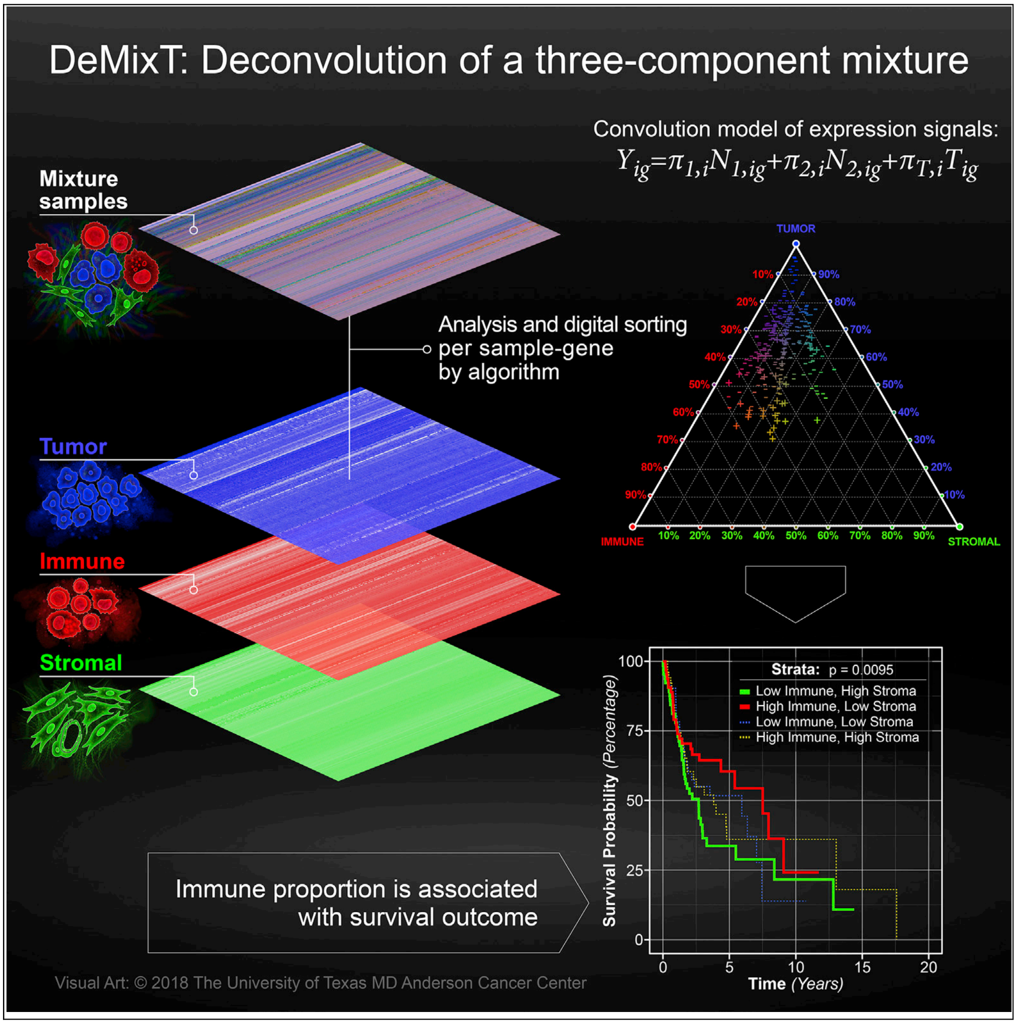
iScience

CellPress

Article

Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration


DeMixT: Deconvolution of a three-component mixture
Convolution model of expression signals:
$Y_{ig}=\pi_{1,i}N_{1,ig}+\pi_{2,i}N_{2,ig}+\pi_{T,i}T_{ig}$
Mixture samples
Analysis and digital sorting per sample-gene by algorithm
Tumor
Immune
Stromal
TUMOR
IMMUNE
STROMAL
10% 20% 30% 40% 50% 60% 70% 80% 90%
90% 80% 70% 60% 50% 40% 30% 20% 10%
10% 20% 30% 40% 50% 60% 70% 80% 90%
Strata: p = 0.0095
Low Immune, High Stroma
High Immune, Low Stroma
Low Immune, Low Stroma
High Immune, High Stroma
Survival Probability (Percentage)
100
75
50
25
0
0
5
10
15
20
Time (Years)
Immune proportion is associated with survival outcome
Visual Art: © 2018 The University of Texas MD Anderson Cancer Center

Zeya Wang, Shaolong Cao, Jeffrey S. Morris, ..., Giovanni Parmigiani, Chris C. Holmes, Wenyi Wang

wwang7@mdanderson.org

HIGHLIGHTS

A new tool *DeMixT* for efficient and accurate transcriptome deconvolution

Individual-level gene expression deconvolution of three components in cancer samples

Accurate estimation of both component-specific proportions and expression profiles

New insight in head and neck cancer prognosis and immune infiltration

Wang et al., iScience 9, 451–460
November 30, 2018
https://doi.org/10.1016/j.isci.2018.10.028

iScience

CellPress

Article

Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration

Zeya Wang,[1,2] Shaolong Cao,[1] Jeffrey S. Morris,[3] Jaeil Ahn,[4] Rongjie Liu,[3] Svitlana Tyekucheva,[5,11] Fan Gao,[1,2] Bo Li,[5,6] Wei Lu,[7] Ximing Tang,[7] Ignacio I. Wistuba,[7] Michaela Bowden,[8] Lorelei Mucci,[9] Massimo Loda,[8,10] Giovanni Parmigiani,[5,11] Chris C. Holmes,[12] and Wenyi Wang[1,13,*]

SUMMARY

Transcriptome deconvolution in cancer and other heterogeneous tissues remains challenging. Available methods lack the ability to estimate both component-specific proportions and expression profiles for individual samples. We present *DeMixT*, a new tool to deconvolve high-dimensional data from mixtures of more than two components. *DeMixT* implements an iterated conditional mode algorithm and a novel gene-set-based component merging approach to improve accuracy. In a series of experimental validation studies and application to TCGA data, *DeMixT* showed high accuracy. Improved deconvolution is an important step toward linking tumor transcriptomic data with clinical outcomes. An R package, scripts, and data are available: https://github.com/wwylab/DeMixTallmaterials.

INTRODUCTION

Heterogeneity of malignant tumor cells adds confounding complexity to cancer treatment. The evaluation of individual components of tumor samples is complicated by the tumor-stroma-immune interaction. Anatomical studies of the tumor-immune cell contexture have demonstrated that it primarily consists of a tumor core, lymphocytes, and the tumor microenvironment (Pages et al., 2009; Fridman et al., 2012). Further research supports the association of infiltrating immune cells with clinical outcomes for individuals with ovarian cancer, colorectal cancer, and follicular lymphoma (Dave et al., 2004; Galon et al., 2006; Zhang et al., 2003). The use of experimental approaches such as laser-capture microdissection (LCM) and cell sorting is limited by the associated expense and time. Therefore, understanding the heterogeneity of tumor tissue motivates a computational approach to integrate the estimation of type-specific expression profiles for tumor cells, immune cells, and the tumor microenvironment. Most commonly available deconvolution methods assume that malignant tumor tissue consists of two distinct components, epithelium-derived tumor cells and surrounding stromal cells (Ahn et al., 2013; Gong and Szustakowski, 2013). Other deconvolution methods for more than two compartments require knowledge of cell-type-specific gene lists (Liebner et al., 2014), i.e., reference genes, with some of these methods applied to estimate subtype proportions within immune cells (Li, et al., 2016a, 2016b; Newman et al., 2015). Therefore, there is still a need for methods that can jointly estimate the proportions and compartment-specific gene expression for more than two compartments in each tumor sample.

The existing method, ISOpure (Quon et al., 2013), may address this important problem. However, ISOpure assumes a linear mixture of raw expression data and represents noncancerous profiles in the mixed tissue samples by a convex combination of all the available profiles from reference samples. A drawback of this modeling approach is that the variance for noncancerous profiles is not compartment specific; therefore (1) the variances that are needed for estimating sample- and compartment-specific expressions cannot be estimated and (2) not accounting for sample variances can result in large bias in the estimated mixing proportions and mean expressions. As we aim to address the need for both gene-specific variance parameters and two unknown mixing proportions per sample in the three-component scenario, our previous heuristic search algorithm developed for two components (Ahn et al., 2013) is inadequate for the computation.

[1] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[2] Department of Statistics, Rice University, Houston, TX 77005, USA

[3] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[4] Department of Biostatistics and Bioinformatics, Georgetown University, Washington, DC 20057, USA

[5] Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA 02215, USA

[6] Department of Statistics, Harvard University, Cambridge, MA 02138, USA

[7] Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[8] Department of Oncologic Pathology, Dana Farber Cancer Institute, Boston, MA 02215, USA

[9] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[10] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[11] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[12] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

[13] Lead Contact

*Correspondence: wwang7@mdanderson.org

https://doi.org/10.1016/j.isci.2018.10.028

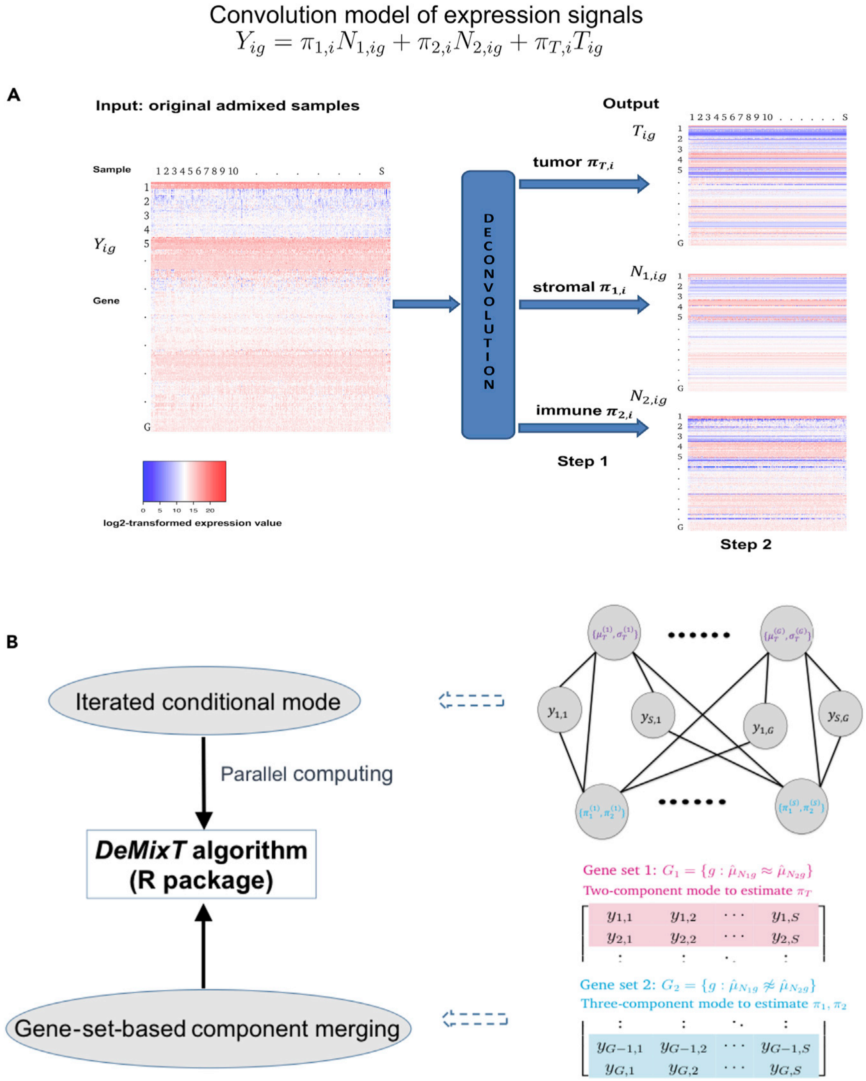

Convolution model of expression signals
$Y_{ig} = \pi_{1,i}N_{1,ig} + \pi_{2,i}N_{2,ig} + \pi_{T,i}T_{ig}$
A
Input: original admixed samples
Sample
1 2 3 4 5 6 7 8 9 10 . . . . S
Y_{ig}
Gene
1 2 3 4 5 . . . . . G
log2-transformed expression value
0 5 10 15 20
DECONVOLUTION
Output
T_{ig}
tumor $\pi_{T,i}$
$N_{1,ig}$
stromal $\pi_{1,i}$
$N_{2,ig}$
immune $\pi_{2,i}$
Step 1
Step 2
B
Iterated conditional mode
Parallel computing
DeMixT algorithm
(R package)
Gene-set-based component merging
$(\mu_T^{(1)}, \sigma_T^{(1)})$
$(\mu_T^{(G)}, \sigma_T^{(G)})$
$y_{1,1}$
$y_{S,1}$
$y_{1,G}$
$y_{S,G}$
$(\pi_1^{(1)}, \pi_2^{(1)})$
$(\pi_1^{(S)}, \pi_2^{(S)})$
Gene set 1: $G_1 = \{g : \hat{\mu}_{N_{1g}} \approx \hat{\mu}_{N_{2g}}\}$
Two-component mode to estimate π_T
$y_{1,1}$ $y_{1,2}$ $\cdots$ $y_{1,S}$
$y_{2,1}$ $y_{2,2}$ $\cdots$ $y_{2,S}$
Gene set 2: $G_2 = \{g : \hat{\mu}_{N_{1g}} \not\approx \hat{\mu}_{N_{2g}}\}$
Three-component mode to estimate π_1, π_2
$y_{G-1,1}$ $y_{G-1,2}$ $\cdots$ $y_{G-1,S}$
$y_{G,1}$ $y_{G,2}$ $\cdots$ $y_{G,S}$

Figure 1. The Model and Algorithm of *DeMixT*

(A) *DeMixT* performs three-component deconvolution to output tissue-specific proportions and isolated expression matrices of tumor (T-component), stromal (N_1-component), and immune cells (N_2-component). Heatmaps of expression levels correspond to the original admixed samples, the deconvolved tumor component, stromal component, and immune component.

(B) *DeMixT*-based parameter estimation is achieved by using the iterated conditional modes (ICM) algorithm and a gene-set-based component merging (GSCM) approach. The top graph describes the conditional dependence between the unknown parameters, which can be assigned to two groups: genome-wise parameters (top row, red superscript) and sample-wise parameters (bottom row, blue superscript). They are connected by edges, which suggest conditional dependence. The unconnected nodes on the top row are independent of each other when conditional on those on the bottom row, and vice versa. Because of conditional independence, we implemented parallel computing to substantially increase computational efficiency. The bottom graph illustrates the GSCM approach, which first runs a two-component deconvolution on gene set G_1 (red), where $\hat{\mu}_{N_1g} \approx \hat{\mu}_{N_2g}$ to estimate π_T, and then runs a three-component deconvolution on gene set G_2 (blue), where $\hat{\mu}_{N_{1g}} \not\approx \hat{\mu}_{N_{2g}}$ and π_T is given by the prior step, to estimate π_1 and π_2.

We have developed a new computational tool, *DeMixT*, to accurately and efficiently estimate the desired high-dimensional parameters in a linear additive model that accounts for variance in the gene expression levels in each compartment (Figure 1A). The corresponding R package and data for *DeMixT* is freely available for downloading at https://github.com/wwylab/DeMixTallmaterials.

RESULTS

The *DeMixT* Model and Algorithm

Here, we summarize our convolution model as follows (Figure 1A; see further details in Transparent Methods). The observed signal Y_{ig} is written as $Y_{ig} = \pi_{1,i}N_{1,ig}+\pi_{2,i}N_{2,ig}+(1-\pi_{1,i}-\pi_{2,i})T_{ig}$ for each gene g and each sample i, where Y_{ig} is the expression for the observed mixed tumor samples and $N_{1,ig}$, $N_{2,ig}$, and T_{ig} represent unobserved raw expression values from the constituents. We assume that $N_{1,ig}$, $N_{2,ig}$, and T_{ig} each follow a $\log_2$-normal distribution with compartment-specific means and variances (Ahn et al., 2013; Löonnstedt and Speed, 2002). The N_1-component and the N_2-component are the first two components, the distributions of which need to be estimated from available reference samples, and $\pi_{1,i}$ and $\pi_{2,i}$ are the corresponding proportions for sample i. The last component is the T-component, the distribution of which is unknown. In practice, the T-component can be any of the following three cell types: tumor, stromal, or immune cells. For inference, we calculate the full likelihood and search for parameter values that maximize the likelihood. Our previously developed heuristic search algorithm (Ahn et al., 2013) for a two-component model is inadequate for a three-component model, which is exponentially more complex: (1) there are two degrees of freedom in the mixing proportions, which is unidentifiable in a large set of genes that are not differentially expressed between any two components, and (2) in each iteration in the parameter search, we need to perform tedious numerical double integrations to calculate the full likelihood. The *DeMixT* algorithm introduce two new elements that help ensure estimation accuracy and efficiency (Figure 1B). We first apply an optimization approach, iterated conditional modes (ICM) (Besag, 1986), which cyclically maximizes the probability of each set of variables conditional on the rest, for which we have observed rapid convergence (Besag, 1986) to a local maximum (see the pseudo-code in Figure S1). The ICM framework further enables parallel computing, which helps compensate for the expensive computing time used in the repeated numerical double integrations. However, this is not sufficient for accurate parameter estimation. We observed that including genes that are not differentially expressed between the N_1 and N_2 components in the three-component deconvolution can introduce large biases in the estimated π_1 and π_2 (Figure S2), whereas the π_T estimation is little affected. We therefore introduce a novel gene-set-based component merging (GSCM) approach (Figure 1B). Here, we first select gene set 1, where $\mu_{N1g}\approx\mu_{N2g}$, and run the two-component model to estimate $\pi_{T,i}$. Then we select gene set 2, where $\mu_{N1g}\not\approx\mu_{N2g}$, and run the three-component model with fixed π_T from the above-mentioned equation, to estimate $\{\pi_{1,i},\pi_{2,i}\}$. Our goal is to avoid searching in the relatively flat regions of the full likelihood (model unidentifiable, Figure S3) and focus on regions where the likelihood tends to be convex. Using this approach, we not only improve the estimation accuracy but also further reduce the computing time, as only a small part of the entire parameter space needs to be searched.

Validation Using Data with Known Truth

We validated *DeMixT* in two datasets with known truth in proportions and mean expressions: a publicly available microarray dataset (Shen-Orr et al., 2010) generated using mixed RNAs from rat brain, liver, and lung tissues in varying proportions and an RNA sequencing (RNA-seq) dataset generated using mixed RNAs from three cell lines, lung adenocarcinoma (H1092), cancer-associated fibroblasts (CAFs), and tumor infiltrating lymphocytes (TILs).

We used GEO: GSE19830 (Shen-Orr et al., 2010) as our first dataset for benchmarking. This microarray experiment was designed for the expression profiling of samples from *Rattus norvegicus* with the Affymetrix Rat Genome 230 2.0 Array, including 30 mixed samples of liver, brain, and lung tissues in 10 different mixing proportions with three replicates (Table S1). To run *DeMixT*, we used the samples with 100% purity to generate the respective reference profiles for the N_1-component, N_2-component, and T-component. We ran the deconvolution for the 30 mixed samples under three scenarios, respectively, assuming the liver, brain, and lung tissues to be the unknown T-component tissue. To generate the second dataset in RNA-seq, we performed a mixing experiment, in which we mixed mRNAs from three cell lines, lung adenocarcinoma in humans (H1092), CAFs, and TILs, at different proportions to generate 32 samples, including 9 samples that correspond to three repeats of a pure cell-line sample for the three cell lines (Table S2). The RNA amount of each tissue in the mixture samples was calculated on the basis of real RNA concentrations tested in the biologist's laboratory. We assessed our deconvolution approach through a number of statistics, e.g., concordance correlation coefficients (CCCs) (Lawrence and Lin, 1989), root mean square errors, and a summary statistic for measuring the reproducibility of the estimated π across scenarios when a different component is unknown (see Transparent Methods). We showed that *DeMixT* performed

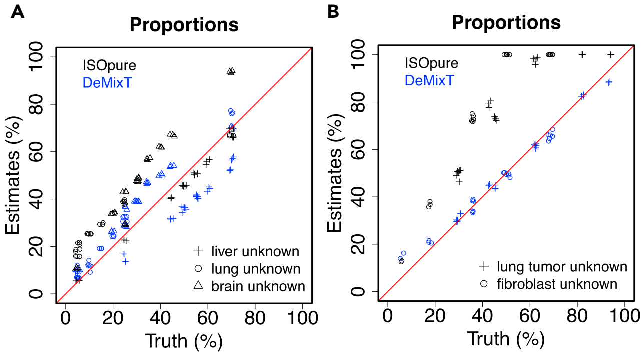


Figure 2. Validation Results using Microarray and RNA-seq Data from Tissue and Cell-Line Mixture Experiments

(A) Scatterplot of estimated tissue proportions versus the truth when liver (plus), brain (triangle), or lung (circle) tissue is assumed to be the unknown tissue in the microarray experiments mixing the three; estimates from ISOpure are also presented.

(B) Scatterplot of estimated tissue proportions versus the truth when either lung tumor (plus) or fibroblast (circle) cell lines are assumed to be the unknown tissue in the RNA-seq experiments mixing lung tumor, fibroblast, and lymphocyte cell lines.

See also Figures S4 and S6 and Tables S3–S7.

well and outperformed ISOpure in terms of accuracy and reproducibility (Figures 2A and 2B; see Transparent Methods for further details, Figures S4–S7, Tables S3–S7).

Validation Using LCM Data

We then applied *DeMixT* to a "gold standard" validation dataset from real tumor tissue that has known proportions, mean expressions, and individual component-specific expressions. This dataset (GEO: GSE97284) was generated at Dana Farber Cancer Institute through Laser capture microdissection (LCM) experiments on tumor samples from patients with prostate cancer. It consists of 25 samples of isolated tumor tissues, 25 samples of isolated stromal tissues, and 23 admixture samples (Tyekucheva et al., 2017b). LCM was performed on formalin-fixed paraffin embedded (FFPE) tissue samples from 23 patients with prostate cancer, and microarray gene expression data were generated using the derived and the matching dissected stromal and tumor tissues (GEO: GSE97284 [Tyekucheva et al., 2017a]). Owing to the low quality of the FFPE samples, we selected a subset of probes (see Transparent Methods) and ran *DeMixT* under a two-component mode. *DeMixT* obtained concordant estimates of the tumor proportions when the proportion of the stromal component was unknown and when the proportion of tumor tissue was unknown (CCC = 0.87) (Figure 3A). *DeMixT* also tended to provide accurate component-specific mean expression levels (Figures 3B, 3C, and S8) and yielded standard deviation estimates that are close to those from the dissected tumor samples (Figure S9). As a result, the *DeMixT* individually deconvolved expressions achieved high CCCs (mean = 0.96) for the tumor component (Figures 3D and S10). The expressions for the stromal component were more variable than those for a common gene expression dataset, hence both *DeMixT* and ISOpure gave slightly biased estimates of the means and standard deviations.

Application to the Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma Data

A recent study of head and neck squamous cell carcinoma (HNSCC) showed that the infiltration of immune cells, both lymphocytes and myelocytes, is positively associated with viral infection in virus-associated tumors (Li, et al., 2016; 2016b). We downloaded HNSCC RNA-seq data from The Cancer Genome Atlas (TCGA) data portal (Cancer Genome Atlas Network, 2015) and ran *DeMixT* for deconvolution. We normalized the expression data with the total count method and filtered out genes with zero count in any sample. There was a total of 44 normal tissue and 269 tumor samples in the HNSCC dataset. We collected the information of human papillomavirus (HPV) infection status for the HNSCC samples. Samples were classified as HPV-positive (HPV+) using an empiric definition of the detection of >1,000 mapped RNA-seq reads, primarily aligning to viral genes E6 and E7, which resulted in 36 HPV+ samples (Cancer Genome Atlas Network, 2015). Since only reference samples for the stromal component are available from TCGA (i.e., 44 normal samples and 269 tumor samples), we devised an analytic pipeline for *DeMixT* to run successfully on the HNSCC samples (for details, see Transparent Methods and Figure S11). In brief, we first used data from the HPV+ tumors to derive reference samples for the immune component and then ran

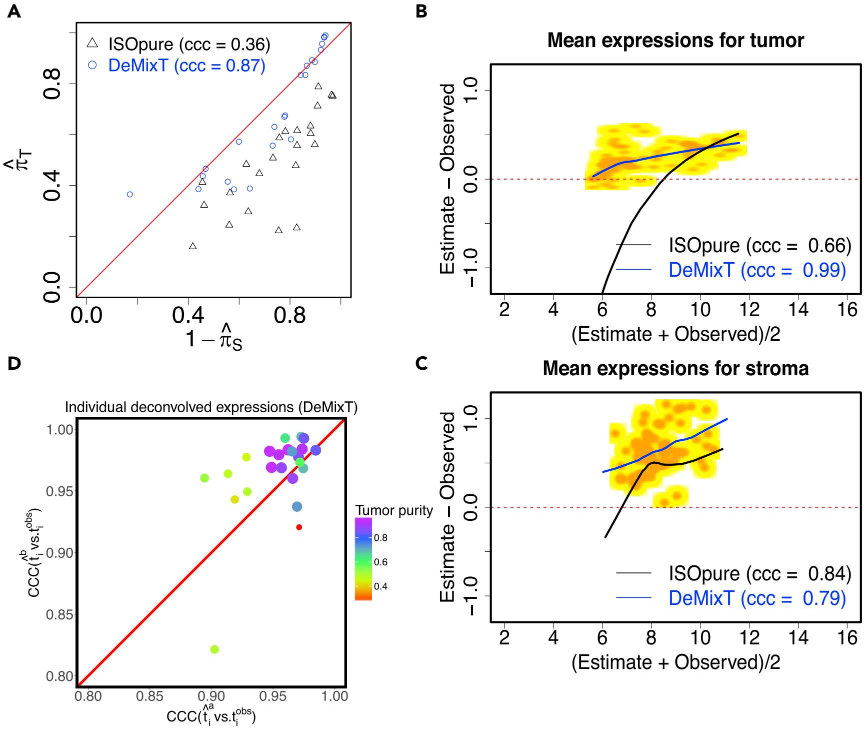


Figure 3. Analyses of Real Data Using *DeMixT* through Validation Using LCM Data in Prostate Cancer

(A) Scatterplot of estimated tumor proportions versus 1- estimated stromal proportions; estimates from *DeMixT* (blue) are compared with those from ISOpure (black).

(B) Smoothed scatter MA plots between observed and deconvolved mean expression values at the log2 scale from *DeMixT* for the tumor component (yellow for low values and orange for high values). The lowest smoothed curves for *DeMixT* are shown in blue and those for ISOpure in black. (C) Smoothed scatter MA plots between observed and deconvolved mean expression values at the Log2 scale from *DeMixT* for the stromal component.

(D) Scatterplot of concordance correlation coefficient (CCC) between individual deconvolved expression profiles for the tumor component ($\hat{t}_i$) and observed values (t_i^{obs}) for 23 LCM matching prostate cancer samples. Superscript *a*: stromal component is represented by reference samples; *b*: tumor component is represented by reference samples. Color gradient and size of each point corresponds to the estimated tumor proportion.

the three-component *DeMixT* on the entire dataset to estimate the proportions for both HPV-negative (HPV-) and HPV+ samples. For all tumor samples, we obtained the immune (mean = 0.22, standard deviation = 0.10), the tumor (mean = 0.64, standard deviation = 0.13), and the stromal proportions (mean = 0.14, standard deviation = 0.07; see Figure 4A). The distribution of stromal proportions seems independent, whereas the tumor and immune proportions are inversely correlated. As expected, HPV+ tumor samples had significantly higher immune proportions than those that tested as HPV- (Li, et al., 2016, 2016b; Fakhry et al., 2008) (p value = 2 × 10^{-8}; Figures 4A, 4B, and S12). To further evaluate the performance of our deconvolved expression levels, we performed differential expression tests for immune versus stromal tissue and immune versus tumor tissue, respectively, on 63 infiltrating immune cell-related genes (CD and HLA genes). For example, Figure 4C illustrates that the deconvolved expressions were much higher in the immune component than in the other two components for three important immune marker genes, CD4, CD14, and HLA-DOB. What we observed with the purified expression levels of these genes is as expected. Overall, 51 of 63 genes were significantly more highly expressed in the immune component than in the other two components (adjusted p values are listed in Data S1; also see Figure 4D). In addition, we divided the patient samples into four groups based on their estimated immune and stromal proportions, using simply the median values as cutoffs. The corresponding four groups of patient samples are significantly different in terms of overall survival outcomes. The Cox proportional hazards regression coefficient of the high-immune-low-stroma group versus the low-immune-high-stroma group is −0.66 with the Wald test (p value = 0.001). As expected, the high-immune-low-stroma group of patients have the best prognosis as compared with the other groups. In comparison, we performed the same survival analysis on patients who are categorized by dichotomizing the immune and stromal scores of ESTIMATE (Yoshihara et al.,

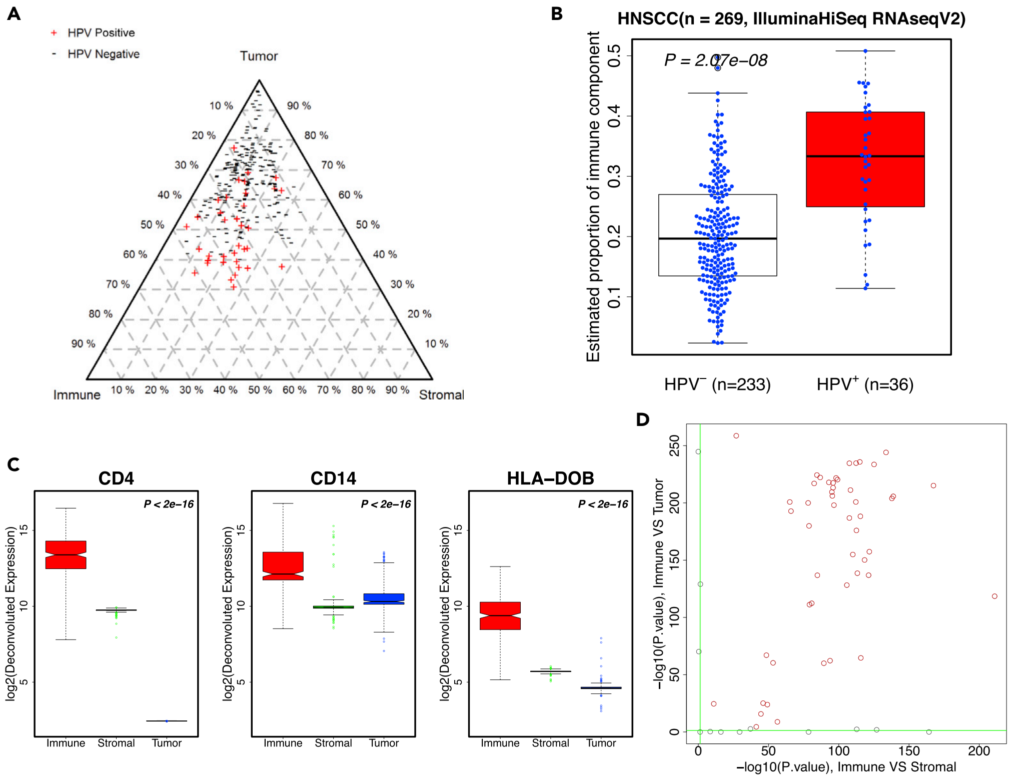


Figure 4. Analyses of Real Data Using *DeMixT* through Application to TCGA RNA-seq Data in HNSCC

(A) A triangle plot of estimated proportions (%) of the tumor component (top), the immune component (bottom left), and the stromal component (bottom right) in the HNSCC data. Points closer to a component's vertex suggests higher proportion for the corresponding component, whose quantity equals the distance between the side opposite the vertex and a parallel line (illustrated as dashed gray lines for the multiples of 10th percentile) that a point is sitting on. The "+" and "−" signs correspond to the infectious status of HPVs.

(B) Boxplots of estimated immune proportions for HNSCC samples in the test set display differences between HPV+ (red) and HPV− (white) samples.

(C) Boxplots of log2-transformed deconvolved expression profiles for three important immune genes (CD4, CD14, HLA-DOB) in the test set of HNSCC samples. Red: immune component; green: stromal component; blue: tumor component. P values of differential tests are at the top right corner for each gene: the first p value is for immune versus stromal component; second p value is for immune versus tumor component.

(D) Scatterplot of negative log-transformed p values for comparing deconvolved expression profiles between immune component and the other two components of 63 immune cell-related genes. The x axis: immune component versus stromal component; y axis: immune component versus tumor component. Genes in red are significant in both comparisons. Green horizontal and vertical lines: cutoff value for statistical significance.

2013), also in four groups. Although the ESTIMATE-defined high-immune-low-stroma group remains on top of all four survival curves, we did not observe a statistically significant difference between these groups. Therefore, *DeMixT*-based immune and stroma proportions is more useful in categorizing patients with different prognosis outcomes (Figure S13).

DISCUSSION

We present a novel statistical method and software, *DeMixT* (R package at https://github.com/wwylab/DeMixTallmaterials), for dissecting a mixture of tumor, stromal, and immune cells based on the gene expression levels and providing an accurate solution. Our method allows us to simultaneously estimate both cell-type-specific proportions and reconstitute patient-specific gene expression levels with little prior information. Distinct from the input data of most other deconvolution methods such as CIBERSORT and ESTIMATE, our input data consist of gene expression levels from (1) observed mixtures of tumor samples and (2) a set of reference samples from p-1 compartments (where p is the total number of compartments). Our different model assumptions and goal for individual-level deconvolved expression levels have brought unique analytical challenges that are not relevant for deconvolution methods aforementioned, which use input from all p compartments and are regression based. Our output data provide the mixing proportions, the means and variances of expression levels for genes in each compartment, as well as the expression levels for each gene in each compartment and each sample. The full gene-compartment sample-specific output allows for the application of all pipelines previously developed for downstream analyses, such as clustering and feature selection methods in cancer biomarker studies, which are still applicable to the deconvolved gene expressions. We achieved this output by modeling compartment-specific variance and addressing the associated inferential challenges. Our model assumes a linear mixture of data before a log2-transformation (Ahn et al., 2013; Löönnstedt and Speed, 2002), thereby introducing nonlinear associations into the log-space of the data. Beyond extending the *DeMix* model (Ahn et al., 2013) from two-component to three-component deconvolution, *DeMixT* also proposes new features as summarized below, resulting in an overall better performance (Figure S14). *DeMixT* addresses transcriptome deconvolution in two steps. In the first step, rather than using a heuristic search as before, we now estimate the mixing proportions and the gene-specific distribution parameters for each compartment using an ICM method (Besag, 1986), which can quickly converge and is guaranteed to find a local maximum. We have further proposed a novel GSCM approach and integrated it with ICM for three-component deconvolution, to substantially improve model identifiability and computational efficiency. In the second step, we reconstitute the expression profiles for each sample and each gene in each compartment based on the parameter estimates from the first step. The success of the second step relies largely on the success of the first. We have overcome the otherwise significant computational burden for searching the high-dimensional parameter space and numerical double integration, owing to our explicit modeling of variance through parallel computing and gene-set-based component merging. On a PC with a 3.07-GHz Intel Xeon processor with 20 parallel threads, *DeMixT* takes 14 min to complete the full three-component deconvolution task of a dataset consisting of 50 samples and 500 genes (see Table S8). Our new design makes it possible to first select a subset of genes for accurate and efficient proportion estimation and then estimate gene expression for any gene set or for the whole transcriptome. This overcomes the deficiency of most existing deconvolution tools that enforce using the same set of genes in the estimations of both proportions and gene expression levels. Our method can be applied to other data types that are generated from mixed materials.

We have used a series of experimental datasets to validate the performance of *DeMixT*. These datasets were generated from a mixture of normal tissues, a mixture of human cell lines, and LCM of FFPE tumor samples. *DeMixT* succeeded in recapitulating the truth in all datasets. When compared with ISOpure, *DeMixT* gave more accurate estimations of proportions in all datasets. *DeMixT* more explicitly accounts for sample variances, an assumption that adheres more closely to the real biological samples. Even for the *in vitro* dataset of admixed rat tissues, which generated only technical replicates that had very small variances so that assuming no variance becomes reasonable, we showed that the estimation of gene expression by *DeMixT* is still comparable with the estimation by ISOpure. On the dataset of mixed human cell lines, *DeMixT* performed as well as CIBERSORT (in estimating the tumor and the fibroblast components), a popular method for estimating only the proportions of cell types in complex tissues (Figure S6), even though *DeMixT* used reference profiles from one less component than CIBERSORT. We further demonstrated tumor-stroma-immune deconvolution by *DeMixT* using TCGA HNSCC data. We were able to correlate our immune proportion estimates with the available HPV infection status in HNSCC, as is consistent with previous observations that a high level of immune infiltration appears with viral infection in cancer (Li, et al., 2016; 2016b). For this dataset, *DeMixT* is the first to provide a triangular view of tumor-stroma-immune proportions (Figure 4A), the interesting dynamics of which may shed new light on predicting the prognosis of HNSCC.

Here, we discuss four major factors that would potentially impact the performance of deconvolution, regardless of the model and method used. (1) The number and diversity of tumor samples and reference profiles. Some cancer types, such as breast cancer, are more heterogeneous within the tumor component than others. Some cancer types show more genomic rearrangements and copy number changes, which impact transcriptomic activities, whereas others, such as prostate cancer, are less often so. There exist large variations in the availability of the number and type of reference profiles across cancer types. We recently applied *DeMixT* to the datasets from the TCGA PanCanAtlas project across 16 cancer types. Among them, we used RNA-seq data generated from the corresponding normal tissues for 15 cancer types, with the sample size for normal samples ranging from 10 to 98. With the remaining cancer types in TCGA, there are <10 normal samples available, for which we have not run *DeMixT*, except for one cancer type (pancreatic cancer, PAAD). In PAAD, we used tumor samples that had been determined to have very low tumor content as the reference profiles (n = 7). In both scenarios of normal controls, we obtained reasonable results, based on which we performed clustering analysis, pathway analysis, and variable selection for gene sets associated with survival outcomes. Our analyses suggested that the estimated mixing proportions and individual expression levels are useful to identify biological signals that were previously diluted in the mixed measures (unpublished results). Generally, our model assumptions will be mildly violated in most studies (e.g., in the TCGA datasets) and strongly violated in some studies. Assuming there is a reasonable level of homogeneity within a component, increasing the sample size will increase the reliability of parameter estimations (i.e., $\widehat{\mu}_{N_{1g}}, \widehat{\mu}_{N_{2g}}; \widehat{\sigma}_{N_{1g}}, \widehat{\sigma}_{N_{2g}}$). (2) The platforms used to profile gene expression. We observed good performances of *DeMixT* on data generated from real tumor samples using both Affymetrix microarray and Illumina RNA sequencing platforms. Testing *DeMixT* on other platforms should involve a first step of checking whether the linearity combination of the log-normal distributions still holds. (3) The tissues from which the various input profiles were derived. We found that expression measurements from FFPE samples are much noisier than those from fresh-frozen samples, and in the analysis of the LCM data, had to devise a more stringent filtering criteria on the set of genes to be used for deconvolution. (4) The genes selected for the sequential steps of the *DeMixT* algorithm. In a two-component setting, we observed that both variances and mean differences in the expression levels between the two components for each gene can affect how accurately the mixing proportions are estimated, whereas not all genes are needed for the proportion estimation. We therefore proposed to select genes that have moderate variances and large differences between the two components to estimate proportions. In a three-component setting, using the GSCM approach to reduce to a pseudo-two-component problem allowed us to apply a similar strategy. The GSCM approach is general in sequentially merging components through gene selections and can be extended to deconvolution problems with more than three components but will incur high computational cost. Currently, our gene selection and GSCM strategy follow the principle of focusing on a subspace of the high-dimensional parameters for model identifiability but are heuristic and may need adaptation across datasets. We observed the performance of GSCM is robust to the number of genes selected within the range of hundreds. Future work includes systematically evaluating the impact of each set of high-dimensional parameters on the full likelihood underlying our convolution model and search for a unified gene selection method for the deconvolution of datasets that range over a wide spectrum of biological phenomena. Future work also includes development of a numerical measure of confidence to filter out potentially unreliable expression estimates.

Reference gene-based deconvolution is popular for estimating immune subtypes within immune cells (Liebner et al., 2014; Newman et al., 2015). Our method does not require reference genes, which we consider as difficult to obtain for the tumor component; however, *DeMixT* can take reference genes when available. With the reference sample approach, we assume that the first *p*-1 compartments in the observed mixture are similar to those in reference samples, whereas the remaining compartment is unknown and so may end up capturing most of the heterogeneity. The reference samples can be derived from historical patient data or from the corresponding healthy tissues, such as data from GTEx (Lonsdale et al., 2013) (e.g., RNA-seq data from sun-exposed skin as reference samples for melanoma, unpublished results). Furthermore, each of the three components may contain more than one type of cell, in particular, the immune component. It was reported that, although the immune cell subtypes are heterogeneous, their relative proportions within the immune component are consistent across patient samples (Gentles et al., 2015), which supports our approach that models the pooled immune cell population using one distribution. Estimating low proportions is more prone to biases in methods without reference genes than those with reference genes, as observed in our cell-line mixed RNA-seq dataset in which the immune cell component is consistently low. However, it occurred only in this artificially mixed dataset, whereas in real data, such as the HNSCC dataset, there are samples presenting a high level of immune infiltration, thus improving the accuracy for all parameter estimations, including those in samples presenting a low level of immune infiltration. In future work, we will consider expanding to a hierarchical model for immune subpopulations that will include dynamic immune components. For optimized performance of *DeMixT*, the data analysis should be linked with cancer-specific biological knowledge.

Limitations of the Study

Here we are focused on resolving statistical challenges in a new concept of jointly estimating component-specific proportions and distributions of gene expression, as well as individual gene expression levels in a mixture of three components. Our approach has been comprehensively benchmarked using multiple datasets. However, *DeMixT* needs further studies to improve its utility in real cancer data, including (1) a unified gene selection method that automatically detects, in a high-dimensional likelihood space, the most identifiable region for parameter estimation; (2) a numerical measure of confidence to filter out potentially unreliable expression estimates; (3) extension to a hierarchical model to accommodate multiple immune cell subtypes; (4) cancer-specific data analyses to further understand and remedy for the potential impact of available normal tissues as input reference profiles.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, 15 figures, 9 tables, and 1 data file and can be found with this article online at https://doi.org/10.1016/j.isci.2018.10.028.

ACKNOWLEDGMENTS

Z.W. and W.W. are supported by the U.S. National Cancer Institute through grant numbers R01CA174206, R01 CA183793, and P30 CA016672. W.W. is supported by U.S. National Cancer Institute 2R01 CA158113. Z.W. and J.S.M. are supported by NIH grants R01 CA178744 and P30 CA016672 and NSF grant 1550088. J.S.M. is supported by U.S. NSF 1550088 and the MD Anderson Colorectal Cancer Moonshot. S.C., J.A., X.T., and I.I.W. are supported by NIH grant 1R01CA183793. X.T. and I.I.W. are supported by The University of Texas Lung Specialized Programs of Research Excellence grant P50CA70907. S.T. is supported by NIH grant R01 CA174206 and Prostate Cancer Foundation Challenge Award. G.P. is supported by NIH grants 5R01 CA174206-05 and 4P30CA006516-51. M.L. is supported by NIH grants RO1CA131945, R01CA187918, DoD PC130716, P50 CA90381 and the Prostate Cancer Foundation, United States.

We thank Vesteinn Porsson, Ilya Shmulevich, David Gibbs, Liuqing Yang, and Hongtu Zhu for useful discussions and valuable suggestions.

AUTHOR CONTRIBUTIONS

Z.W. developed and coded the algorithms in *DeMixT*, analyzed the data, and performed the validation studies. S.C. performed the application study and analyzed the data using *DeMixT*. J.A. proposed the assumptions of linearity and model distributions. F.G. and R.L. helped build the *DeMixT R* package. S.T., B.L., W.L., X.T., I.I.W., M.B, L.M., and M.L. contributed data/materials for the validation and application studies. Z.W. and W.W. wrote the manuscript. J.S.M., S.T., G.P., and C.C.H. contributed to the discussion of results and revision of the manuscript. W.W. supervised the whole study. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 20, 2018
Revised: July 13, 2018
Accepted: October 27, 2018
Published: November 30, 2018

REFERENCES

Gong, T., and Szustakowski, J.D. (2013). DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics *29*, 1083–1085.

Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, MB., Diao, L., Wistuba, I.I., and Wang, W. (2013). De Mix: deconvolution for mixed cancer transcriptomes using raw measured data. Bioinformatics *29*, 1865–1871.

Besag, J. (1986). On the statistical analysis of dirty pictures. J. R. Stat. Soc. Series B Stat. Methodol. *48*, 259–302.

Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature *517*, 576.

Dave, S.S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R.D., Chan, W.C., Fisher, R.I., Braziel, R.M., Rimsza, L.M., Grogan, T.M., et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N. Engl. J. Med. *351*, 2159–2169.

Fakhry, C., Westra, W.H., Li, S., Cmelak, A., Ridge, J.A., Pinto, H., Forastiere, A., and Gillison, M.L. (2008). Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. J. Natl. Cancer Inst. *100*, 261–269.

Fridman, W.H., Pages, F., Sautes-Fridman, C., and Galon, J. (2012). The immune contexture in human tumours: impact on clinical outcome. Nat. Rev. Cancer *12*, 298–306.

Galon, J., Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., and Zinzindohoué, F. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science *313*, 1960–1964.

Gentles, A.J., Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D., and Diehn, M. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat. Med. *21*, 938–945.

Lawrence, I., and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics *45*, 255–268.

Li, B., Li, T., Li, T., Pignon, J.C., Wang, B., Wang, J., Shukla, S.A., Dou, R., Chen, Q., Hodi, F.S., et al. (2016a). Landscape of tumor-infiltrating T cell repertoire of human cancers. Nat. Genet. *48*, 725–732.

Li, B., Severson, E., Pignon, J.C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J.C., Rodig, S., et al. (2016b). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. *17*, 174.

Liebner, D.A., Huang, K., and Parvin, J.D. (2014). MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinformatics *30*, 682–689.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

Löonnstedt, I., and Speed, T. (2002). Replicated microarray data. Stat. Sin. *12*, 31–46.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Pages, F., Galon, J., Dieu-Nosjean, M.C., Tartour, E., Sautès-Fridman, C., and Fridman, W.H. (2009). Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene *29*, 1093–1102.

Quon, G., Haider, S., Deshwar, A.G., Cui, A., Boutros, P.C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. *5*, 29.

Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., and Butte, A.J. (2010). Cell type-specific gene expression differences in complex tissues. Nat. Methods *7*, 287–289.

Tyekucheva S., Bowden M., Bango C., Giunchi F., Huang Y., Zhou C., Bondi A., Lis R., Van Hemelrijck M., Andrén O., et al., (2017a). Data accessible at NCBI GEO database; Accession GSE97284. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97284.

Tyekucheva, S., Bowden, M., Bango, C., Giunchi, F., Huang, Y., Zhou, C., Bondi, A., Lis, R., Van Hemelrijck, M., Andrén, O., et al. (2017b). Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. Nat. Commun. *8420*, https://doi.org/10.1038/s41467-017-00460-4.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612.

Zhang, L., Conejo-Garcia, J.R., Katsaros, D., Gimotty, P.A., Massobrio, M., Regnani, G., Makrigiannakis, A., Gray, H., Schlienger, K., Liebman, M.N., et al. (2003). Intratumoral T Cells, recurrence, and survival in epithelial ovarian cancer. N. Engl. J. Med. *348*, 203–213.

ISCI, Volume *9*

Supplemental Information

Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration

Zeya Wang, Shaolong Cao, Jeffrey S. Morris, Jaeil Ahn, Rongjie Liu, Svitlana Tyekucheva, Fan Gao, Bo Li, Wei Lu, Ximing Tang, Ignacio I. Wistuba, Michaela Bowden, Lorelei Mucci, Massimo Loda, Giovanni Parmigiani, Chris C. Holmes, and Wenyi Wang

Table S1. Summary of datasets GEO19830 with the mixture proportions (%) of rat liver, brain and lung tissues, related to Figure 2.

Mixture	Number of Technical Replicates	Tissue Type	Liver	Brain	Lung
1	3	Pure	100	0	0
2	3	Pure	0	100	0
3	3	Pure	0	0	100
4	3	Mixed	5	25	70
5	3	Mixed	70	5	25
6	3	Mixed	25	70	5
7	3	Mixed	70	25	5
8	3	Mixed	45	45	10
9	3	Mixed	55	20	25
10	3	Mixed	50	30	20
11	3	Mixed	55	30	15
12	3	Mixed	50	40	10
13	3	Mixed	60	35	5

Table S2. Summary of datasets in the mixed cell line experiment with the mixture proportions (%) of lung adenocarcinoma in humans (H1092), cancer-associated fibroblasts (CAFs) and tumor infiltrating lymphocytes (TIL), related to Figure 2.

Mixture	Tissue Type	H1092	CAF	TIL
1	Pure	100	0	0
2	Pure	100	0	0
3	Pure	100	0	0
4	Pure	0	100	0
5	Pure	0	100	0
6	Pure	0	100	0
7	Pure	0	0	100
8	Pure	0	0	100
9	Pure	0	0	100
10	Mixed	45.6	50.8	3.6
11	Mixed	45.6	50.8	3.6
12	Mixed	45.6	50.8	3.6
13	Mixed	61.9	35.6	2.5
14	Mixed	61.9	35.6	2.5
15	Mixed	61.9	35.6	2.5
16	Mixed	29.6	68	2.4
17	Mixed	29.6	68	2.4
18	Mixed	29.6	68	2.4
19	Mixed	43.2	49.7	7.1
20	Mixed	43.2	49.7	7.1
21	Mixed	43.2	49.7	7.1
22	Mixed	63	36.2	0.9
23	Mixed	63	36.2	0.9
24	Mixed	63	36.2	0.9
25	Mixed	30	69.1	0.8
26	Mixed	30	69.1	0.8
27	Mixed	30	69.1	0.8
28	Mixed	81.9	17.7	0.4
29	Mixed	81.9	17.7	0.4
30	Mixed	81.9	17.7	0.4
31	Mixed	93.6	6	0.4
32	Mixed	93.6	6	0.4

Table S3. Measures of reproducibility for estimated proportions across different scenarios in the GSE19830 dataset and the mixed cell line RNA-seq dataset, related to Figure 2.

Estimated Tissue	*DeMixT*	*ISOpure*
Brain	0.03	0.10
Lung	0.03	0.08
Liver	0.03	0.07
H1092	0.05	0.40
CAF	0.06	0.41
TIL	0.02	0.02

Table S4. Concordance correlation coefficients between estimated and true proportions in the GSE19830 dataset. The 95% confidence interval is in parentheses, related to Figure 2.

Estimated Tissue	Brain	Lung	Liver	Average
DeMixT (Brain Unknown)	0.88 (0.80, 0.93)	0.95 (0.91, 0.97)	0.74 (0.61, 0.83)	0.86
DeMixT (Lung Unknown)	0.84 (0.71, 0.91)	0.97 (0.95, 0.98)	0.75 (0.63, 0.84)	0.85
DeMixT (Liver Unknown)	0.77 (0.65, 0.86)	0.96 (0.94, 0.97)	0.74 (0.62, 0.83)	0.82
ISOpure (Brain Unknown)	0.69 (0.55, 0.79)	1 (1.00, 1.00)	0.72 (0.58, 0.81)	0.80
ISOpure (Lung Unknown)	0.97 (0.94, 0.99)	0.74 (0.61, 0.83)	0.84 (0.75, 0.90)	0.85
ISOpure (Liver Unknown)	0.93 (0.88, 0.96)	0.98 (0.96, 0.99)	0.98 (0.96, 0.99)	0.96

Table S5. Root mean squared errors (RMSEs) between estimated and true proportions in the GSE19830 dataset, related to Figure 2.

Estimated Tissue	Brain	Lung	Liver	Average
DeMixT (Brain Unknown)	0.08	0.06	0.13	0.09
DeMixT (Lung Unknown)	0.1	0.05	0.13	0.09
DeMixT (Liver Unknown)	0.12	0.05	0.13	0.10
ISOpure (Brain Unknown)	0.18	0.02	0.16	0.12
ISOpure (Lung Unknown)	0.04	0.14	0.11	0.10
ISOpure (Liver Unknown)	0.07	0.04	0.04	0.05

Table S6. Concordance correlation coefficients between estimated and true proportions in the mixed cell line RNA-seq dataset. The 95% confidence interval is given in parentheses. H1092: lung tumor adenocarcinoma; CAF: cancer-associated fibroblasts; TIL: tumor infiltrating lymphocytes, related to Figure 2.

Estimated Tissue	Lung Tumor (H1092)	Fibroblast (CAF)	Immune (TIL)	Average
DeMixT (H1092 Unknown)	0.99 (0.99, 1.00)	0.91 (0.84, 0.95)	0.14 (0.05, 0.22)	0.68
DeMixT (CAF Unknown)	0.91 (0.84, 0.95)	0.98 (0.97, 0.99	0.08 (0.02, 0.14)	0.66
ISOpure (H1092 Unknown)	0.51 (0.31, 0.66)	0.54 (0.35, 0.69)	0.26 (0.13, 0.38)	0.44
ISOpure (CAF Unknown)	0.51 (0.33, 0.65)	0.45 (0.28, 0.60)	-0.01 (-0.03, 0.01)	0.32

Table S7. Root mean squared errors between estimated proportions and true proportions in RNA-seq data from mixed cell line experiment, related to Figure 2.

Estimated Tissue	Lung Tumor (H1092)	Fibroblast (CAF)	Immune (TIL)	Average
DeMixT (H1092 Unknown)	0.02	0.08	0.09	0.06
DeMixT (CAF Unknown)	0.09	0.04	0.08	0.07
ISOpure (H1092 Unknown)	0.27	0.25	0.03	0.18
ISOpure (CAF Unknown)	0.34	0.36	0.03	0.24

H1092, lung tumor adenocarcinoma; CAF, cancer-associated fibroblasts; TIL, tumor infiltrating lymphocytes

Table S8. Computing time for *DeMixT*. *DeMixT* was run on a simulated dataset consisting of 50 samples and 500 genes using 2 or 20 threads. Of all genes, 400 belong to gene set 1 (G_1) and the remaining 100 belong to gene set 2 (G_2), as defined in our gene-set-based component merging approach, related to Figure 1b.

	w/o CM	w/CM		
	Total	Two-component step: G1	Three-component: G2	Total
2 threads	16.1 h	37 min	48 min	85 min
20 threads	2.5h	6 min	8 min	14 min

Table S9. Number of probes/genes with different relationships between different component tissues, related to Figure 1.

GEO19830, mixed tissue microarray data:

Unknown Tissue	Number of Probes	Percentage of Probes
$\hat{\mu}_{liver} \approx \hat{\mu}_{brain} \approx \hat{\mu}_{lung}$	10928/31099	35.1%
$\hat{\mu}_{liver} \not\approx \hat{\mu}_{brain} \approx \hat{\mu}_{lung}$	4321/31099	13.9%
$\hat{\mu}_{liver} \approx \hat{\mu}_{brain} \not\approx \hat{\mu}_{lung}$	2978/31099	9.6%
$\hat{\mu}_{liver} \not\approx \hat{\mu}_{brain} \not\approx \hat{\mu}_{lung}$	4671/31099	15.0%

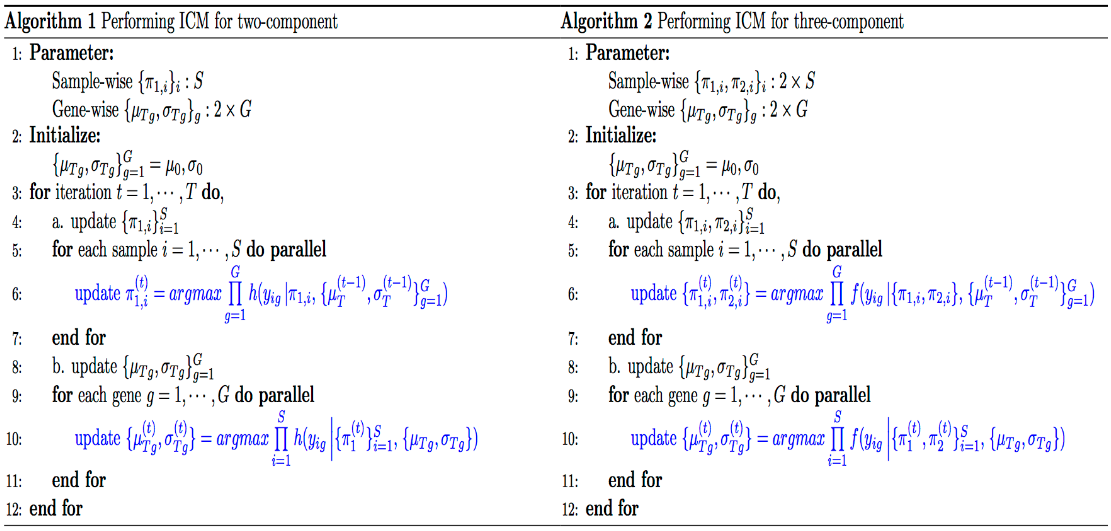
Mixed cell line RNA-seq data:

Unknown Tissue	Number of Genes	Percentage of Genes
$\hat{\mu}_{H1092} \approx \hat{\mu}_{CAF} \approx \hat{\mu}_{TIL}$	490/5715	8.6%
$\hat{\mu}_{H1092} \not\approx \hat{\mu}_{CAF} \approx \hat{\mu}_{TIL}$	752/5715	13.2%
$\hat{\mu}_{H1092} \approx \hat{\mu}_{CAF} \not\approx \hat{\mu}_{TIL}$	958/5715	16.8%
$\hat{\mu}_{H1092} \not\approx \hat{\mu}_{CAF} \not\approx \hat{\mu}_{TIL}$	2373/5715	41.5%

Microarray data from laser capture microdissected FFPE prostate cancer patient samples:

Unknown Tissue	Number of Genes	Percentage of Genes
$\hat{\mu}_{Tumor} \approx \hat{\mu}_{Normal}$	31149/32321	96.4%
$\hat{\mu}_{Tumor} \not\approx \hat{\mu}_{Normal}$	1172/32321	3.6%

* Here we define the relationship $\hat{\mu}_1 \approx \hat{\mu}_2$ as $0.95 < \frac{\hat{\mu}_1}{\hat{\mu}_2} < 1.05$ in the table, where μ denotes the sample mean of log2-transformed expression data.



Algorithm 1 Performing ICM for two-component

```
Parameter:
    Sample-wise $\{\pi_{1,i}\}_i : S$
    Gene-wise $\{\mu_{Tg}, \sigma_{Tg}\}_g : 2 \times G$
Initialize:
    $\{\mu_{Tg}, \sigma_{Tg}\}_{g=1}^{G} = \mu_0, \sigma_0$
for iteration $t = 1, \cdots, T$ do,
    a. update $\{\pi_{1,i}\}_{i=1}^{S}$
    for each sample $i = 1, \cdots, S$ do parallel
        update $\pi_{1,i}^{(t)} = argmax \prod_{g=1}^{G} h(y_{ig} | \pi_{1,i}, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^{G})$
    end for
    b. update $\{\mu_{Tg}, \sigma_{Tg}\}_{g=1}^{G}$
    for each gene $g = 1, \cdots, G$ do parallel
        update $\{\mu_{Tg}^{(t)}, \sigma_{Tg}^{(t)}\} = argmax \prod_{i=1}^{S} h(y_{ig} \left| \{\pi_1^{(t)}\}_{i=1}^{S}, \{\mu_{Tg}, \sigma_{Tg}\}\right.)$
    end for
end for
```

Algorithm 2 Performing ICM for three-component

```
Parameter:
    Sample-wise $\{\pi_{1,i}, \pi_{2,i}\}_i : 2 \times S$
    Gene-wise $\{\mu_{Tg}, \sigma_{Tg}\}_g : 2 \times G$
Initialize:
    $\{\mu_{Tg}, \sigma_{Tg}\}_{g=1}^{G} = \mu_0, \sigma_0$
for iteration $t = 1, \cdots, T$ do,
    a. update $\{\pi_{1,i}, \pi_{2,i}\}_{i=1}^{S}$
    for each sample $i = 1, \cdots, S$ do parallel
        update $\{\pi_{1,i}^{(t)}, \pi_{2,i}^{(t)}\} = argmax \prod_{g=1}^{G} f(y_{ig} | \{\pi_{1,i}, \pi_{2,i}\}, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^{G})$
    end for
    b. update $\{\mu_{Tg}, \sigma_{Tg}\}_{g=1}^{G}$
    for each gene $g = 1, \cdots, G$ do parallel
        update $\{\mu_{Tg}^{(t)}, \sigma_{Tg}^{(t)}\} = argmax \prod_{i=1}^{S} f(y_{ig} \left| \{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^{S}, \{\mu_{Tg}, \sigma_{Tg}\}\right.)$
    end for
end for
```

Figure S1. Outline of the ICM implementation in *DeMixT*, related to Figure 1.

The h() represents the full likelihood based on a single integral for a two-component model; and g() represents the full likelihood based on a double integral for a three-component model.

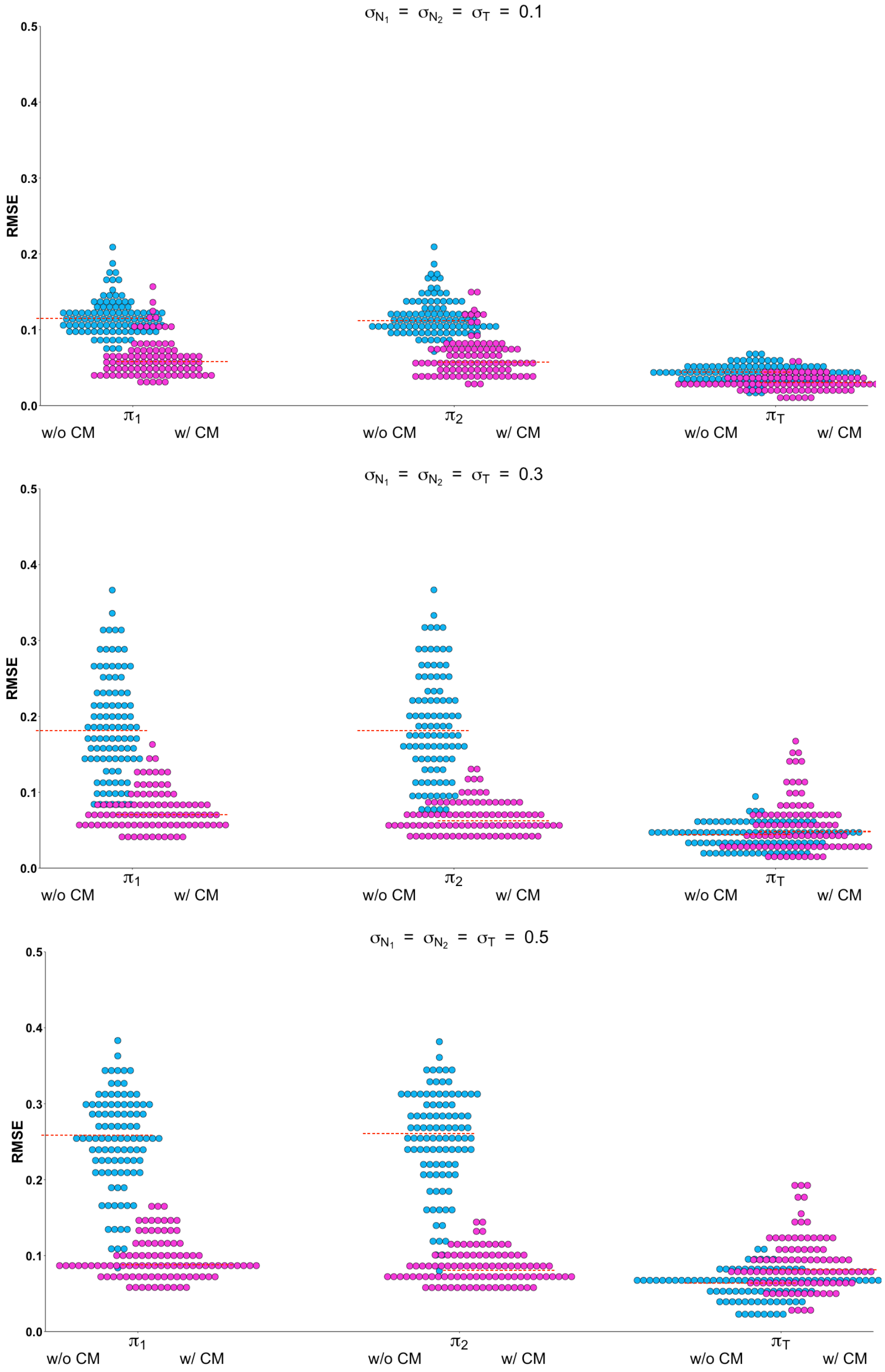


Figure S2. Dot plots of root mean square errors (RMSEs) between true and estimated proportions, using DeMixT with (w/) and without (w/o) component merging (CM), related to Figure 1.

We simulated 500 samples for 475 genes with $\mu_{N_1} \approx \mu_{N_1}$ and 25 genes with $\mu_{N_1} \not\approx \mu_{N_1}$, and repeated 25 times. Blue dots: deconvolution results without CM; red dots: those with CM; red dashed lines: median values.

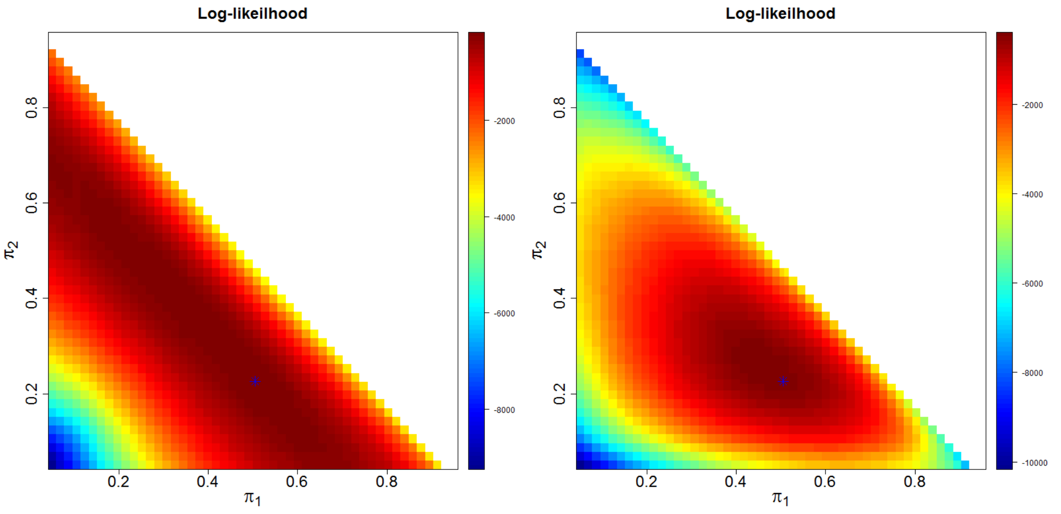


Figure S3. Log-Likelihood surface for π_1 and π_2, related to Figure 1.

The left panel shows that for 100 genes where $\mu_{N_1} \approx \mu_{N_2}$, π_1 and π_2 are not identifiable. The right panel shows for 100 genes where $\mu_{N_1} \not\approx \mu_{N_2}$, π_1 and π_2 are identifiable. The panels are generated from the same dataset, same sample, but on different sets of genes.

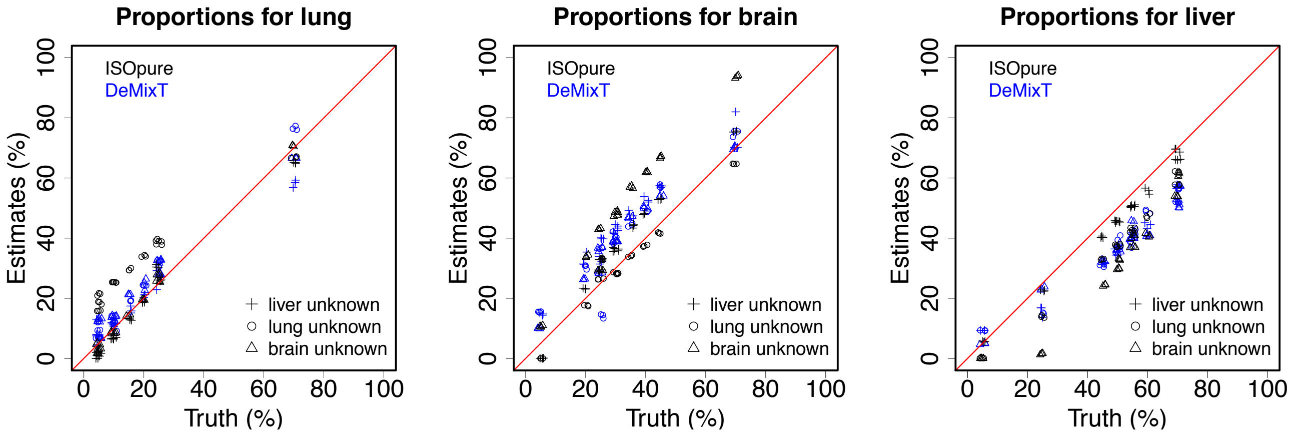


Figure S4. Scatter plots of estimated tissue proportions against true tissue proportions for the GSE19830 dataset, related to Figure 2.

All proportion estimates from running DeMixT are shown when either the liver, brain, or lung tissue is assumed to be the tissue with unknown expression profiles. Plus symbols: liver tissue is unknown; circles: lung tissue is unknown; triangles: brain tissue is unknown; blue: DeMixT; black: ISOpure.

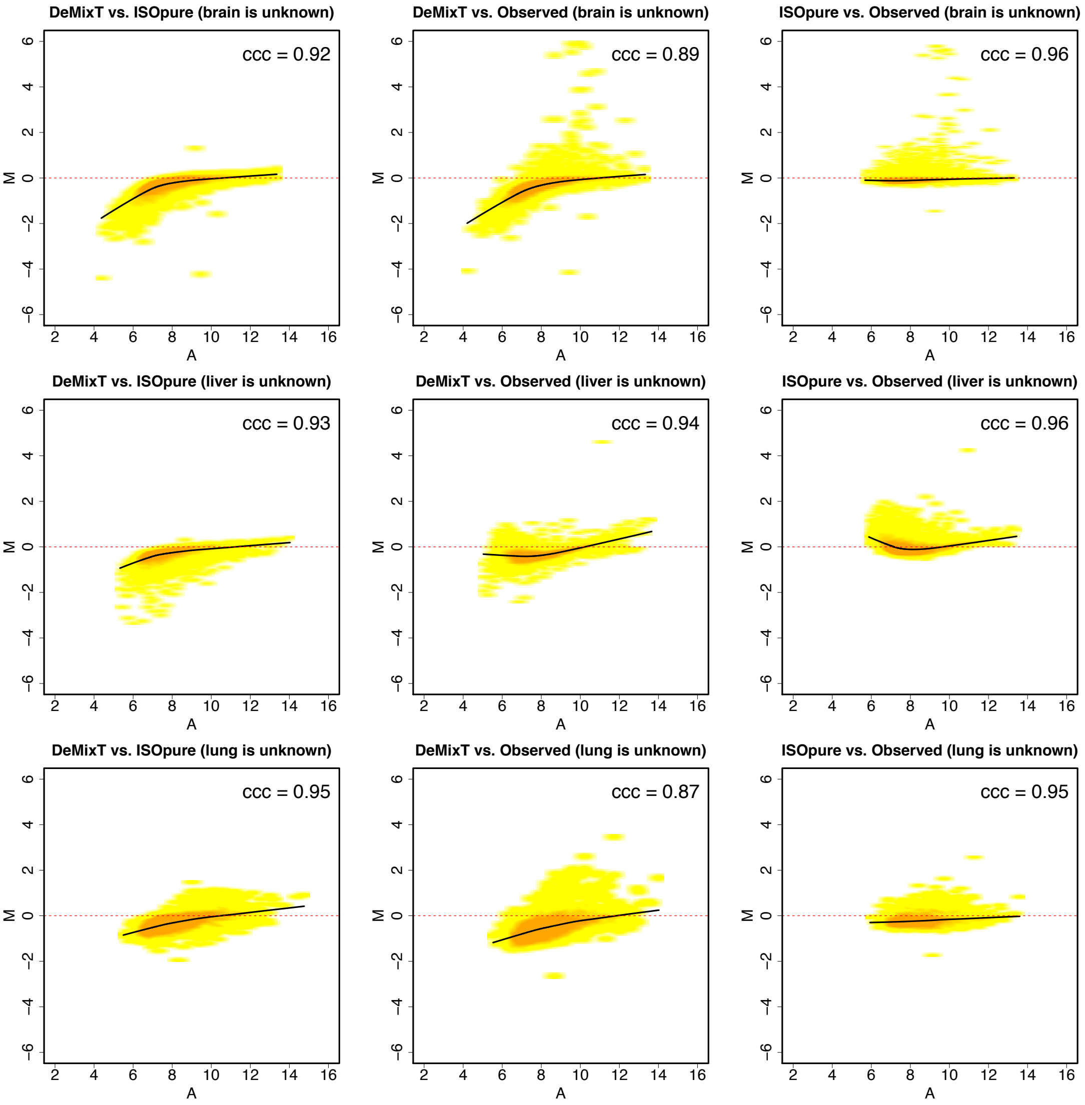


Figure S5: Smoothed scatter MA plots of mean estimated tissue-specific expression (at the log2 scale) from DeMixT and ISOpure in the GSE19830 dataset, related to Figure 2.

The MA plots compare the mean values of log2-transformed deconvolved expression levels across genes for DeMixT vs. ISOpure, DeMixT vs. observed samples, and ISOpure vs. observed samples, when either liver, lung or brain tissue was the unknown component. M: the difference in the two values; A: the average of the two values.

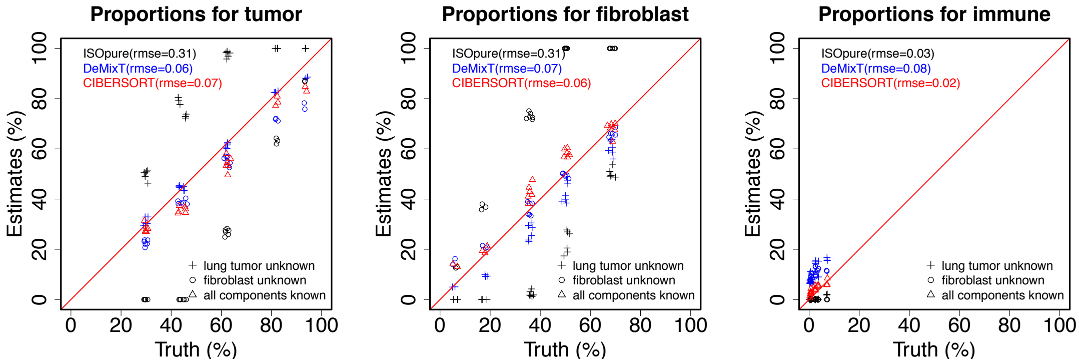


Figure S6. Scatter plots of estimated versus true proportions for the mixed cell line RNA-seq dataset, related to Figure 2.

All estimated proportions from DeMixT and ISOpure are shown when either lung tumor or fibroblast was the unknown component. Plus symbols: reference profiles of the lung cancer cell line are unknown; circles: reference profiles of the fibroblast cell line are unknown; triangle: the reference profiles of all the cell lines are known (only for CIBERSORT). Blue: DeMixT; black: ISOpure; red: CIBERSORT. Since CIBERSORT does not allow for any unknown component, the estimated proportions of CIBERSORT are based on the known reference genes from each component. DeMixT yielded proportion estimates with similar RMSE as CIBERSORT and much lower than ISOpure when compared to the truth.

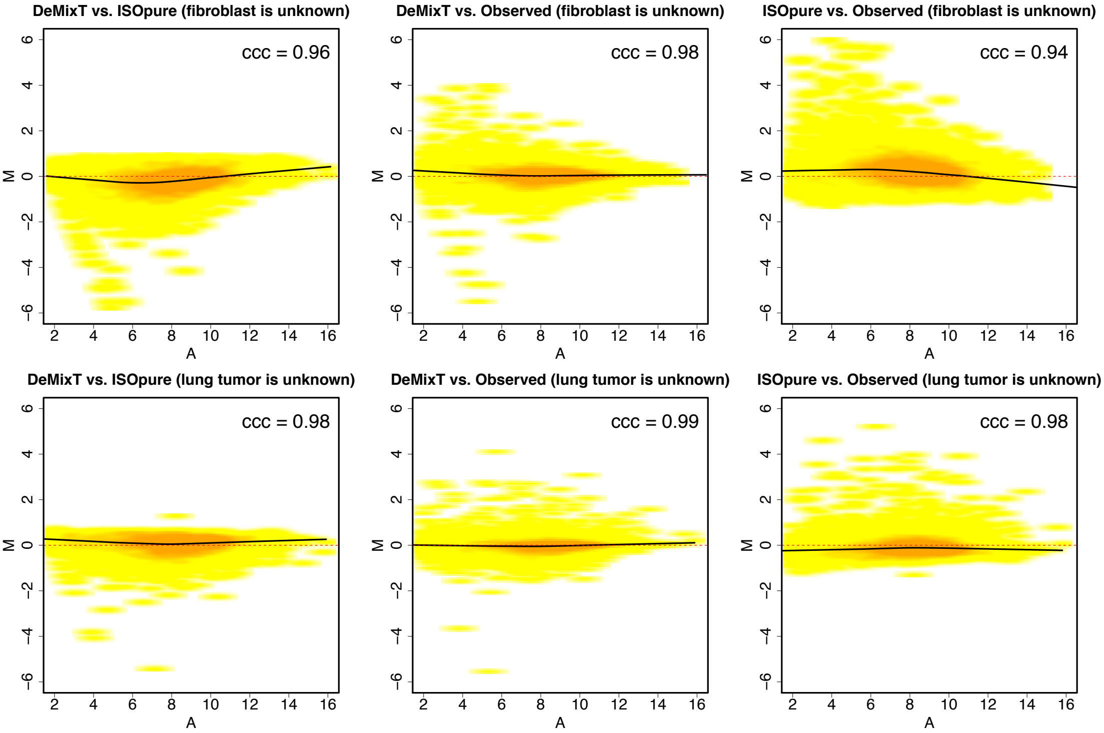


Figure S7. Smoothed scatter MA plots of mean estimated tissue-specific expression levels (at the log2 scale) from DeMixT and ISOpure in the mixed cell line RNA-seq dataset, related to Figure 2.

The MA plots compare mean values of log2-transformed deconvolved expression across genes for DeMixT vs. ISOpure, DeMixT vs. observed samples, and ISOpure vs. observed samples, when either lung cancer or fibroblast cell line was the unknown component. M: difference in the two values; A: average of the two values.

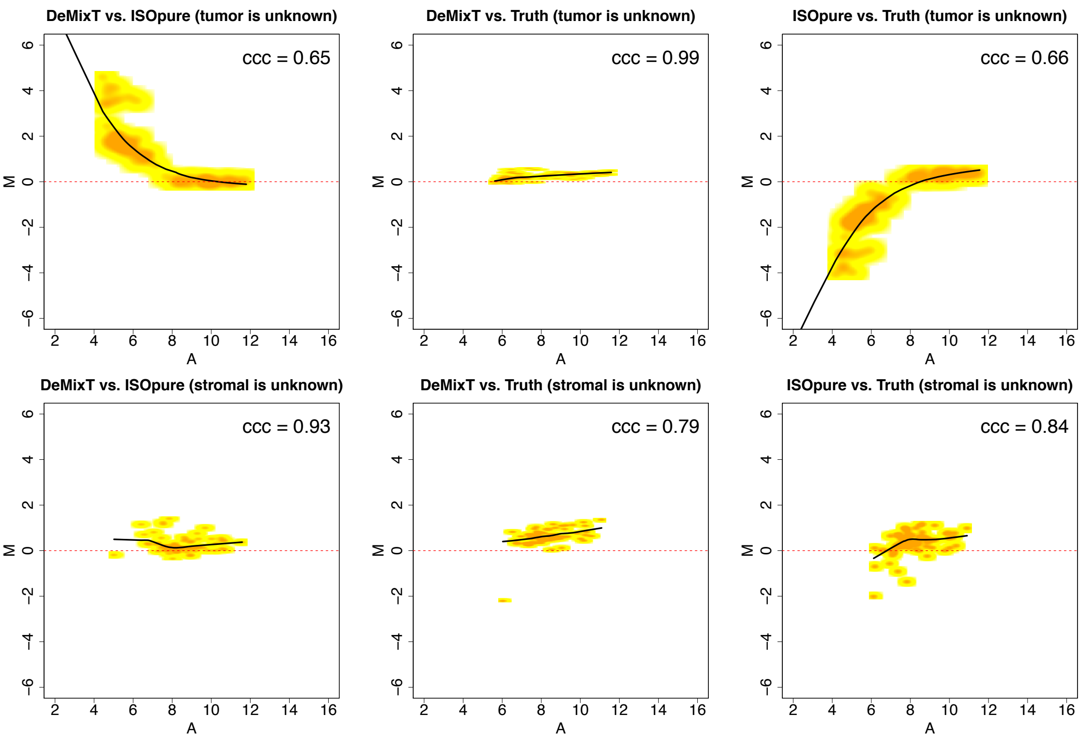


Figure S8: Smoothed scatter MA plots of mean estimated tissue-specific expression (at the log2 scale) between DeMixT and ISOpure in the LCM FFPE prostate cancer microarray dataset, related to Figure 3.

The MA plots compare mean values of log2-transformed deconvolved expression across genes for DeMixT vs. ISOpure, DeMixT vs. observed samples, and ISOpure vs. observed samples, when either tumor or stromal tissue was the unknown component. Shown are results from a pre-selected list of probesets (80 probesets) with the most differential expression between tumor and stromal tissues.

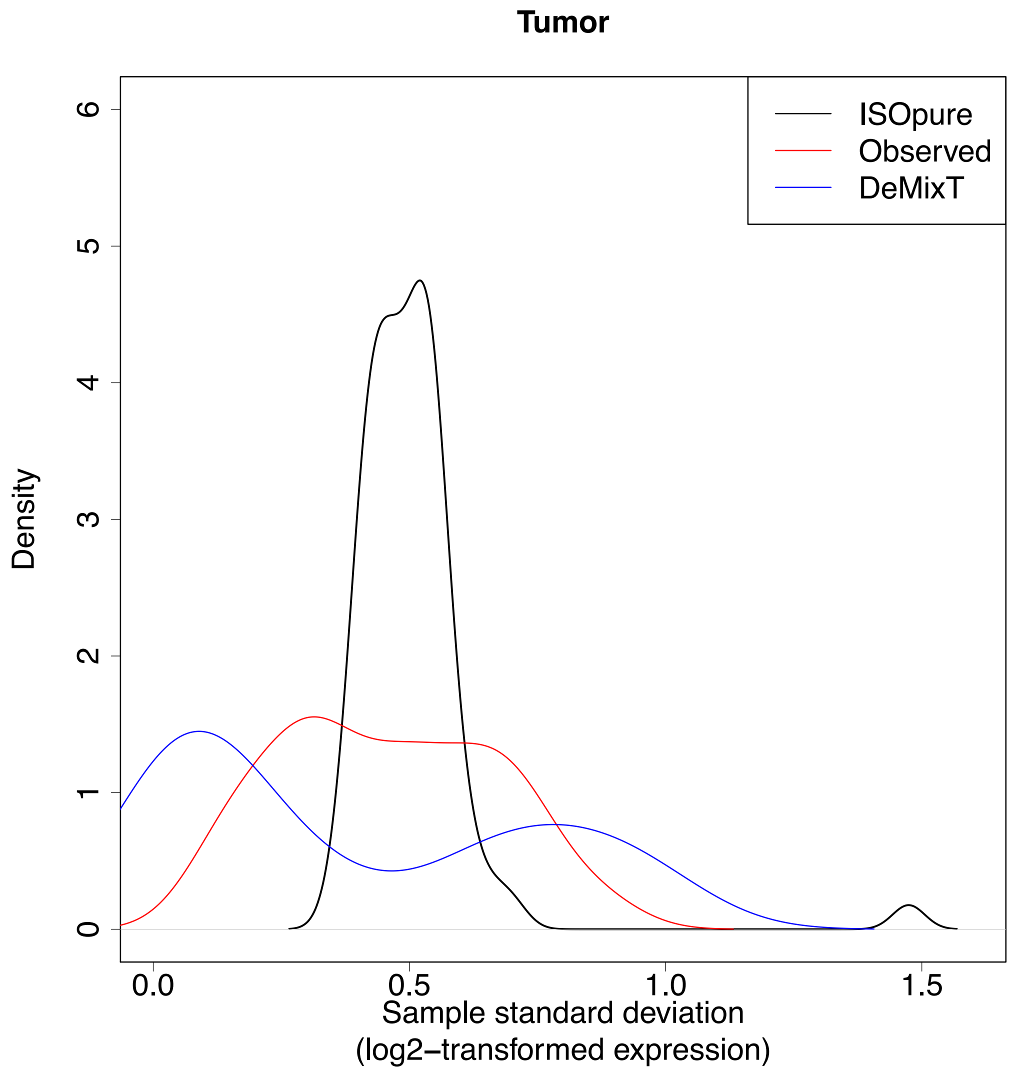


Figure S9. Density plot comparing sample standard deviations between deconvolved expression profiles of subset probes for DeMixT and ISOpure in the LCM FFPE prostate cancer microarray dataset when tumor tissue was assumed to be the unknown component; with measured expression profiles of isolated tumor tissues, related to Figure 3.

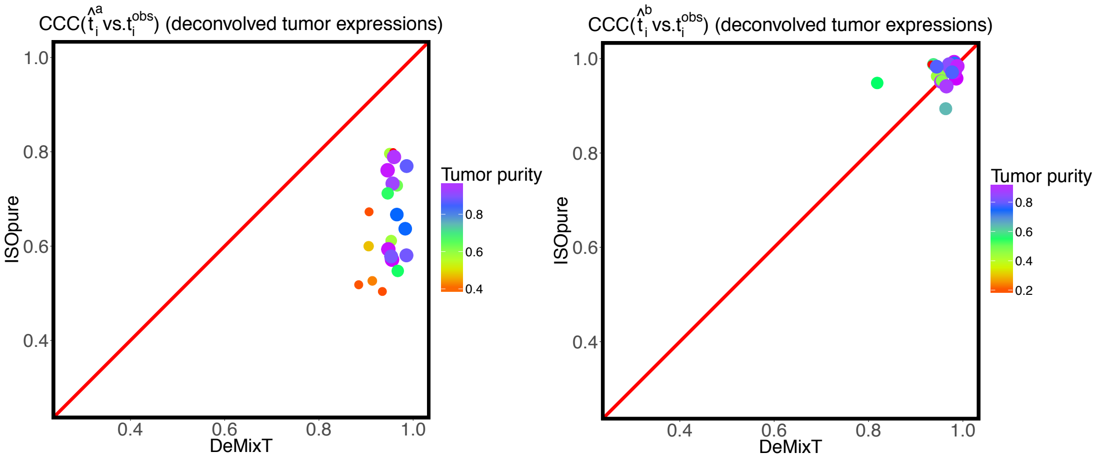


Figure S10. Scatter plots of concordance correlation coefficient (CCC) between individual deconvolved expressions and observed values for the tumor component in 23 LCM prostate samples, related to Figure 3. Each point corresponds to a sample. We compared results from ISOpure with those from DeMixT. Left panel shows the results when the expression data from stromal samples were taken as the input. Right panel shows the results when the expression data from tumor samples were taken as the input. The color gradient and size in each point corresponds to the estimated tumor proportions from DeMixT.

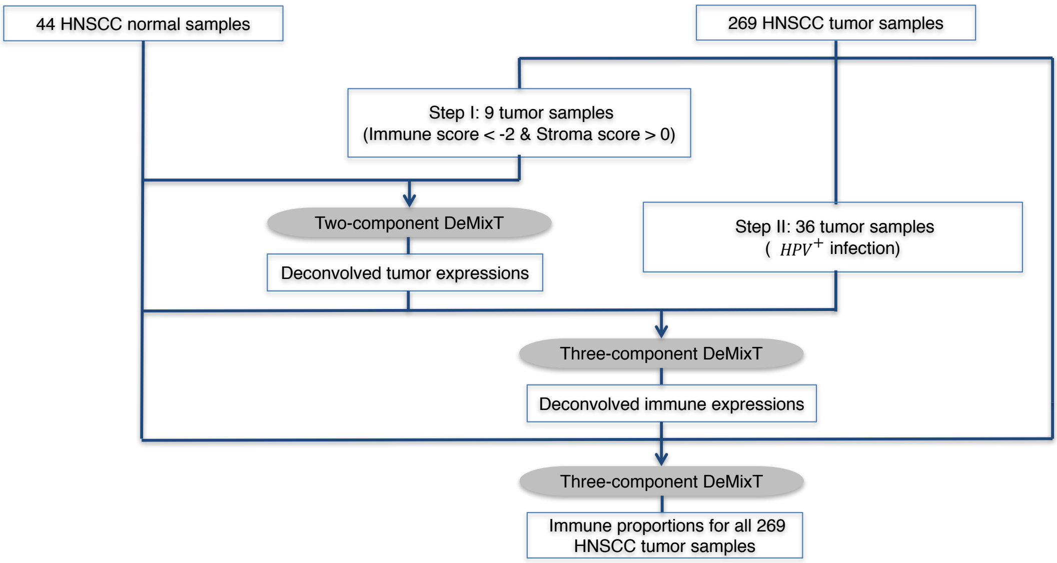

44 HNSCC normal samples
269 HNSCC tumor samples
Step I: 9 tumor samples
(Immune score < -2 & Stroma score > 0)
Two-component DeMixT
Deconvolved tumor expressions
Step II: 36 tumor samples
(HPV^+ infection)
Three-component DeMixT
Deconvolved immune expressions
Three-component DeMixT
Immune proportions for all 269 HNSCC tumor samples

Figure S11. Workflow for analysis of immune infiltration in the HNSCC dataset, related to Figure 4.

We obtained immune scores and stromal scores for all samples using the ESTIMATE method.

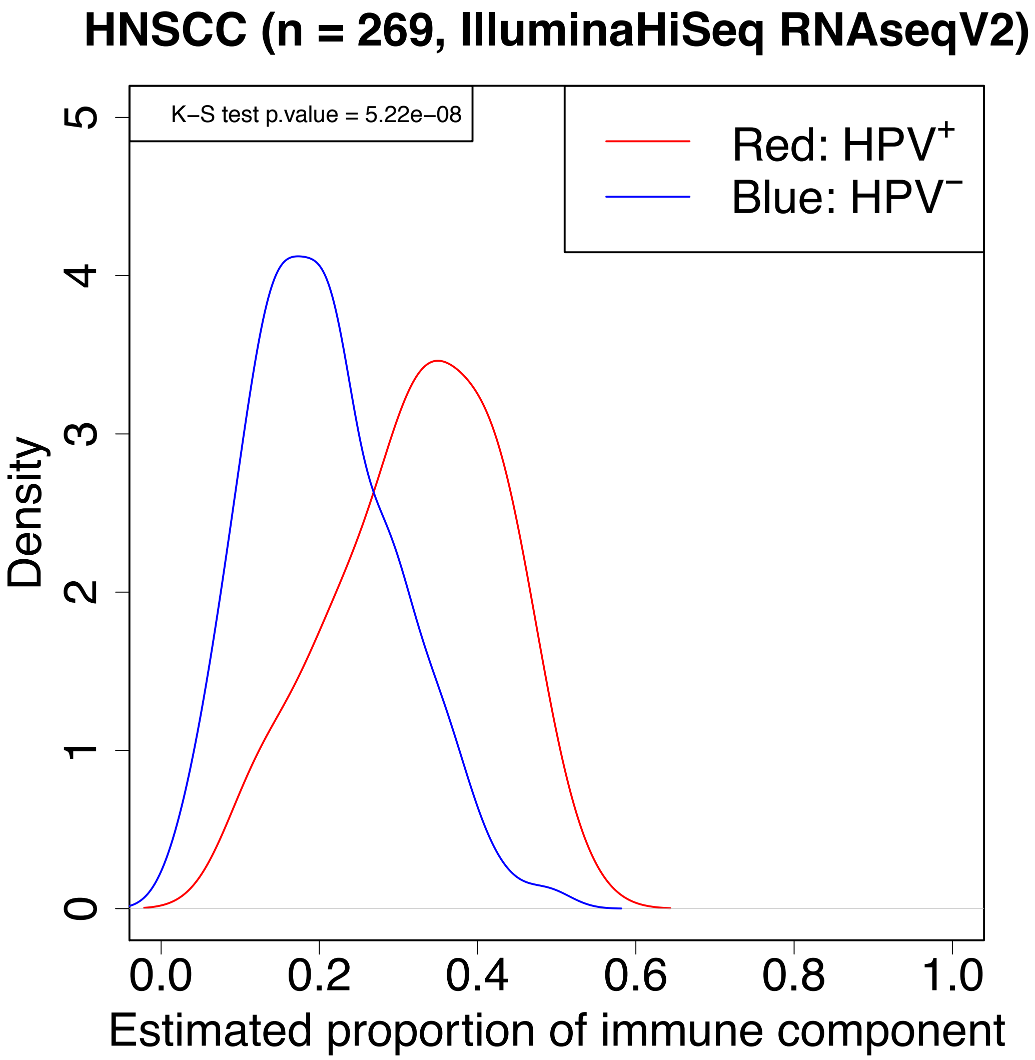


Figure S12. Density plot of estimated immune proportions for tumor samples with HPV test results, related to Figure 4.
Red curve: for those with HPV+ status; blue curve: for those with HPV-.

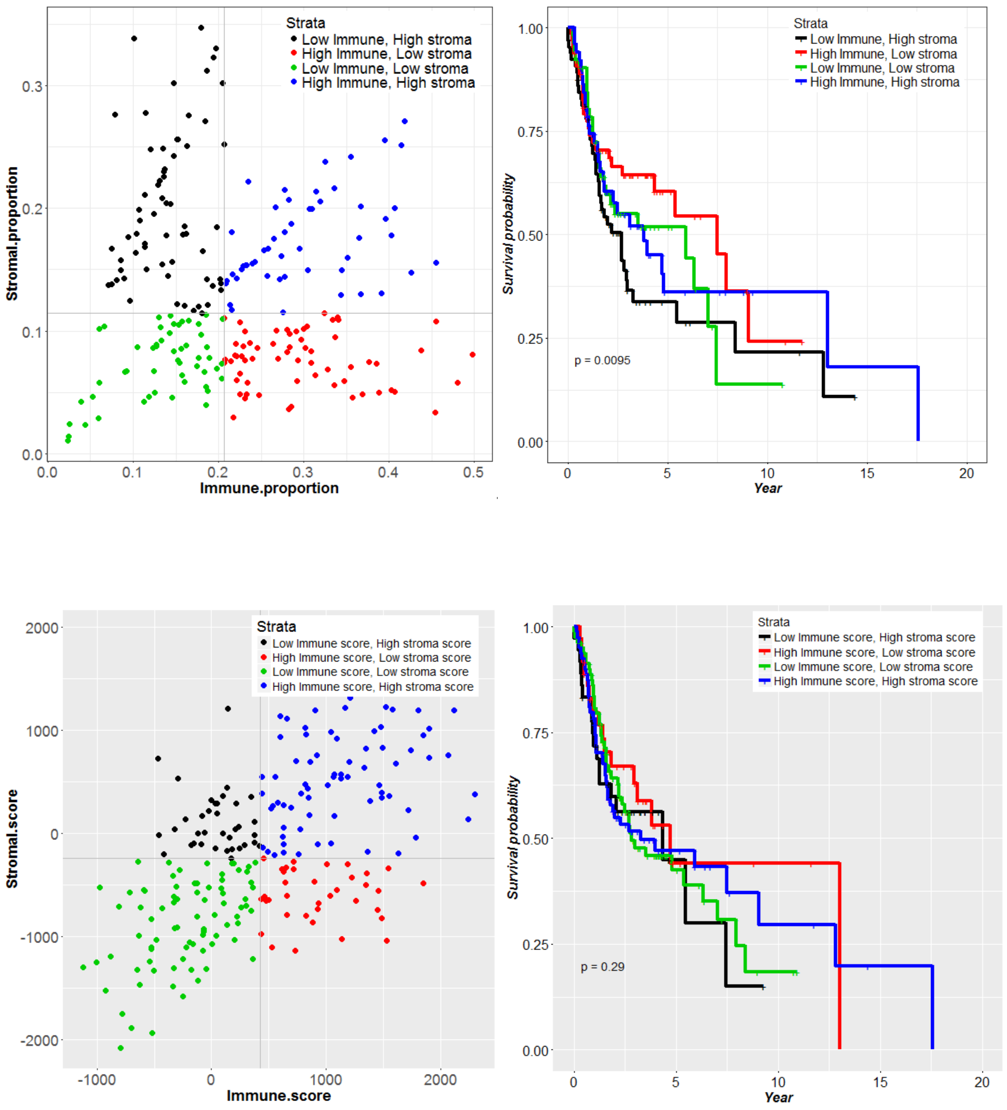


Figure S13. Association of immune-stroma-proportions from DeMixT with overall survival compared with association of immune-stroma-scores from ESTIMATE with overall survival in HNSCC, related to Figure 4.

Upper-left panel: a scatter plot of estimated immune- and stroma- proportions. Each point represents an HNSCC sample. Grey lines represent cutoffs that are used to divide patient samples into four groups. Upper-right panel: Kaplan-Meier curves of overall survival for HNSCC by those four patient groups given by the upper-left figure. The p-value of Cox regression model is calculated based on the Wald test. Bottom-left panel: a scatter plot of estimated immune- and stroma- scores from ESTIMATE. Grey lines represent cutoffs to divide patient samples into another four groups. Bottom-right panel: Kaplan-Meier curves of overall survival for HNSCC by those four patient groups given by the bottom-left figure.

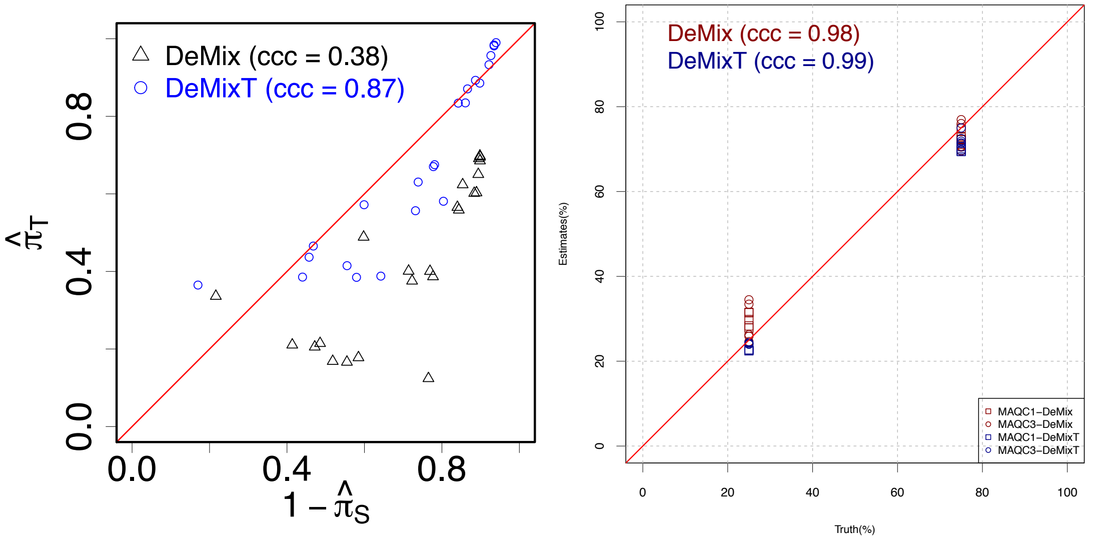


Figure S14. Comparison of proportion estimation between DeMixT and DeMix, related to Figure 1 and 2.

Left panel shows the scatter plot of estimated tumor proportions versus 1-estimated stromal proportions for the validation using LCM data in prostate cancer; estimates from DeMixT (blue) are compared with those from DeMix (black). Right panel shows the estimation of proportions, between DeMixT (blue) and DeMix (red), of the unknown component tissues from two available data sources that are given in the DeMix paper: MAQC1: MAQC site 1, MAQC3: MAQC site 3.

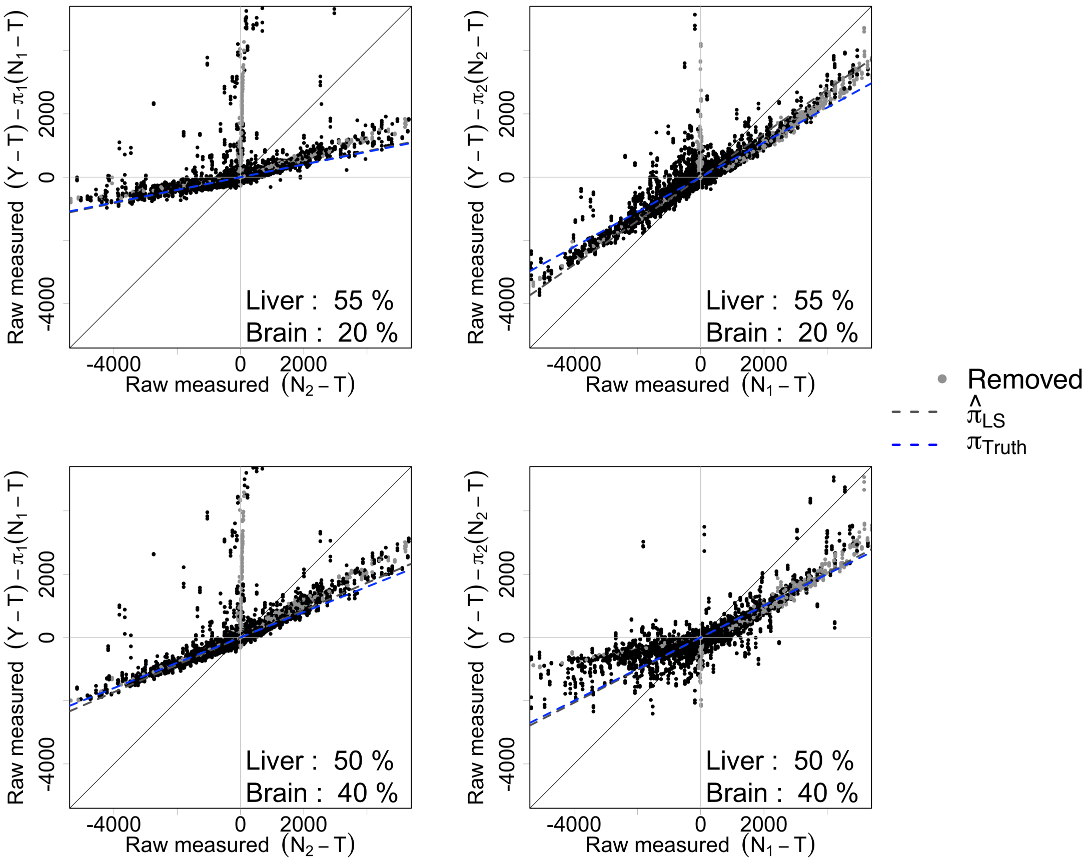


Figure S15. Scatter plots of $Y_{ig} - \overline{T}_g - \pi_{2,i}(\overline{N}_{2,g} - \overline{T}_g)$ versus $\overline{N}_{1,g} - \overline{T}_g$ for $\pi_{1,i}$ and $Y_{ig} - \overline{T}_g - \pi_{1,i}(\overline{N}_{1,g} - \overline{T}_g)$ versus $\overline{N}_{2,g} - \overline{T}_g$ for $\pi_{2,i}$ using the raw measured data from GSE19830 in two mixture scenarios, related to Figure 2.

Dark grey dashed line: fitted regression coefficient for all probes by least squares; blue dashed line: true mixing proportion; light grey dots: probesets removed with the criterion that the mean expression after log2-transformation is less than 7 in either *N₁* or *N₂*; black dots: remaining probes. If the linearity assumption holds, the fitted line should lie approximately on the truth.

Transparent Methods

Model

Let Y_{ig} be the observed expression levels of the raw measured data from clinically derived malignant tumor samples for gene $g, g = 1, \cdots, G$ and sample $i, i = 1, \cdots, S$. G denotes the total number of probes/genes and S denotes the number of samples. The observed expression levels for solid tumors can be modeled as a linear combination of raw expression levels from three components:

$$Y_{ig} = \pi_{1,i} N_{1,ig} + \pi_{2,i} N_{2,ig} + (1 - \pi_{1,i} - \pi_{2,i}) T_{ig} \tag{1}$$

Here $N_{1,ig}$, $N_{2,ig}$ and T_{ig} are the unobserved raw expression levels from each of the three components. We call the two components for which we require reference samples the N_1-component and the N_2-component. We call the unknown component the T-component. We let $\pi_{1,i}$ denote the proportion of the N_1-component, $\pi_{2,i}$ denote the proportion of the N_2-component, and $1 - \pi_{1,i} - \pi_{2,i}$ denote the proportion of the T-component. We assume that the mixing proportions of one specific sample remain the same across all genes. Our model allows for one component to be unknown, and therefore does not require reference profiles from all components. A set of samples for $N_{1,ig}$ and $N_{2,ig}$, respectively, needs to be provided as input data. This three-component deconvolution model is applicable to the linear combination of any three components in any type of material. It can also be simplified to a two-component model, assuming there is just one N-component. For application in this paper, we consider tumor (T), stromal (N_1) and immune components (N_2) in an admixed sample (Y). Following the convention that $\log_2$-transformed microarray gene expression data follow a normal distribution, we assume that the raw measures $N_{1,ig} \sim LN(\mu_{N_1 g}, \sigma^2_{N_1 g})$, $N_{2,ig} \sim LN(\mu_{N_2 g}, \sigma^2_{N_2 g})$ and $T_{ig} \sim LN(\mu_{Tg}, \sigma^2_{Tg})$, where LN denotes a $\log_2$-normal distribution and $\sigma^2_{N_1 g}$,$\sigma^2_{N_2 g}$,σ^2_{Tg} reflect the variations under $\log_2$-transformed data (Ahn et al., 2013; Lönnstedt and Speed, 2002). Consequently, our model can be expressed as the convolution of the density function for three $\log_2$-normal distributions. Because there is no closed form of this convolution, we use numerical integration to evaluate the complete likelihood function.

Our model expressed as the convolution of the density function for three log2-normal distributions.

$$\begin{aligned}
L &= \prod_{i=1}^{S}\prod_{g=1}^{G} f(y_{ig}\,|\mu_{T_g}, \mu_{N_{1g}}, \mu_{N_{2g}}, \sigma_{T_g}, \sigma_{N_{1g}}, \sigma_{N_{2g}}, \pi_{1,i}, \pi_{2,i}) \\
&\propto \prod_{i=1}^{S}\prod_{g=1}^{G}\{\int_0^y \frac{1}{n'_{2,ig}\sigma_{N_{2g}}}\exp[-\frac{\{\log_2(n'_{2,ig}) - \mu_{N_{2g}} - \log_2(\pi_{2,i})\}^2}{2\sigma^2_{N_{2g}}}]\frac{1}{n'_{1,ig}\sigma_{N_{1g}}} \\
&\times \int_0^{y-n'_2}\exp[-\frac{\{\log_2(n'_{1,ig}) - \mu_{N_{1g}} - \log_2(\pi_{1,i})\}^2}{2\sigma^2_{N_{1g}}}]\frac{1}{(y_{ig} - n'_{1,ig} - n'_{2,ig})\sigma_{T_g}} \\
&\times \exp[-\frac{\{\log_2(y_{ig} - n'_{1,ig} - n'_{2,ig}) - \mu_{T_g} - \log_2(1 - \pi_{1,i} - \pi_{2,i})\}^2}{2\sigma^2_{T_g}}]dn'_{1,ig}dn'_{2,ig}\}
\end{aligned} \tag{2}$$

where $n'_{1,ig} = \pi_{1,i}n_{1,ig}$ and $n'_{2,ig} = \pi_{2,i}n_{2,ig}$.

The *DeMixT* algorithm for deconvolution

DeMixT estimates all distribution parameters and cellular proportions and reconstitutes the expression profiles for all three components for each gene and each sample, as shown in equation (1). The estimation procedure (summarized in **Figure 1b**) has two main steps as follows.

1. Obtain a set of parameters $\{\pi_{1,i}, \pi_{2,i}\}_{i=1}^{S}$, $\{\mu_T, \sigma_T\}_{g=1}^{G}$ to maximize the complete likelihood function, for which $\{\mu_{N_{1,g}}, \sigma_{N_{1,g}}, \mu_{N_{2,g}}, \sigma_{N_{2,g}}\}_{g=1}^{G}$ were already estimated from the available unmatched samples of the N_1 and N_2 component tissues. This step is described in further details below in parameter estimation and the GSCM approach.

2. Reconstitute the expression profiles by searching each set of $\{n_{1,ig}, n_{2,ig}\}$ that maximizes the joint density of $N_{1,ig}$, $N_{2,ig}$ and T_{ig}

$$\begin{aligned}
&\underset{n_{1,ig},n_{2,ig}}{\arg\max}\,\phi(\frac{y_{ig} - \hat{\pi}_{1,i}n_{1,ig} - \hat{\pi}_{2,i}n_{2,ig}}{1 - \hat{\pi}_{1,i} - \hat{\pi}_{2,i}}\,|\hat{\mu}_{T_g}, \hat{\sigma}_{T_g}) \\
&\times \quad \phi(n_{1,ig}\,|\hat{\mu}_{N_{1g}}, \hat{\sigma}_{N_{1g}})\phi(n_{2,ig}\,|\hat{\mu}_{N_{2g}}, \hat{\sigma}_{N_{2g}})
\end{aligned} \tag{3}$$

where $\phi(.|\mu, \sigma^2)$ is a log2-normal distribution density with location parameter μ and scale parameter σ.

In step 2, we combined the golden section search method with successive parabolic interpolations to find the maximum of the joint density function with respect to $n_{1,ig}$ and $n_{2,ig}$ that are positively bounded and constrained by $\hat{\pi}_{1,i}n_{1,ig} + \hat{\pi}_{2,i}n_{2,ig} \le y_{ig}$. The value of t_{ig} is solved as $y_{ig} - \hat{\pi}_{1,i}n_{1,ig} - \hat{\pi}_{2,i}n_{2,ig}$.

Parameter estimation using iterated conditional modes (ICM)

In step 1, the unknown parameters to be estimated can be divided into two groups: gene-wise parameters, $\{\mu_T, \sigma_T\}_{g=1}^{G}$, and sample-wise parameters, $\{\pi_1, \pi_2\}_{i=1}^{S}$. These two groups of parameters are conditionally independent (**Figure 1b**). For each pair of gene-wise parameters, we have
$\{\pi_1, \pi_2\}_i \perp\!\!\!\perp \{\pi_1, \pi_2\}_j | \{\mu_T, \sigma_T\}_{k=1}^{G}$, for all $i \neq j \in \{1, \cdots, S\}$, and similarly for each pair of sample-wise parameters, we have $\{\mu_T, \sigma_T\}_i \perp\!\!\!\perp \{\mu_T, \sigma_T\}_j | \{\pi_1, \pi_2\}_{k=1}^{S}$, for all $i \neq j \in \{1, \cdots, G\}$. These relationships allow us to implement an optimization method, ICM, to iteratively derive the conditional modes of each pair of gene-wise or sample-wise parameters, conditional on the others (Besag, 1986). Here, π_1, π_2 are constrained between 0 and 1, and μ_T, σ_T are positively bounded. We combined a golden section search and successive parabolic interpolations to find a good local maximum (Brent, 1973) in each step. As shown by Besag (Besag, 1986), for ICM, the complete likelihood never decreases at any iteration and the convergence to the local maximum is guaranteed. Our ICM implementation is described in **Figure S1**.

The GSCM approach to improve model identifiability

Due to the high dimension of the parameter search space, and often flat likelihood surfaces in certain regions of the true parameters (e.g., $\mu_1 \approx \mu_2$) that will be encountered by ICM (**Figure S3**), we have developed a GSCM approach (illustrated in **Figure 1b**) to focus on the hilly part of the likelihood space. This reduces the parameter search space and improves the accuracy and computational efficiency. Here, we describe our general strategy. As there are large variations in the number of genes that are differentially expressed across datasets, the actual cutoffs may be adjusted for a given dataset.

Stage 1 We first combine the N_1 and N_2 components and assume a two-component mixture instead of three. This allows us to quickly estimate π_T.

a: We select a gene set containing genes with small standard deviations (< 0.1 or 0.5) for both the N_1 and N_2 components. Among these genes, we further select genes with $\overline{LN}_{1g} \approx \overline{LN}_{2g}$ (mean difference < 0.25 or 0.5), where the $\overline{LN}$ is the sample mean for the log2-transformed data. Within this set, we further select genes with the largest sample standard deviations of Y_g (top 250), suggesting differential expression between T and N.

b: We run *DeMixT* in the two-component setting to estimate μ_{Tg}, σ_{Tg}^2 and π_T.

Stage 2 We then fix the values of $\{\pi_T\}_i$ as derived from Stage 1, and further estimate $\{\pi_1\}_i$ and $\{\pi_2\}_i$ in the three-component setting.

a: We select genes with the greatest difference in the mean expression levels between the N_1 and N_2 components as well as those with the largest sample standard deviations of Y_g (top 250).

b: We run *DeMixT* in the three-component setting over the selected genes to estimate π_1 and π_2 given π_T.

c: We estimate the gene-wise parameters for all genes given the fixed π's. Finally, given all parameters, per gene per sample expression level, $n_{1,ig}$, $n_{2,ig}$ and t_{ig} are reconstituted.

Simulation study for the GSCM approach

To demonstrate the utility of GSCM for parameter estimation, we simulated a dataset with expression levels from 500 genes and 90 samples, 20 of pure N_1-type, 20 of pure N_2-type and 50 mixed samples. For the 50 mixed samples, we generated their proportions for all three components $(\pi_1, \pi_2, \pi_T) \sim Dir(1, 1, 1)$, where Dir is a Dirichlet distribution. For each mixed sample, we simulated expression levels of 500 genes for the N_1 and T-component from a $\log_2$-normal distribution with $\mu_{N_{1g}}$ and μ_{Tg} from $N_{[0,+\infty]}(7, 1.5^2)$, and with equal variance. For the N_2-component, we generated $\mu_{N_{2g}}$ from $\mu_{N_{1g}} + d_g$, where $d_g \sim N_{[-0.1,0.1]}(0, 1.5^2)$ for 475 genes ($\hat{\mu}_{N1g} \approx \hat{\mu}_{N2g}$) and $d_g \sim N_{[0.1,3]}(0, 1.5^2) \cup N_{[-3,-0.1]}(0, 1.5^2)$ for 25 genes ($\hat{\mu}_{N1g} \not\approx \hat{\mu}_{N2g}$). Then we mixed the N_1, N_2 and T-component expression levels linearly at the generated proportions according to our convolution model. We created a full matrix consisting of 20 N_1-type reference samples (generated separately from the N_1 distribution), 20 N_2-type reference samples (generated separately from the N_2 distribution) and 50 mixed samples at each simulation and repeated the simulation 100 times for each of the three variance values $\sigma \in \{0.1, 0.3, 0.5\}$ to finally obtain 300 simulation repeats. We first ran *DeMixT* with GSCM, where we used 475 genes with simulated $\hat{\mu}_{N1g} \approx \hat{\mu}_{N2g}$ to run the two-component deconvolution (N versus T) and used the remaining 25 genes to run the three-component deconvolution with estimated $\hat{\pi}_T$. We also ran *DeMixT* without GSCM using all 500 genes.

Data analysis

All analyses were performed using the open-source environment R (http://cran.r-project.org). Documentation (knitr-html) of all scripts is provided at the GitHub repository.

Mixed tissue microarray dataset

We downloaded dataset GSE19830 (Shen-Orr et al., 2010a) from the GEO browser. We used the R package *{affy}* to summarize the raw probe intensities with quantile normalization but without background correction as recommended in previous studies (Liebner, K. Huang, and Parvin, 2014). We evaluated the performance of *DeMixT* with regard to tissue proportions and deconvolved expression levels on the set of genes that were selected based on the GSCM approach. Specifically, we selected genes with sample standard deviation < 0.1 in N_1 and N_2 components, among which we used those with $\overline{LN}_{1g} - \overline{LN}_{2g} < 0.25$ for running the 2-compoment model, and used the top 250 genes with largest $\overline{LN}_{1g} - \overline{LN}_{2g}$ and largest sample standard deviation in Y for running the 3-component model. Then we ran ISOpure for the purpose of comparison.

Analysis of microarray data from mixed RNA from rat tissues: brain, liver and lung (Table S1).

Checking for the linearity assumption. Our DeMixT model relies on the assumption that the tissue-specific expression levels are combined linearly to create the observed Y. In the mixed tissue data, we can check for the validity of this assumption when T_{ig}'s and N_{ig}'s are known. Based on the linear equation, we have

$$Y_{ig} = \pi_{1,i}N_{1,ig} + \pi_{2,i}N_{2,ig} + (1 - \pi_{1,i} - \pi_{2,i})T_{ig} \Leftrightarrow \begin{cases} \pi_{1,i} = \frac{Y_{ig}-T_{ig}-\pi_{2,i}(N_{2,ig}-T_{ig})}{N_{1,ig}-T_{ig}} \\ \pi_{2,i} = \frac{Y_{ig}-T_{ig}-\pi_{1,i}(N_{1,ig}-T_{ig})}{N_{2,ig}-T_{ig}} \end{cases} \tag{4}$$

Thus, we generated scatter plots with a regression line to compare $Y_{ig} - \bar{T}_g - \pi_{2,i}(\bar{N}_{2,g} - \bar{T}_g)$ with $\bar{N}_{1,g} - \bar{T}_g$ and $Y_{ig} - \bar{T}_g - \pi_{1,i}(\bar{N}_{1,g} - \bar{T}_g)$ with $\bar{N}_{2,g} - \bar{T}_g$, where the sample mean for $N_{1,g}$(e.g. Liver), $N_{2,g}$(e.g. Brain) and T_g(e.g. Lung) were used instead of each $N_{1,ig}$, $N_{2,ig}$ and T_{ig}. In this dataset, the repeats were technical and presented little variation across samples, which allowed us to simply use sample means as surrogates for the expressions from individual samples.

As illustrated in **Figure S15** with 2 mixture scenarios (liver: brain: lung at 55:20:25 and 50:40:10), the linearity assumption holds reasonably within most samples; however, there was always a small set of probes that deviated from the linear line and formed a vertical line at 0 on the x-axis: N_1-T or N_2-T. We found that a criterion on probesets with mean expression (log2-transformed) < 7 in either N_1 or N_2 can accurately identify this set and therefore remove them, suggesting a potential cause of such behavior is the expression levels below the reliable detection range of microarrays, with noise overtaking the signal in the N-components in these probesets.

Deconvolution results. DeMixT showed high concordance correlations and small root mean squared errors (RMSEs) between the estimates and the true proportions of all three tissues in deconvolution, irrespective of which tissue was assumed as the unknown component that was without available knowledge for expression profiles. DeMixT gave accurate estimates for the proportions of the unknown component. ISOpure also performed well in estimating the proportions of the unknown tissues **Supplementary Tables 4-5**). A stable deconvolution algorithm should provide similar estimates of tissue-specific proportions no matter which component is assumed to be unknown. We assessed this through a reproducibility statistic and found that DeMixT was more stable than ISOpure (**Table S3, Figure 2a** and **Figure S4**). Both DeMixT and ISOpure yielded accurate estimates of the mean expression levels for each tissue component (**Figure S5**).

Mixed cell line RNA-seq dataset

This dataset was generated in house by mixing RNAs from three cell lines at fixed proportions. We mapped raw reads generated from paired-end Illumina sequencing to the human reference genome build 37.2 from NCBI through TopHat (default parameters and supplying the -G option with the GTF annotation file downloaded from the NCBI genome browser). The mapped reads obtained from the TopHat output were cleaned by SAMtools to remove improperly mapped and duplicated reads. We then used Picard tools to sort the cleaned SAM files according to their reference sequence names and create an index for the reads. The gene-level expression was quantified by applying the R packages GenomicFeatures and GenomicRanges. We generated a reference table from the human reference genome hg19 and then used the function findOverlaps to count the number of reads mapped to each exon for all the samples. This count dataset was pre-processed by total count normalization, and genes that contained zero counts were removed. The pre-processed count data were used as input for *DeMixT* and ISOpure. We performed the same GSCM step as in the analysis of mixed tissue microarray data.

Analysis of RNA-seq data from RNA from mixed cell lines: H1092, CAF and TIL (Table S2).

DeMixT yielded proportion estimates with higher CCC and smaller errors (average RMSE = 0.06, 0.07) than ISOpure (average RMSE = 0.18 and 0.24) when compared to the truth (**Figure 2b, Supplementary Tables 6-7**). Proportion estimates were consistent when different components were treated as unknown in our experiments (**Table S3** and **Figure S6**). Both DeMixT and ISOpure overestimated the immune proportions when lymphocytes were unknown, which had low proportions (0.4-7.1%) in all mixed samples, but the degree of overestimation from DeMixT was smaller. In the two scenarios in which DeMixT was able to identify the lymphocyte component, we estimated tissue-specific expressions for all the genes with non-zero counts, and found high concordance (> 0.98) between the deconvolved expression estimates and mean expression levels. Again, we observed smaller differences in mean expression levels across genes when using DeMixT compared to ISOpure (**Figure S7**).

Laser-capture microdissection (LCM) prostate cancer FFPE microarray dataset

This dataset was generated at the Dana Farber Cancer Institute (GSE97284 (Tyekucheva et al., 2017a)). Radical prostatectomy specimens were annotated in detail by pathologists, and regions of interest were identified that corresponded to benign epithelium, prostatic intraepithelial neoplasia (abnormal tissue that is possibly precancerous), and tumor, each with its surrounding stroma. These regions were laser-capture microdissected using the ArcturusXT system (Life Technologies). Additional areas of admixed tumor and adjacent stromal tissue were taken. FFPE samples are known to generate overall lower quality expression data than those from fresh frozen samples. We observed a small proportion of probesets that presented large differences in mean expression levels between the dissected tissues: tumor (T) and stroma (N) in this dataset (**Table S9**). Only 53 probesets presented a mean difference $(|\overline{T} - \overline{N}|) > 1$, as compared to $10,397$ probesets in GSE19830. We therefore chose the top 80 genes with the largest mean differences and ran both *DeMixT* and ISOpure under two settings: tumor unknown and stroma unknown.

TCGA HNSCC data

We downloaded RNA-seq data for HNSCC from TCGA data portal (https://portal.gdc.cancer.gov/). There was a total of 44 normal and 269 tumors samples for HNSCC. We collected the information of HPV infection for the HNSCC samples. Samples were classified as HPV+ using an empiric definition of detection of $>$ 1000 mapped RNA-seq reads, primarily aligning to viral genes E6 and E7, which resulted in 36 HPV+ samples (Cancer Genome Atlas Network, 2015). We then devised a workflow to estimate the immune cell proportions (**Figure S11**). Our workflow included three steps. The downloaded normal samples provided reference profiles for the stromal component in each step. We first downloaded stromal and immune scores from single-sample gene set enrichment analysis for all of our tumor samples (Yoshihara et al., 2013) and selected 9 tumor samples with low immune scores (< -2) and high stromal scores (> 0), which suggested that these samples were likely low in immune infiltration. We then ran *DeMixT* under the two-component mode on these samples, generating the deconvolved expression profiles for the tumor and stromal components. We used these profiles as reference samples for running *DeMixT* under the three-component mode in the 36 HPV^+ samples, generating deconvolved expression profiles for the immune component. In these two steps, we used deconvolved profiles that have smaller estimated standard variations as the reference profiles for the next step. We then ran *DeMixT* under the three-component mode on all 269 samples with reference profiles from normal samples and the deconvolved immune component. We calculated p-values (Benjamini-Hochberg corrected (Benjamini and Hochberg, 1995)) for the differential test of deconvolved expressions for the immune component versus the stromal component, and for the immune component versus the tumor component, respectively, on a set of 63 immune marker genes. We performed gene selection in the GSCM approach (as described above), with a slightly larger threshold to account for the large sample size: sample standard deviation < 0.6 and the top 500 genes for three-component deconvolution to estimate the π's.

Summary statistics for performance evaluation.

Concordance correlation coefficient (CCC). To evaluate the performance of our method, we use the CCC and RMSE. The CCC ρ_{xy} is a measure of agreement between two variables x and y and is defined as $\rho_{xy} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}$, where μ and σ^2 are the corresponding mean and variance for each variable, and ρ is the correlation coefficient between the two variables. We calculate the CCC to compare the estimated and true proportions to evaluate the proportion estimation. We also calculate the CCC to compare the deconvolved and observed expression values (log_2-transformed).

Measure of reproducibility. To assess the reproducibility of the estimated π across scenarios when the different components are unknown (i.e., three scenarios for a three-component model with one unknown component), we define a statistic $R = \frac{1}{S}\sum_i^S(\frac{1}{K-1}\sum_k^K(\epsilon_i^k - \frac{1}{K}\sum_k^K \epsilon_i^k)^2)^{\frac{1}{2}}$, where $\epsilon_i^k = \hat{\pi}_i^k - \pi_i$, $\hat{\pi}_i^k$ is the estimated value for the k-th scenario and π_i is the truth for sample i. S denotes the sample size and K is the number of scenarios. This measures the variations in the estimation errors across different scenarios. We consider a method with a smaller R as more reproducible and therefore more desirable.

Data and software availability

The public data used in this study are GSE19830 (Shen-Orr et al., 2010b) and GSE97284 (Tyekucheva et al., 2017b) from GEO browser, and RNA-seqV2 count data from the Genomic Data Commons Data Portal (*Genomic Data Commons Data Portal: TCGA Head and Neck Squamous Carcinoma* n.d.). The RNA-seq count data used for validation were generated from our lab and can be downloaded from `https://github.com/wwylab/DeMixTallmaterials`. The accession number for the FASTQ files of the RNA-seq count data reported in this paper is GEO: GSE121127. The *DeMixT* source code and the entire analytic pipeline are available at `https://github.com/wwylab/DeMixTallmaterials`.

References

Ahn, J. et al., 2013. DeMix: Deconvolution for Mixed Cancer Transcriptomes Using Raw Measured Data. *Bioinformatics*, 29(15), pp. 1865–1871.

Lönnstedt, I. and Speed, T., 2002. Replicated microarray data. *Statistica sinica*, 12(1), pp. 31–46.

Besag, J., 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302.

Brent, R. P., 1973. *Algorithms for minimization without derivatives*. Courier Corporation.

Shen-Orr, S. S. et al., 2010a. Cell type-specific gene expression differences in complex tissues. *Nat Meth*, 7(4), pp. 287–289.

Liebner, D. A., Huang, K., and Parvin, J. D., 2014. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, 30(5), pp. 682–689.

Tyekucheva, S. et al., 2017a. Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nature Communications*, 8(1), p. 420.

Cancer Genome Atlas Network, 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536), p. 576.

Yoshihara, K. et al., 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Shen-Orr, S. S. et al., 2010b. *Data accessible at NCBI GEO database; Accession GSE19830*. URL: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19830`.

Tyekucheva, S. et al., 2017b. *Data accessible at NCBI GEO database; Accession GSE97284*. URL: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97284`.

Genomic Data Commons Data Portal: TCGA Head and Neck Squamous Carcinoma. URL: `https://portal.gdc.cancer.gov/projects/TCGA-HNSC`.