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

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

Algorithm 2 Performing ICM for three-component

1: **Parameter:**
   Sample-wise $\{\pi_{1,i}, \pi_{2,i}\}_i : 2 \times S$
   Gene-wise $\{\mu_{Tg}, \sigma_{Tg}\}_g : 2 \times G$
2: **Initialize:**
   $\{\mu_{Tg}, \sigma_{Tg}\}_{g=1}^{G} = \mu_0, \sigma_0$
3: **for** iteration $t = 1, \cdots, T$ **do,**
4: a. update $\{\pi_{1,i}, \pi_{2,i}\}_{i=1}^{S}$
5: **for** each sample $i = 1, \cdots, S$ **do parallel**
6: update $\{\pi_{1,i}^{(t)}, \pi_{2,i}^{(t)}\} = argmax \prod_{g=1}^{G} f(y_{ig} | \{\pi_{1,i}, \pi_{2,i}\}, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^{G})$
7: **end for**
8: b. update $\{\mu_{Tg}, \sigma_{Tg}\}_{g=1}^{G}$
9: **for** each gene $g = 1, \cdots, G$ **do parallel**
10: update $\{\mu_{Tg}^{(t)}, \sigma_{Tg}^{(t)}\} = argmax \prod_{i=1}^{S} f(y_{ig} | \{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^{S}, \{\mu_{Tg}, \sigma_{Tg}\})$
11: **end for**
12: **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.

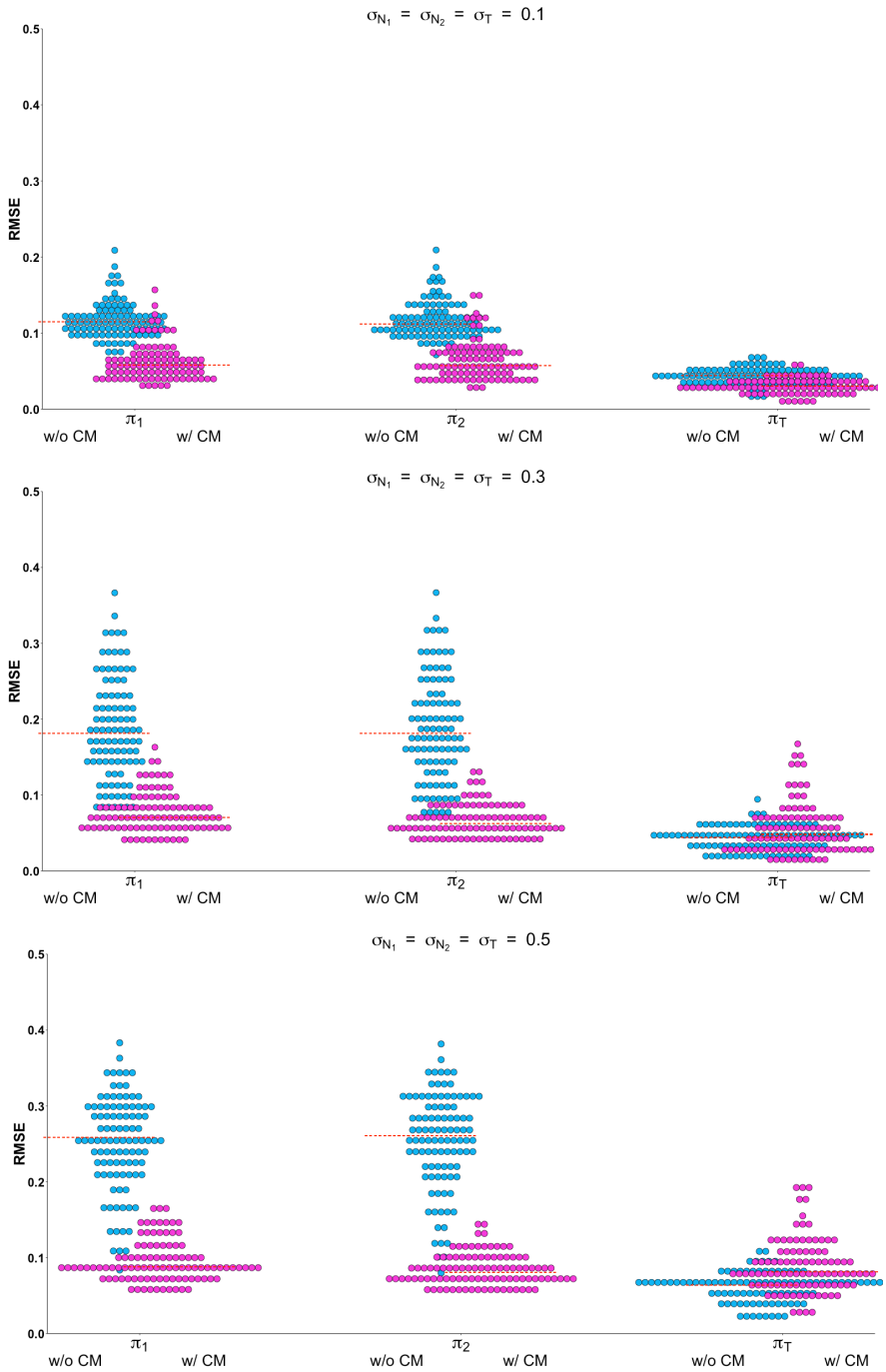

Figure S2. Dot plots of root mean square errors (RMSEs) between true and estimated proportions, using DeMixT with (w/) and without (w/o) component merging (CM), related to Figure 1.

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.

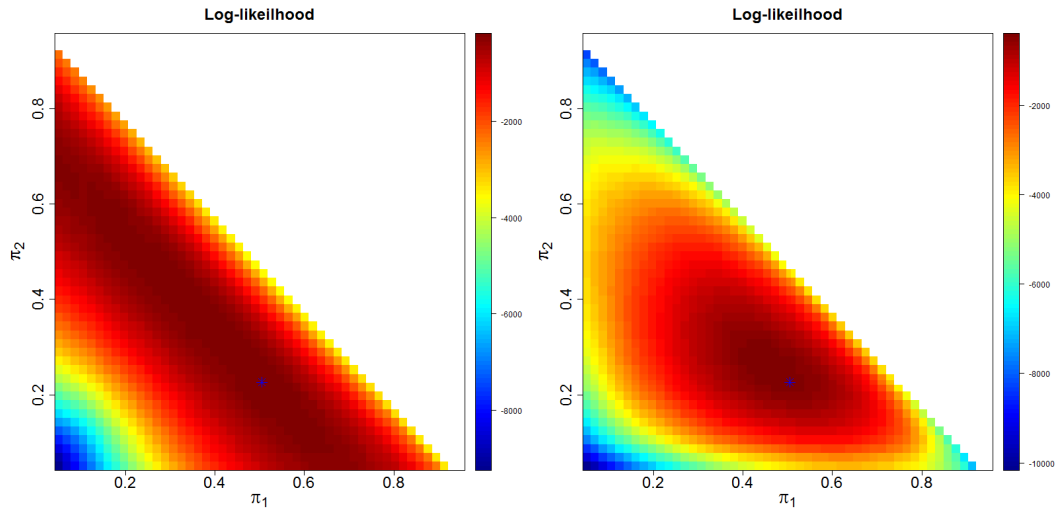

Figure S3. Log-Likelihood surface for π_1 and π_2, related to Figure 1.

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.

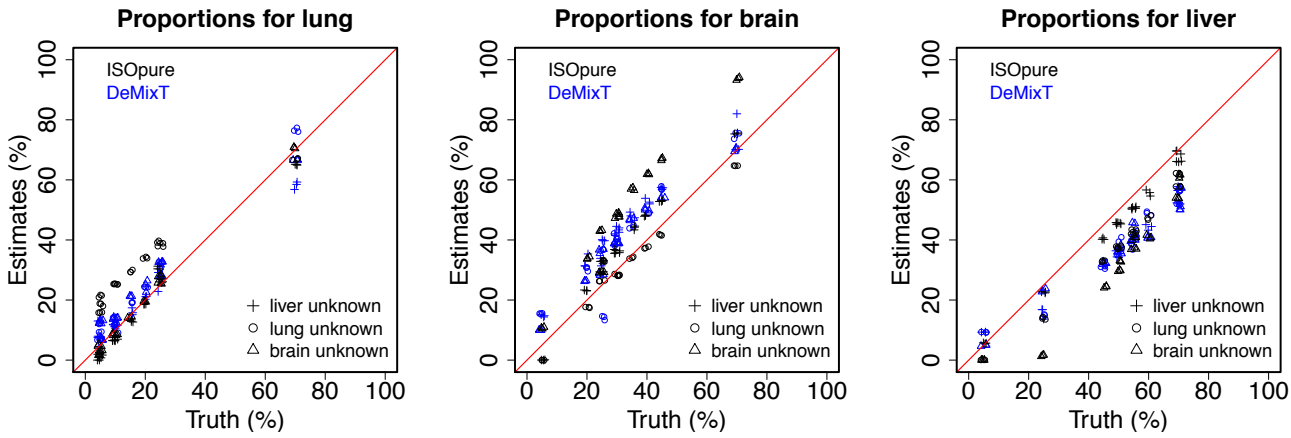

Figure S4. Scatter plots of estimated tissue proportions against true tissue proportions for the GSE19830 dataset, related to Figure 2.
All proportion estimates from running DeMixT are shown when either the liver, brain, or lung tissue is assumed to be the tissue with unknown expression profiles. Plus symbols: liver tissue is unknown; circles: lung tissue is unknown; triangles: brain tissue is unknown; blue: DeMixT; black: ISOpure.

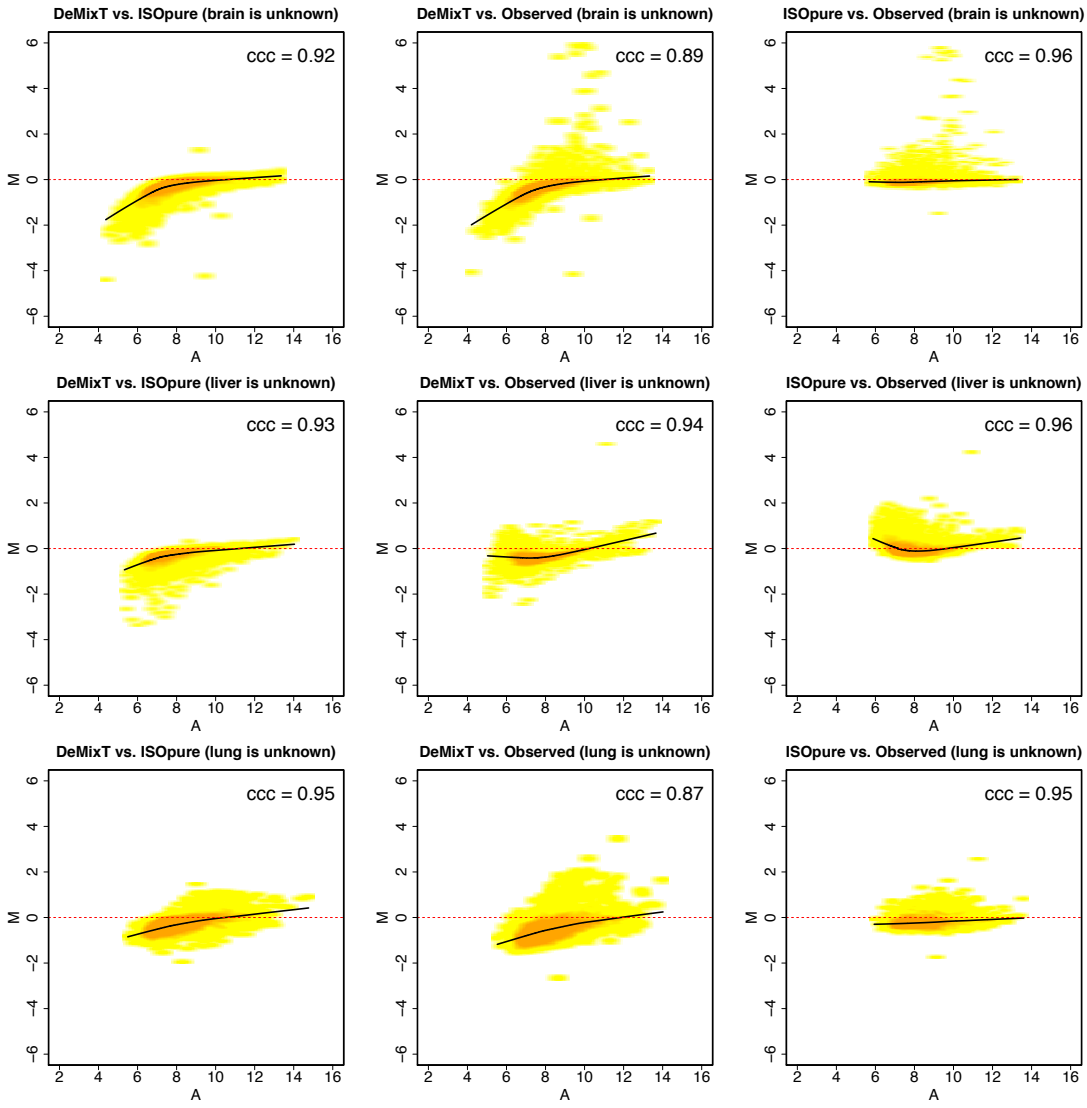

Figure S5: Smoothed scatter MA plots of mean estimated tissue-specific expression (at the log2 scale) from DeMixT and ISOpure in the GSE19830 dataset, related to Figure 2.

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.

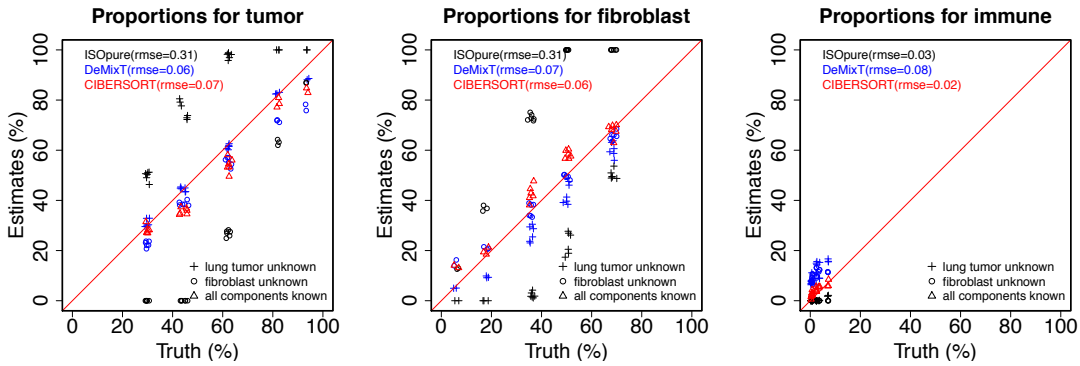

Figure S6. Scatter plots of estimated versus true proportions for the mixed cell line RNA-seq dataset, related to Figure 2.

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.

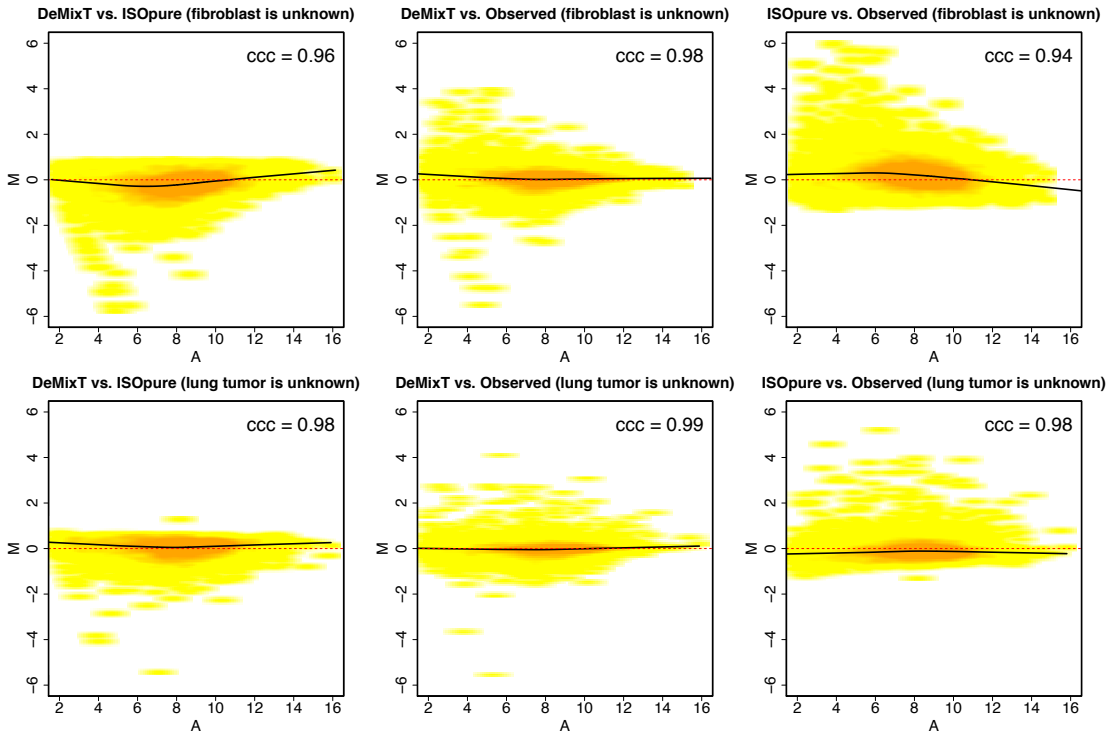

Figure S7. Smoothed scatter MA plots of mean estimated tissue-specific expression levels (at the log2 scale) from DeMixT and ISOpure in the mixed cell line RNA-seq dataset, related to Figure 2.

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.

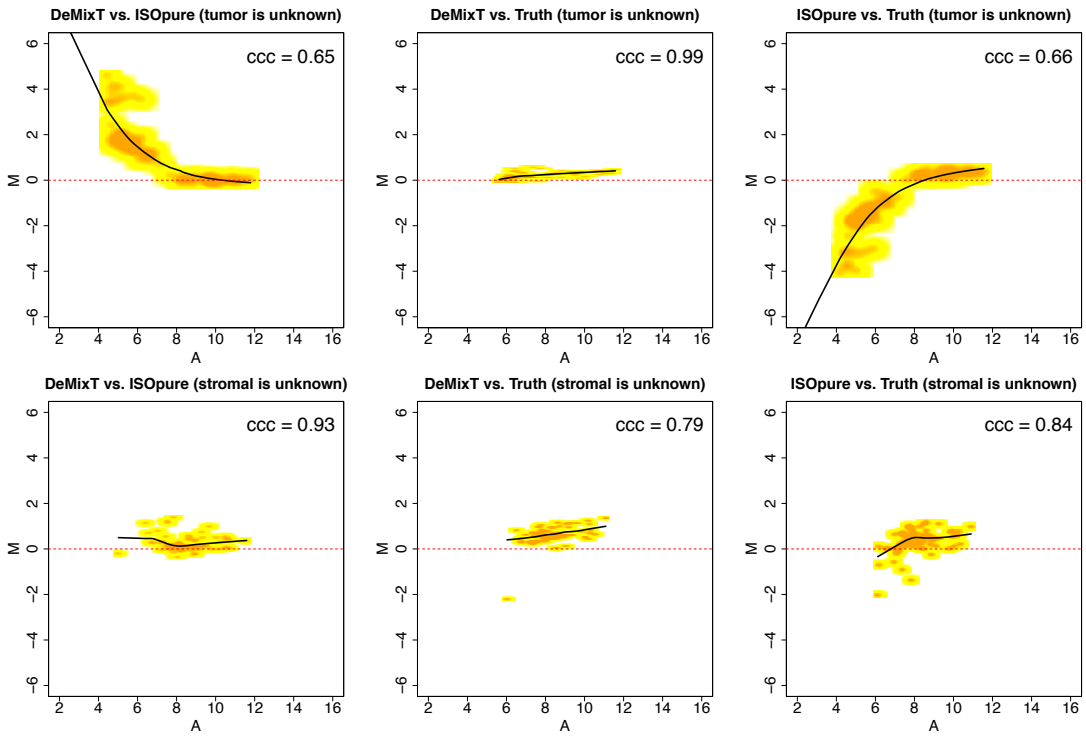

Figure S8: Smoothed scatter MA plots of mean estimated tissue-specific expression (at the log2 scale) between DeMixT and ISOpure in the LCM FFPE prostate cancer microarray dataset, related to Figure 3.
The MA plots compare mean values of log2-transformed deconvolved expression across genes for DeMixT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

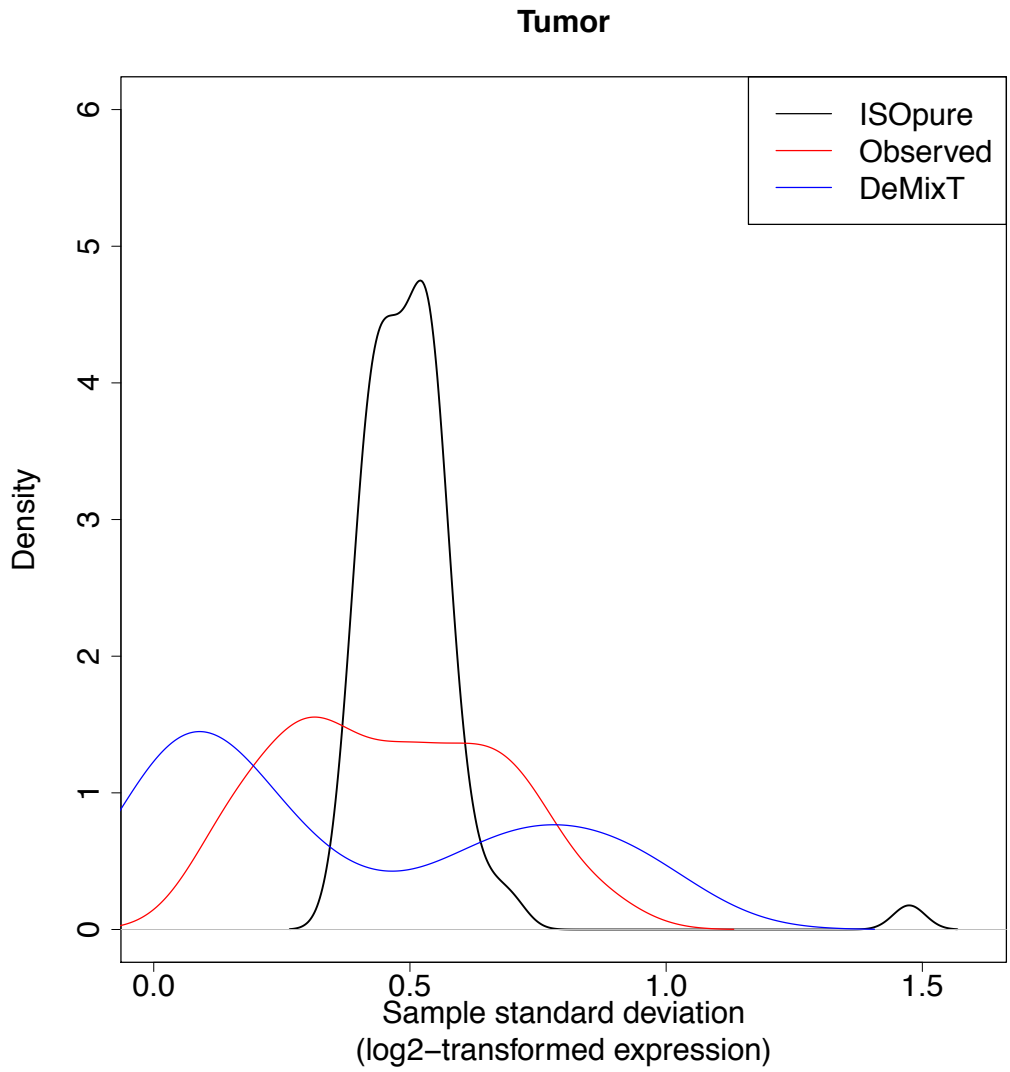

Figure S9. Density plot comparing sample standard deviations between deconvolved expression profiles of subset probes for DeMixT and ISOpure in the LCM FFPE prostate cancer microarray dataset when tumor tissue was assumed to be the unknown component; with measured expression profiles of isolated tumor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

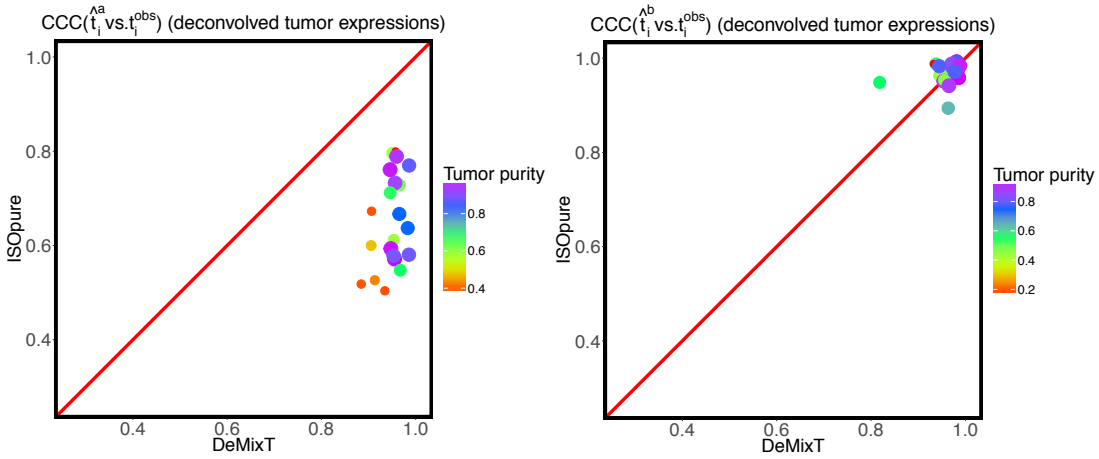

Figure S10. Scatter plots of concordance correlation coefficient (CCC) between individual deconvolved expressions and observed values for the tumor component in 23 LCM prostate samples, related to Figure 3. Each point corresponds to a sample. We compared results from ISOpure with those from DeMixT. Left 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.

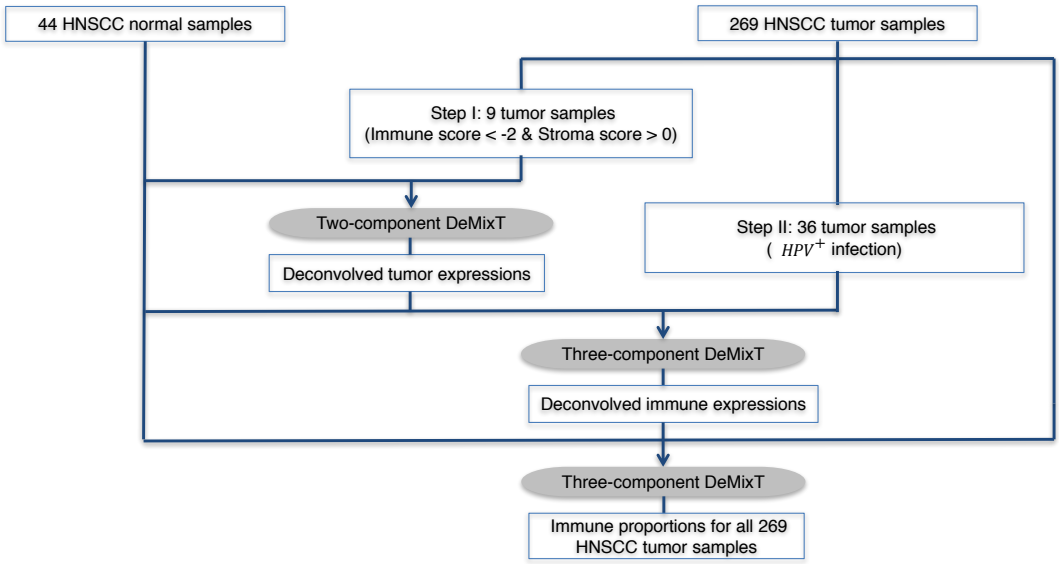

Figure S11. Workflow for analysis of immune infiltration in the HNSCC dataset, related to Figure 4.

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

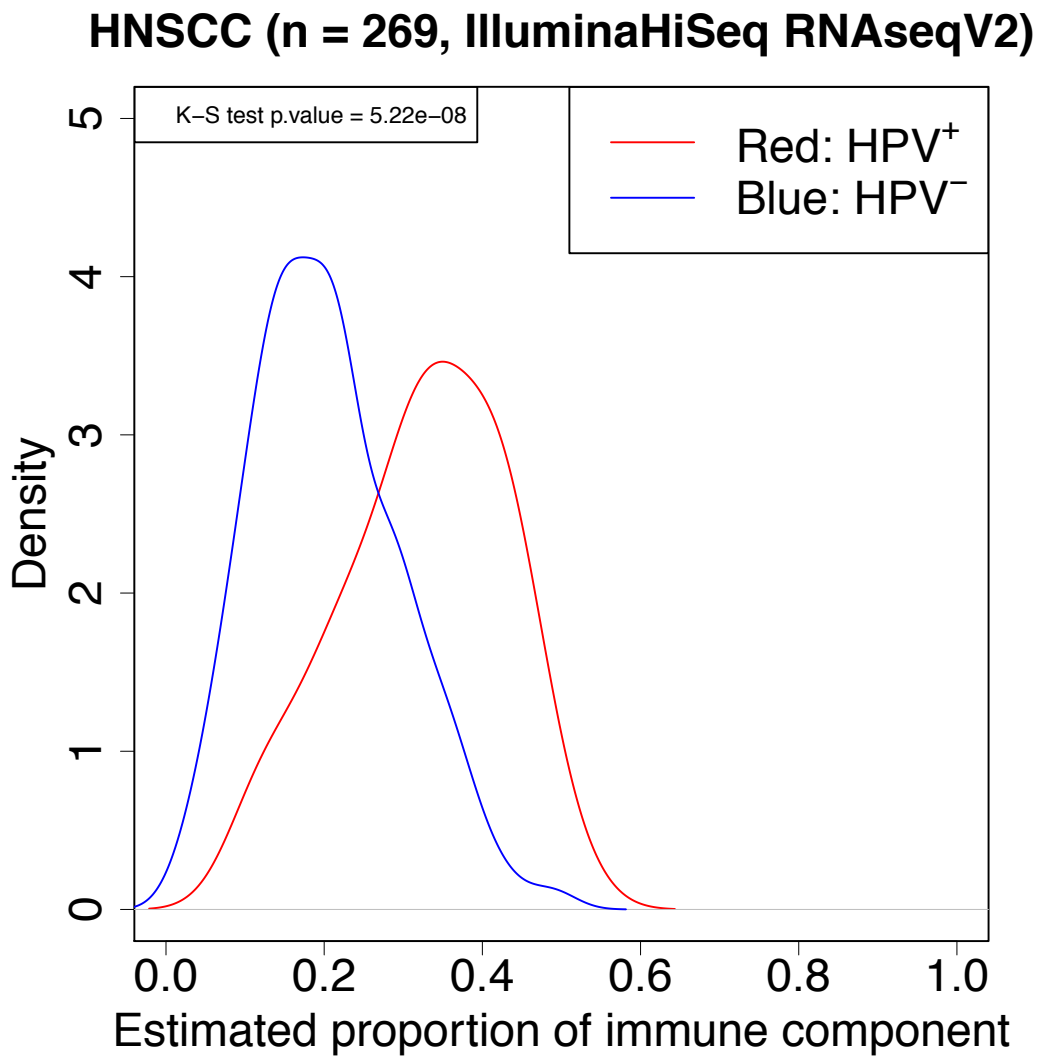

Figure S12. Density plot of estimated immune proportions for tumor samples with HPV test results, related to Figure 4.
Red curve: for those with HPV+ status; blue curve: for those with HPV-.

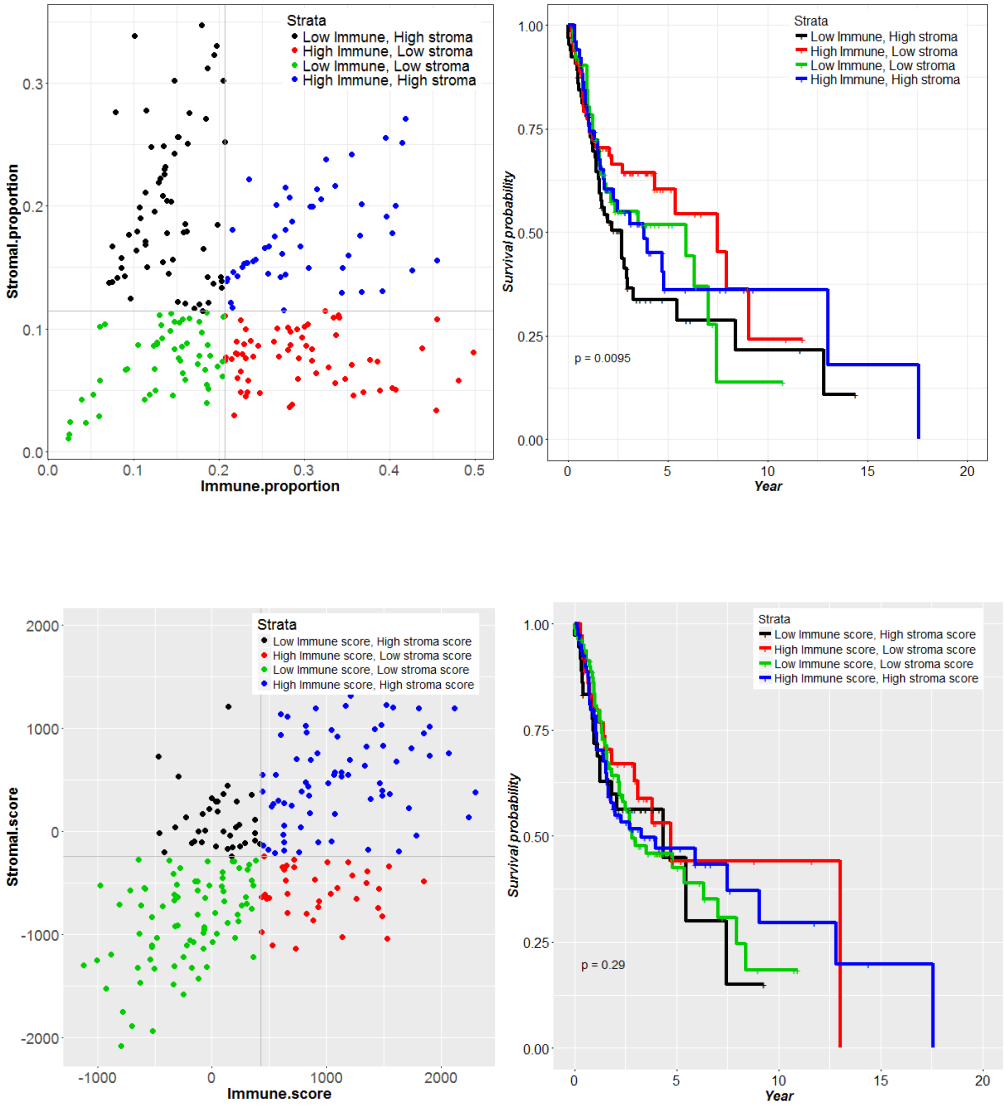

Figure S13. Association of immune-stroma-proportions from DeMixT with overall survival compared with association of immune-stroma-scores from ESTIMATE with overall survival in HNSCC, related to Figure 4.

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.

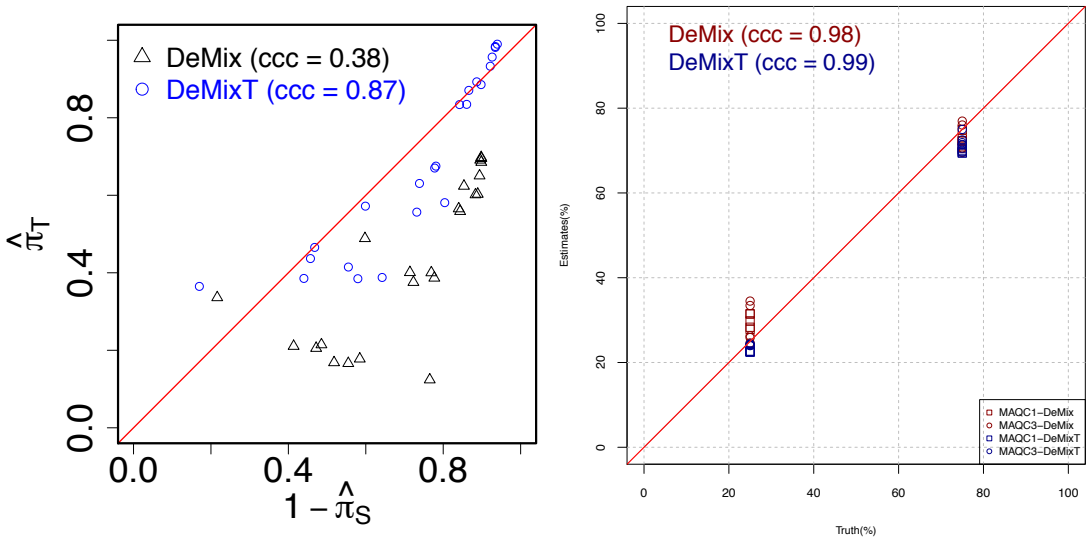

Figure S14. Comparison of proportion estimation between DeMixT and DeMix, related to Figure 1 and 2.

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.

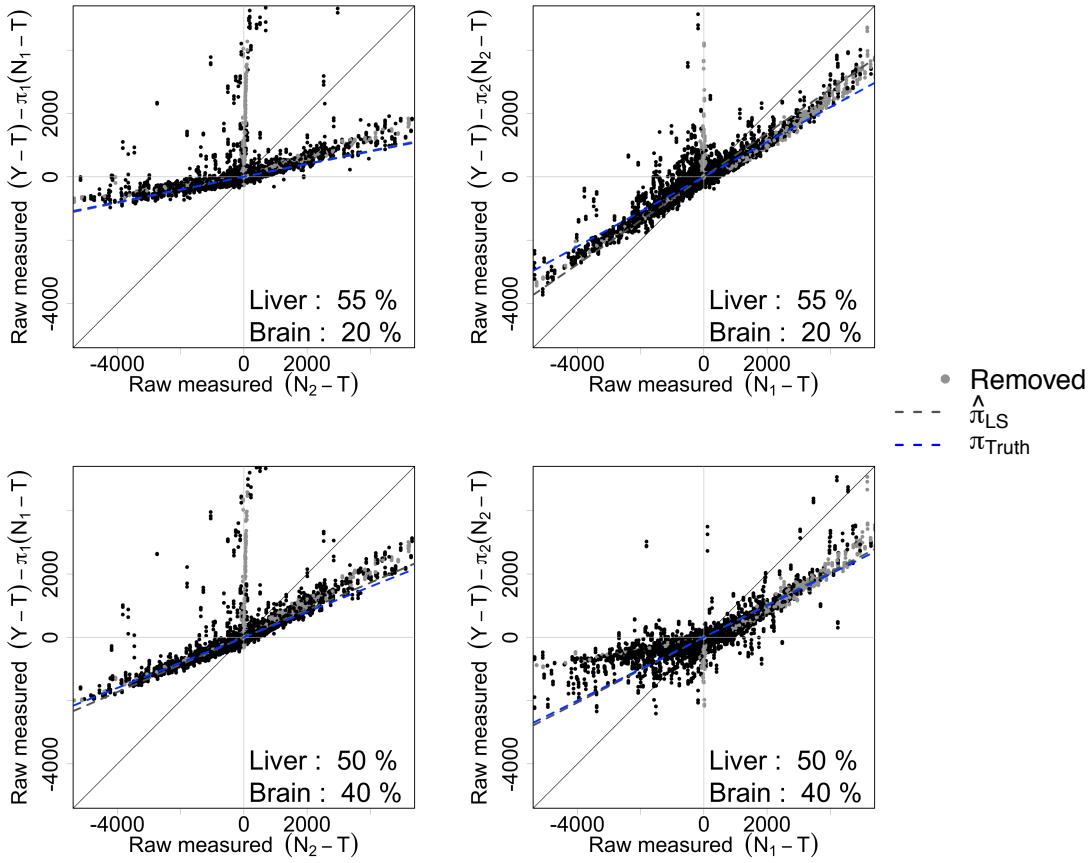

Figure S15. Scatter plots of $Y_{ig} - \overline{T}_g - \pi_{2,i}(\overline{N}_{2,g} - \overline{T}_g)$ versus $\overline{N}_{1,g} - \overline{T}_g$ for $\pi_{1,i}$ and $Y_{ig} - \overline{T}_g - \pi_{1,i}(\overline{N}_{1,g} - \overline{T}_g)$ versus $\overline{N}_{2,g} - \overline{T}_g$ for $\pi_{2,i}$ 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 N_1 or N_2; 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}

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

   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 $\hat{\pi}_{1,i} n_{1,ig} + \hat{\pi}_{2,i} n_{2,ig} \leq y_{ig}$. 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^2_{Tg} 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

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

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.