolution Interactive Tool offers a new option for cell type quantification. GEDIT produces accurate results in both simulated and *in vitro* mixtures, outcompeting other contemporary tools. Moreover, GEDIT offers flexibility because it can be applied to any cell type, provided the proper reference data. GEDIT also accepts data generated from microarray, bulk RNA-Seq, or scRNA-Seq, and supports cross-platform compatibility. Lastly, we present with our tool a comprehensive library of reference matrices. This includes data assembled from 8 distinct sources, spanning a wide range of cell types and platforms. Some of these have been previously published, while others were novelly assembled.

We extensively tested GEDIT on several large public datasets. When applied to skin biopsies, keratinocytes are found to be the most abundant cell type, as expected. However, 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.

Compared to other tools, GEDIT produces accurate results when tested on mixtures of human immune cells. GEDIT produced the lowest error both overall and for 5 of the 6 cell types in the mixtures. Moreover, GEDIT provides increased flexibility over these other tools, in that it can be applied to a greater number of cell types and platforms.

While single cell RNA-Seq is an emerging approach, these methods are not always capable of accurately quantifying cell type populations, due to biases associated with the capture of different cell

types. However, the pure reference profiles they produce can be used by GEDIT to produce accurate estimates of cell type populations. This approach circumvents some of the biases associated with the preparation of samples for both scRNA-Seq and FACS. Moreover, it is more economical, particularly when researchers have already collected bulk RNA-Seq data for other purposes.

Acknowledgments

We acknowledge the Biomedical Big Data Grant (5T32LM012424-03) for supporting Brian Nadel during this research. We also acknowledge the Bruins in Genomics Program for supporting Hannah Waddel and Misha Khan during the summer of 2017, when they contributed to this work. Special thanks to Erin Nadel for assistance with figure preparation.

References

1. Bolen CR, Uduman M, Kleinstein SH. Cell subset prediction for blood genomic studies. BMC Bioinformatics. 2011 Jun 24;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 Aug;21(8):938–45.

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 Aug 22;17(1):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 Mar 15;12(4):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 Nov 17;17(1):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 [Internet]. Vol. 14, Nature Methods. 2017. p. 395–8. Available from: http://dx.doi.org/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 May 21;161(5):1202–14.

8. Hines WC, Su Y, Kuhn I, Polyak K, Bissell MJ. Sorting out the FACS: a devil in the details. Cell Rep. 2014 Mar 13;6(5):779–81.

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 Apr;16(4):327–32.

10. 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 Dec 15;32(24):3842–3.

11. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017 Nov 15;18(1):220.

12. 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 May;12(5):453–7.

13. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol [Internet]. 2019 May 6; Available from: http://dx.doi.org/10.1038/s41587-019-0114-2

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 Aug 1;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 Apr 20;101(16):6062–7.

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 Sep 20;14:632.

17. Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica. 2013 Oct;98(10):1487–9.

18. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004 Oct 22;306(5696):636–40.

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 Jan 16;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 Nov 9;9(1):4735.

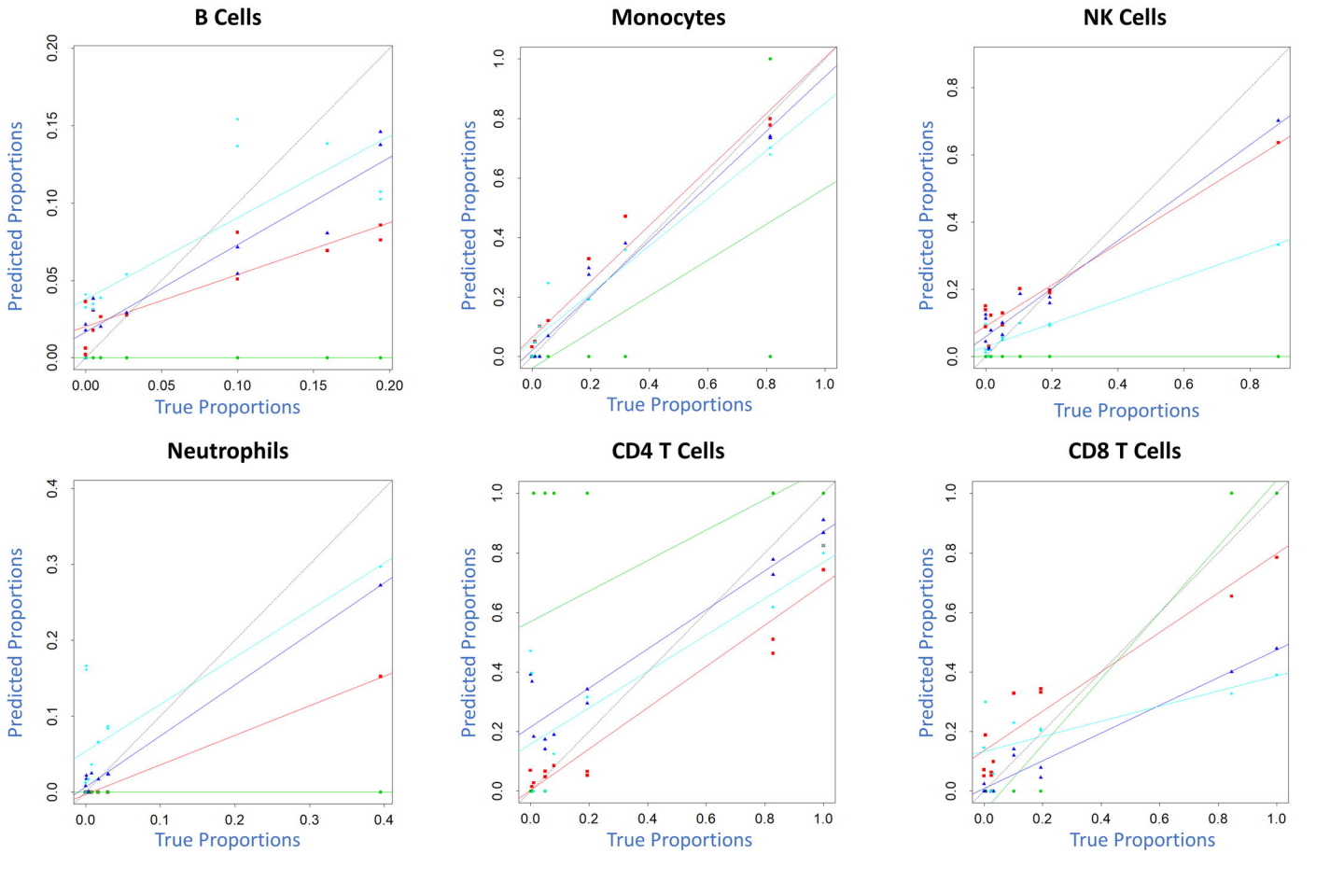
21. Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. dtangle: accurate and robust cell type deconvolution. Bioinformatics [Internet]. 2018 Nov 8 [cited 2019 Jan 17]; Available from: https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty926/5165376

Supplementary File 1. Local version of GEDIT tool.

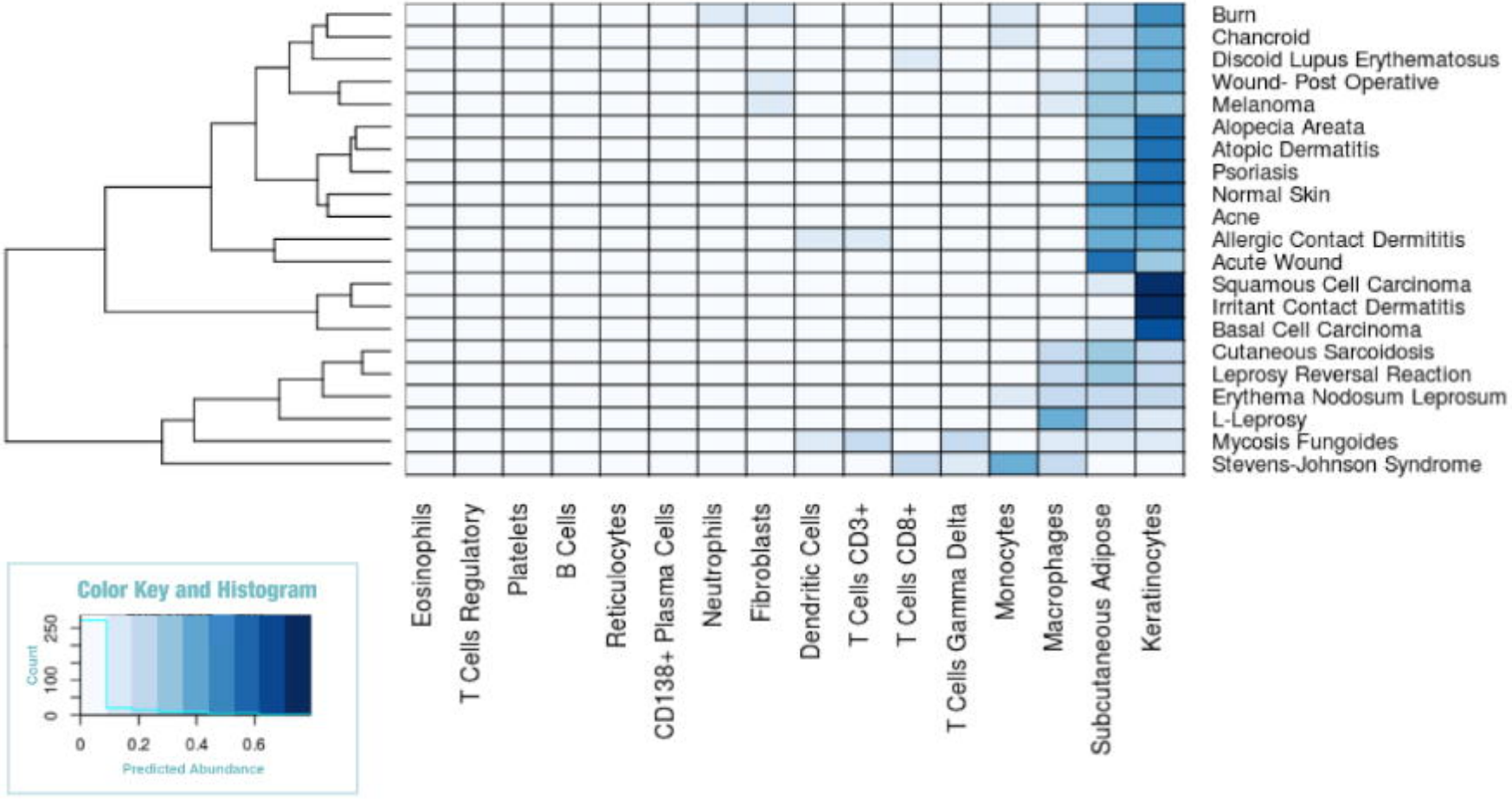
Supplementary File 2. Novel reference matrices assembled as part of this publication.

Supplementary File 3. Pairs of matrices used to generate synthetic mixtures.

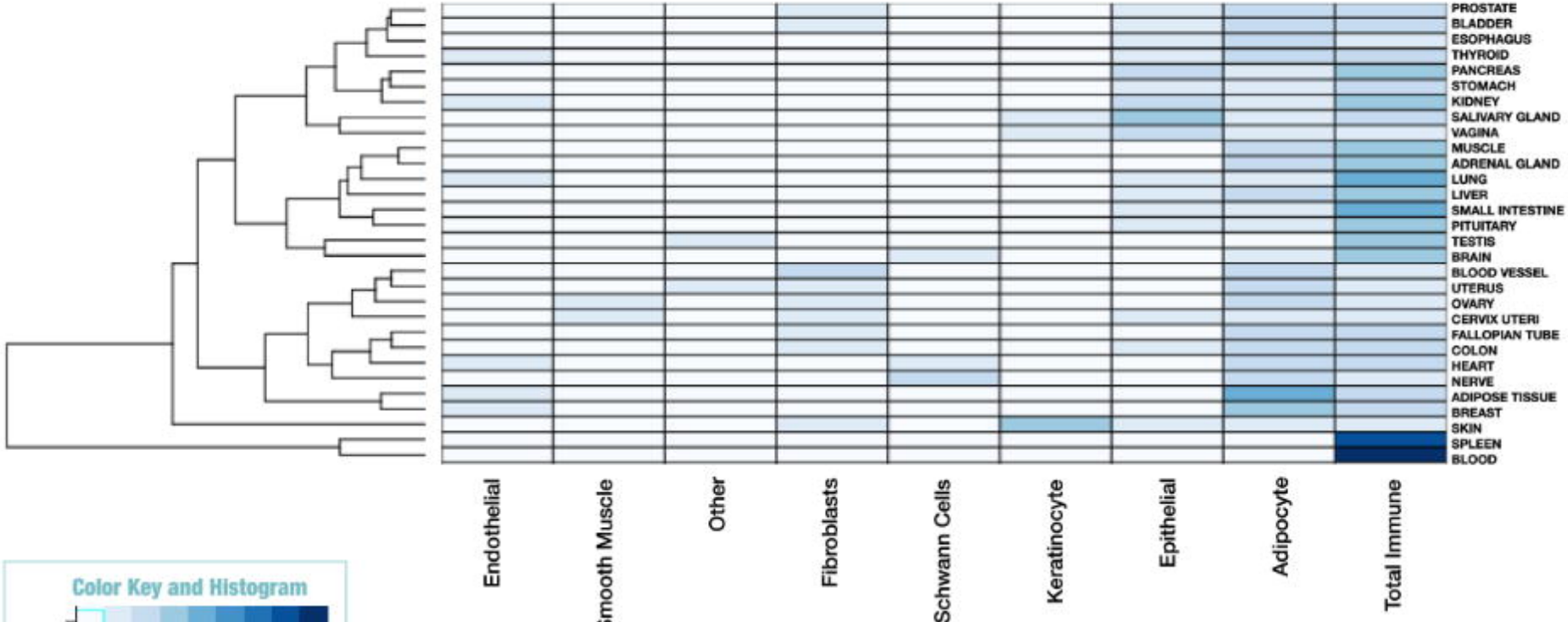
Supplementary Figure 1. Cell Types present in reference matrices used for GTEX data.

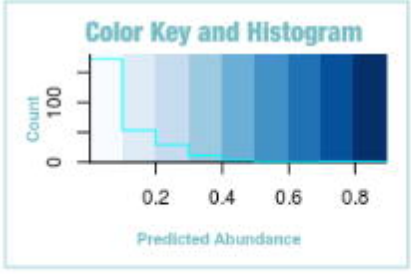

B Cells
Monocytes
NK Cells
Neutrophils
CD4 T Cells
CD8 T Cells
Predicted Proportions
True Proportions

Predicted Cell Types of Skin Diseases



Predicted Cell Types of GTEX Samples

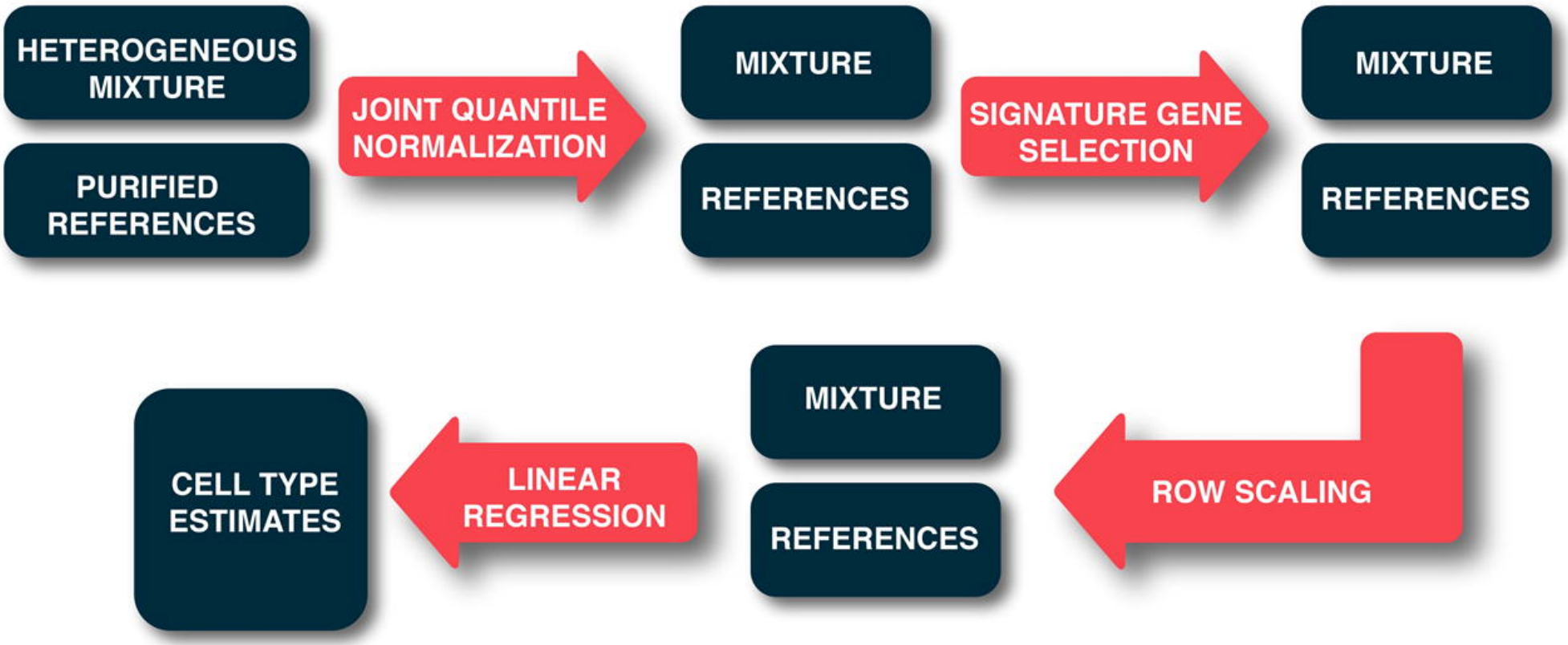


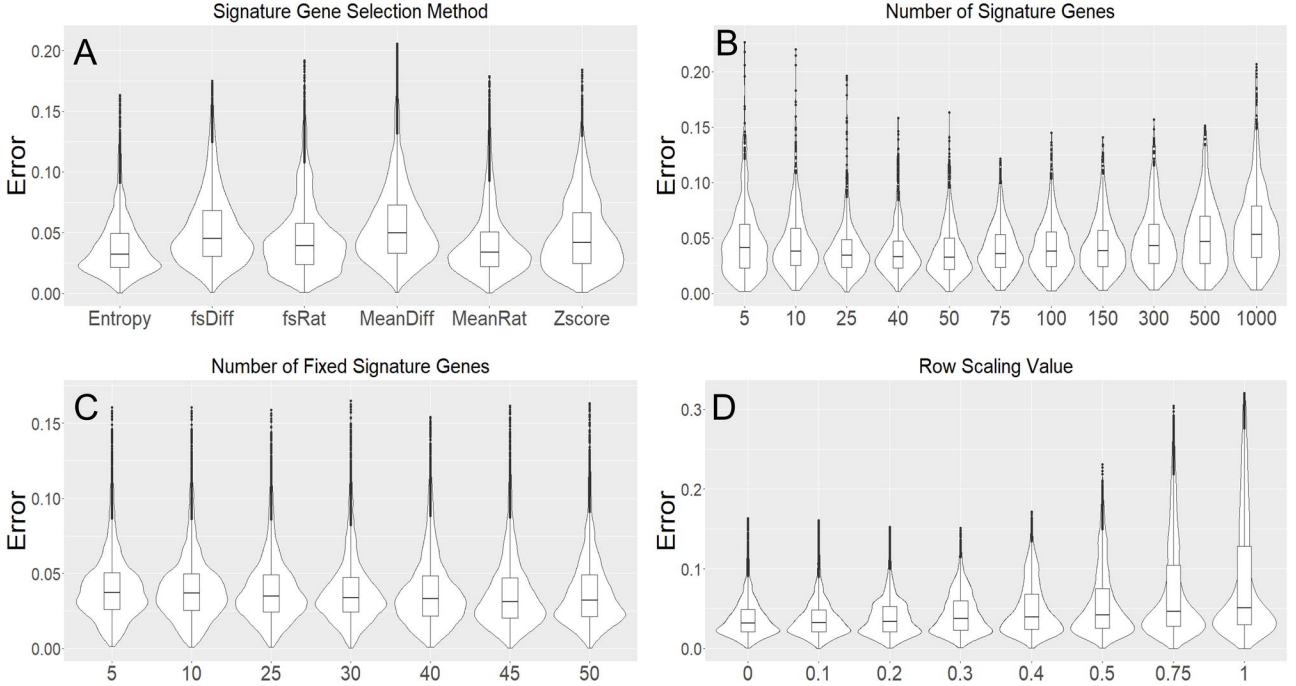


Reference Matrices Used For GTEX

	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

The GEDIT Pipeline




Signature Gene Selection Method
A
Error
0.20
0.15
0.10
0.05
0.00
Entropy
fsDiff
fsRat
MeanDiff
MeanRat
Zscore
Number of Signature Genes
B
Error
0.20
0.15
0.10
0.05
0.00
5
10
25
40
50
75
100
150
300
500
1000
Number of Fixed Signature Genes
C
Error
0.15
0.10
0.05
0.00
5
10
25
30
40
45
50
Row Scaling Value
D
Error
0.3
0.2
0.1
0.0
0
0.1
0.2
0.3
0.4
0.5
0.75
1

Average Errors by Tool, Reference Matrix and Cell Type