Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration

Zeya Wang[1, 2], Jeffrey S. Morris[3], Shaolong Cao[2], Jaeil Ahn[4], Rongjie Liu[3], Svitlana Tyekucheva[5, 11], 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] & Wenyi Wang[2]

[1]Department of Statistics, Rice University, Houston, TX; [2]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX; [3]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX; [4]Department of Biostatistics and Bioinformatics, Georgetown University, Washington, DC; [5]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA; [6]Department of Statistics, Harvard University, Cambridge, MA; [7]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX; [8]Center for Molecular Oncologic Pathology, Dana Farber Cancer Institute, Boston, MA; [9]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; [10] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [11] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; [12]Department of Statistics, University of Oxford, Oxford, United Kingdom. Correspondence should be addressed to W.W. (wwang7@mdanderson.org).

ABSTRACT: 62 words

MAIN TEXT: 1509 words

ABSTRACT

We develop a novel method DeMixT for the gene expression deconvolution of three compartments in cancer patient samples: tumor, immune and surrounding stromal cells. In validation studies using mixed cell line and laser-capture microdissection data, DeMixT yielded accurate estimates for both cell proportions and compartment-specific expression profiles. Application to the head and neck cancer data shows DeMixT-based deconvolution provides an important step to link tumor transcriptome data with clinical outcomes.

MAIN TEXT

Heterogeneity of malignant tumor cells adds confounding complexity to cancer treatment. The evaluation of individual components of tumor samples is complicated by the tumor-stromal-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[3,4]. Further research supports the association of infiltrating immune cells with clinical outcome for individuals with ovarian cancer, colorectal cancer and follicular lymphoma[5-7]. The use of experimental approaches such as laser micro-dissection and cell sorting is limited by the associated expense and time. Therefore, understanding the heterogeneity of tumor cells motivates a computational approach to integrate the estimation of type-specific expression profiles in tumor cells, immune cells and microenvironment. Most commonly available deconvolution methods assume that malignant tumor cells consist of two distinct components, epithelium-derived tumor and surrounding stromal cells[1,2]. Other deconvolution methods for more than two compartments require knowledge of cell-type-specific gene lists[8], i.e. reference genes, with some of these methods focused on application in estimating subtype proportions within immune cells[9-11]). Therefore there is still a need for methods that can provide joint estimation of proportions and compartment-specific gene expression for more than two compartments in each tumor sample. We have developed a new statistical approach, DeMixT, to address this need (**Fig.1a,** R package freely downloadable at https://github.com/wwylab/DeMix).

Previously developed ISOpure[12] may also 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. One 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) in genes where the noncancerous compartments actually bears a substantial variance, not accounting for it can result in large bias in estimated mixing proportions and mean expressions. Our proposed DeMixT approach explicitly models variance for each compartment in order to fulfill our comprehensive goal in deconvolution.

Here, we summarize DeMixT as follows (details in Online 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 observed mixed tumor samples, $N_{1,ig}$, $N_{2,ig}$ and T_{ig} represent unobserved raw expression values from its constituents. We assume$N_{1,ig}$, $N_{2,ig}$ and T_{ig}follow a $\log_2$-normal distribution with compartment-specific means and variances[1,13]. We call the first two components as N_1-component and N_2-component, 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*. We define the last component as T-component, whose distribution is unknown. In practice, the T-component can be any of the three: tumor, stroma 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[1] for a two-component model now becomes inefficient for a three-component model space which is much more complex: $\{\pi_{1,i},\pi_{2,i}\}_{i=1}^{S}, \{\mu_{Tg},\sigma_{Tg}\}_{g=1}^{G}$. We have implemented an optimization approach called iterative conditional modes[14] that cyclically maximizes the probability of each set of variables conditional on the rest, for which we have observed rapid convergence[14]. We further developed a novel two-stage approach to extract reliable expression measurements and improve estimation performance of the vector of proportions, which would then further improve the estimation of means and variances (See Online Methods for details).

We first validated DeMixT in two datasets with known truth in proportions and mean expressions (see **Online Methods**): a publicly available microarray dataset[15] generated using mixed RNA from rat brain, liver and lung tissues in varying proportions **(Supplementary Table**

1); and an RNA-seq dataset generated using mixed RNAs from three cell lines, lung adenocarcinoma (H1092), cancer-associated fibroblasts (CAF) and tumor infiltrating lymphocytes (TIL) (**Supplementary Table 2**). We assessed our approach through a number of statistics, e.g. concordance correlation coefficients, root mean square errors and a summary statistics for measuring reproducibility (**Online Methods**), we showed that DeMixT performed well and outperformed ISOpure in terms of accuracy and stability for estimation irrespective of which component is treated as unknown (**Fig. 1b-c, See Supplementary Information for further details, Supplementary Fig. 1-4, Supplementary Tables 3-7**).

Next, we applied DeMixT to a gold-standard validation dataset from real tumor, which has known proportions, mean expressions and individual component-specific expressions. Laser-capture microdissection (LCM) was performed on Formalin Fixed Paraffin Embedded (FFPE) tissue samples from 23 prostate cancer patients and generate microarray gene expression data using the derived, and the matching dissected stromal and tumor tissues (GSE97284, private link available to reviewers). Due to quality of FFPE samples, we selected a subset of probes (**Online methods**), and ran DeMixT under a two-component mode. DeMixT obtained concordant estimates of tumor proportion for when stroma is unknown and when tumor is unknown (CCC=0.87) (**Fig. 2a**). DeMixT also tended to provide accurate component-specific mean expression levels (**Figs. 2b**, **2c** and **Supplementary Fig. 5**) and yielded standard deviation estimates that are close to those from the dissected tumor samples **(Supplementary Fig. 6)**. As a result, the DeMixT individually deconvolved expressions achieved high CCCs (mean= 0.96) for the tumor component **(Figs. 2d** and **Supplementary Fig. 7).** The expressions for the stromal component here are more variable than a common gene expression dataset hence both DeMixT and ISOpure gave slightly biased estimates on means and standard deviations.

A recent study showed with head and neck squamous cell carcinoma (HNSCC) the infiltration of immune cells, both lymphocytes and myelocytes, is positively associated with viral infection in virus-associated tumors[10]. We downloaded HNSCC RNA-seq data from the TCGA data portal[16] and ran DeMixT for deconvolution. Since only reference samples for the stromal component are available from TCGA (i.e., 44 normal samples and 269 tumor samples), we devised an analysis pipeline for DeMixT to run successfully on the HNSCC samples (See online methods for details, **Supplementary Fig. 8**). Briefly we first used data from the HPV+ tumors to derive reference samples for the immune component, and then ran the three-compoment DeMixT on the entire dataset to estimate proportions for both HPV- and HPV+ samples. For all tumor samples, we obtained the immune (mean = 0.22, sd = 0.10), the tumor (mean = 0.64, sd = 0.13), and the stromal proportions (mean = 0.14, sd = 0.07, see **Supplementary Figure 9**). As expected, tumor samples with HPV+ had significantly higher immune proportions than those tested as HPV- [10,17] (P = 2e-8, **Fig. 2e** and **Supplementary Fig. 9-10**). To further evaluate the performance of our deconvolved expression levels, we performed differential expression tests for immune versus stromal, and immune versus tumor, respectively, on 63 infiltrating immune cell-related genes (CD and HLA genes). For example, **Fig. 2f** illustrates the deconvolved expressions were much higher in the immune component than the other two for three important immune marker genes CD4, CD14, HLA-DOB. Overall, 51 out of 63 genes were significantly more highly expressed in immune than the other two components **(adjusted p-values are listed in Supplementary Table 8**, also see **Supplementary Fig. 11)**.

In this work, we have presented a novel statistical method and software, DeMixT (R package at https://github.com/wwylab/DeMix, Docker container at https://cloud.docker.com/app/rj2016/repository/docker/rj2016/demix/general), for dissecting a mixture of tumor, stroma and immune cells 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. Our input data is distinct from those of other deconvolution methods: gene expressions from 1) observed mixture tumor samples and 2) a set of reference samples from *p*-1 compartments (*p* is the total number of compartments). Our output data is unique as we further provide gene- and compartment-specific expression levels for each tumor sample, essentially allowing for all previously developed downstream analyses pipelines, such as clustering and feature selection in cancer biomarker studies, still applicable to the deconvolved gene expressions. We achieved this unique output by modeling compartment-specific variance and addressing the associated inference challenges. Our method is extendable to more than three components.

The reference gene-based deconvolution is popular for estimating immune subtypes within immune cells[8,11]. We do not require reference genes which we consider as difficult to find for the tumor component, although DeMixT can take reference genes when available. With the reference sample approach, we assume available normal samples to be representative of the non-tumor component in the mixture samples. This assumption is often violated, although we found such violation may not affect the deconvolution results of DeMixT. The reference samples can be derived from historical patient data or from other healthy individuals, such as data from GTEx[18] (unpublished results). Furthermore, each of the three components may contain more than one type of cells, in particular, the immune component. It was reported that although heterogeneous, the relative proportions of immune subtypes within the immune compartment is consistent across patient samples[19], supporting us to model the pooled immune cell population using one distribution. Finally, the performance of DeMixT will be optimized when the data analysis practice is linked with the cancer-specific biological knowledge.

In conclusion, DeMixT helps to resolve the bottleneck arising from sample heterogeneity in cancer genomic studies.

ACKNOWLEDGEMENS

We would like to thank Vésteinn Þórsson, Ilya Shmulevich, David Gibbs, Liuqing Yang and Hongtu Zhu for useful discussions and valuable suggestions on data analysis.

References:

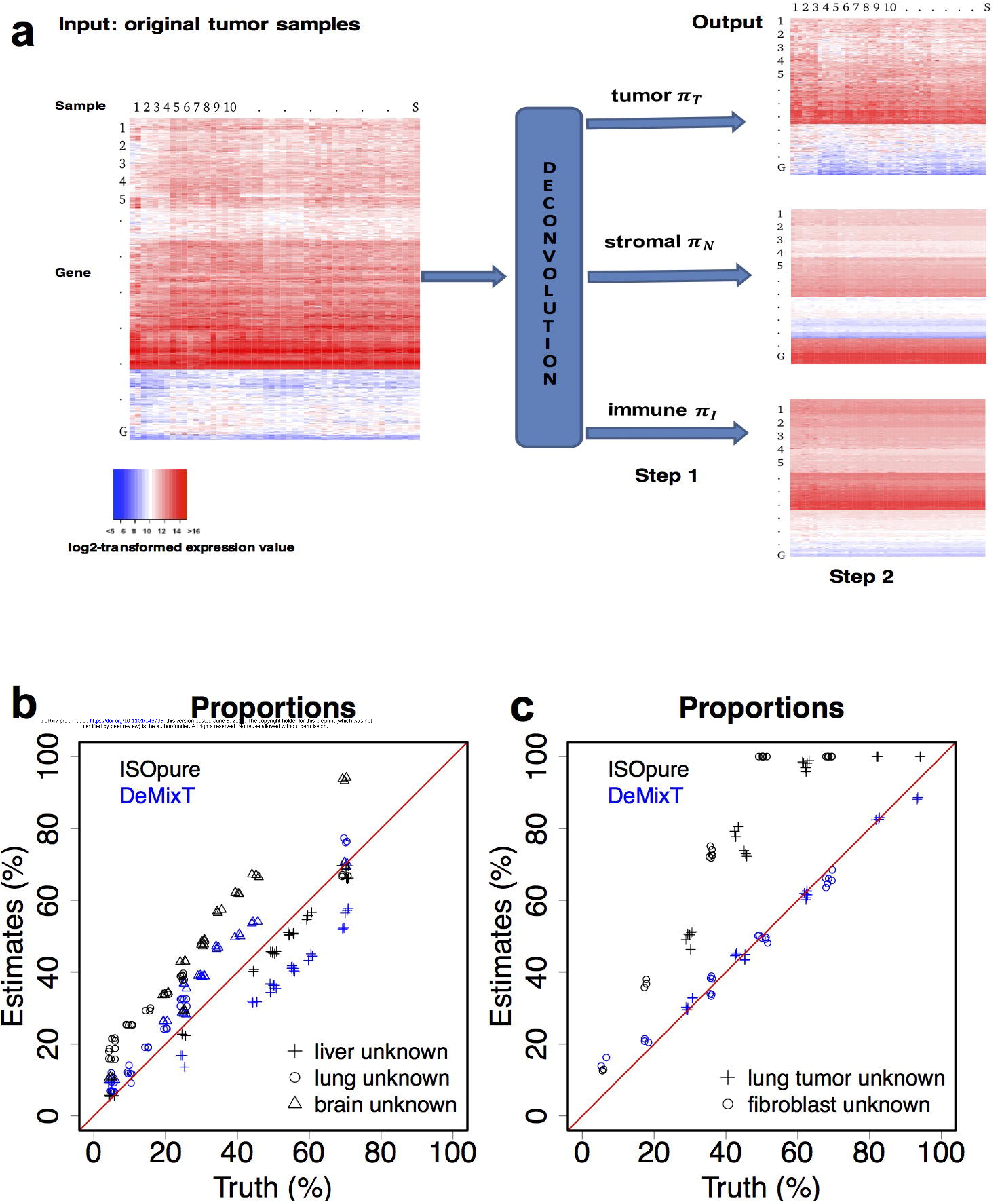
1. Ahn, J., *et al.* DeMix: Deconvolution for Mixed Cancer Transcriptomes Using Raw Measured Data. *Bioinformatics* **29**, 1865-1871 (2013).
2. Gong, T. & Szustakowski, J.D. DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq data. *Bioinformatics* **29**, 1083-1085 (2013).
3. Pages, F., *et al.* Immune infiltration in human tumors: a prognostic factor that should not be ignored. *Oncogene* **29**, 1093-1102 (2009).
4. Fridman, W.H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* **12**, 298-306 (2012).
5. Dave, S.S., *et al.* Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells. *New England Journal of Medicine* **351**, 2159-2169 (2004).
6. Galon, J., *et al.* Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science* **313**, 1960-1964 (2006).

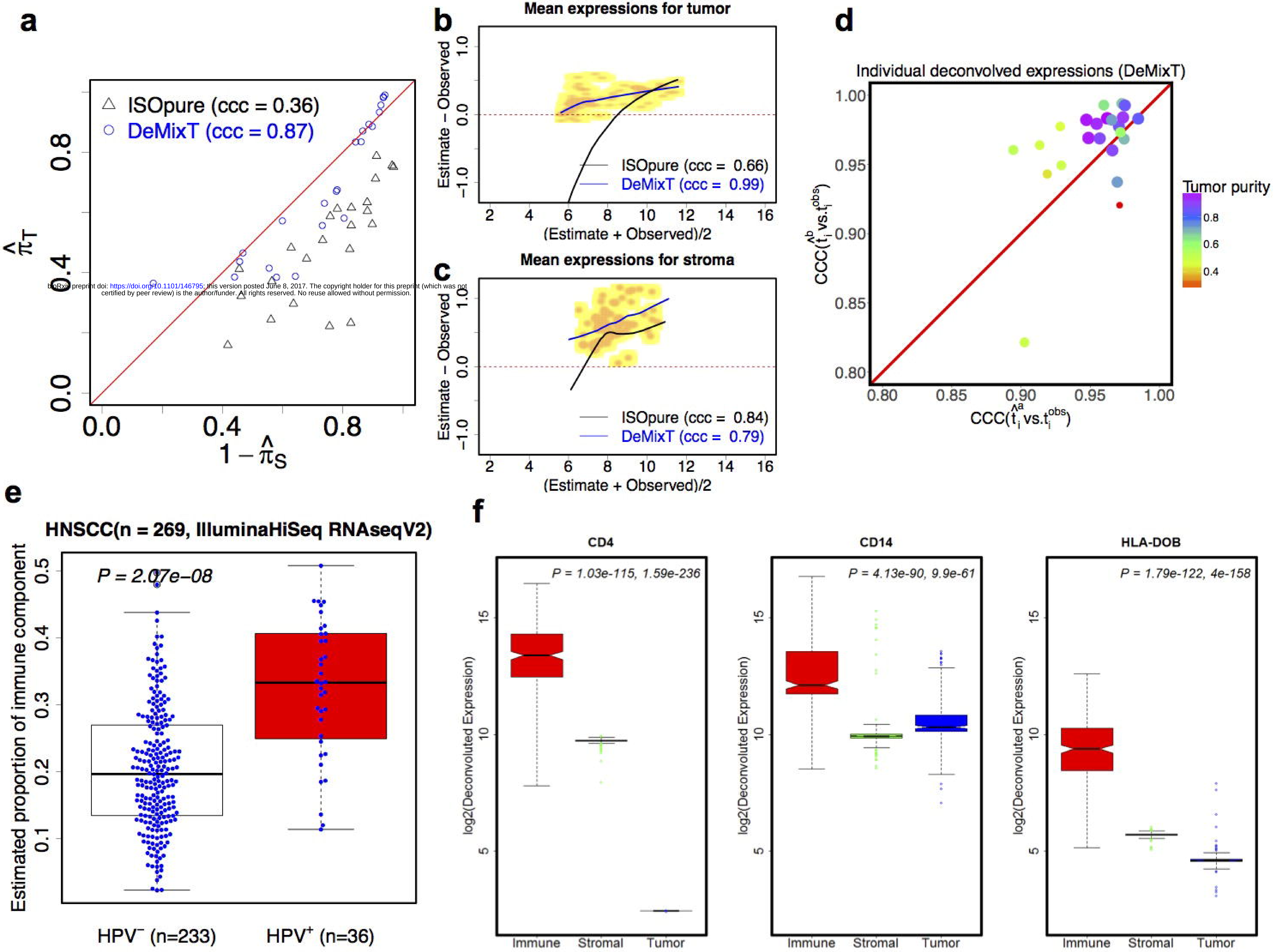
7. Zhang, L., *et al.* Intratumoral T Cells, Recurrence, and Survival in Epithelial Ovarian Cancer. *New England Journal of Medicine* **348**, 203-213 (2003).
8. Liebner, D.A., Huang, K. & Parvin, J.D. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics* **30**, 682-689 (2014).
9. Li, B., *et al.* Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nature genetics* **48**, 725-732 (2016).
10. Li, B., *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biology* **17**, 174 (2016).
11. Newman, A.M., *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457 (2015).
12. Quon, G., *et al.* Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med* **5**, 29 (2013).
13. Lönnstedt, I. & Speed, T. Replicated microarray data. *Statistica sinica* **12**, 31-46 (2002).
14. Besag, J. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 259-302 (1986).
15. Shen-Orr, S.S., *et al.* Cell type-specific gene expression differences in complex tissues. *Nat Meth* **7**, 287-289 (2010).
16. Network, C.G.A. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582 (2015).
17. Fakhry, C., *et al.* Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial. *Journal of the National Cancer Institute* **100**, 261-269 (2008).
18. Lonsdale, J., *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45**, 580-585 (2013).
19. Gentles, A.J., *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature medicine* **21**, 938-945 (2015).

Figure legends

Fig. 1. Description of DeMixT and validation results using microarray and RNA-seq data from tissue and cell line mixture experiments. (a) An illustration of three-component deconvolution to output tissue-specific proportion, and isolated expression matrices of tumor, stromal and immune cells. Heat map of expression levels uncovers the difference in gene expression patterns between original tumor samples, deconvolved tumor components, stromal components and immune components. (b) A scatter plot of estimated tissue proportions against the truths when the liver (cross), brain (triangle), or lung (rectangle) tissue is assumed to be the unknown tissue in the microarray experiments mixing rat liver, brain, and lung tissues; estimates from ISOpure are also presented. (c) Scatter plot of estimated tissue proportions against the truth when either lung tumor (cross) or fibroblast (rectangle) cell lines are assumed to be the unknown tissue in the RNA-seq experiments mixing lung tumor, fibroblast and lymphocyte cell lines.

Fig. 2. Analyses of real data using DeMixT. Validation using LCM data in prostate cancer (a-c) and application to TCGA data in head and neck cancer (d-e). (a) Scatter plot of estimated tumor proportions versus 1- estimated stromal proportions; estimates from DeMixT (blue) are compared with those from ISOpure (black). (b)-(c) Smoothed scatter MA plots between observed and deconvolved mean expression values from DeMixT for the tumor and stromal components, respectively (yellow for low values and orange for high values). The lowess smoothed curves for DeMixT is shown in black and ISOpure in blue. (d) Scatter plot of concordance correlation (CCC) between individual deconvolved expression profiles (tihat)and observed values (tiobs) for 23 LCM matching prostate samples. We use superscript *a* to denote the scenario when the stromal component are reference samples; *b* to denote the scenario when the tumor component are reference samples.. The color gradient of each points corresponds to the estimated tumor proportion.. (e) Boxplots of estimated immune proportions for HNSCC samples in the test set display differences between HPV-positive (red) and HPV-negative (white) samples. (f) Box plots of log2-transformed deconvolved expression profiles for three important immune genes (CD4, CD14, HLA-DOB) in the test set of HNSCC samples. Red stands for the immune component; green stands for the stromal component; and blue stands for the tumor component. P-values of differential tests are given in the top right corner for each gene: the first p-value is for immune vs. stromal; and the second p-value is for immune vs. tumor.


a
Input: original tumor samples
Sample
1 2 3 4 5 6 7 8 9 10 . . . . S
1
2
3
4
5
Gene
G
<5 6 8 10 12 14 >16
log2-transformed expression value
DECONVOLUTION
tumor π_T
stromal π_N
immune π_I
Step 1
Output
1 2 3 4 5 6 7 8 9 10 . . . . S
Step 2
b
Proportions
ISOpure
DeMixT
liver unknown
lung unknown
brain unknown
Estimates (%)
Truth (%)
0
20
40
60
80
100
c
Proportions
ISOpure
DeMixT
lung tumor unknown
fibroblast unknown
Estimates (%)
Truth (%)


a
ISOpure (ccc = 0.36)
DeMixT (ccc = 0.87)
0.0
0.4
0.8
$\hat{\pi}_T$
$1 - \hat{\pi}_S$
b
Mean expressions for tumor
Estimate – Observed
(Estimate + Observed)/2
ISOpure (ccc = 0.66)
DeMixT (ccc = 0.99)
-1.0
0.0
1.0
2
4
6
8
10
12
14
16
c
Mean expressions for stroma
Estimate – Observed
(Estimate + Observed)/2
ISOpure (ccc = 0.84)
DeMixT (ccc = 0.79)
d
Individual deconvolved expressions (DeMixT)
$CCC(\hat{t}^b_i \text{ vs. } t^{obs}_i)$
$CCC(\hat{t}^a_i \text{ vs. } t^{obs}_i)$
0.80
0.85
0.90
0.95
1.00
Tumor purity
0.8
0.6
0.4
e
HNSCC(n = 269, IlluminaHiSeq RNAseqV2)
P = 2.07e−08
Estimated proportion of immune component
0.1
0.2
0.3
0.4
0.5
HPV^- (n=233)
HPV^+ (n=36)
f
CD4
P = 1.03e-115, 1.59e-236
CD14
P = 4.13e-90, 9.9e-61
HLA-DOB
P = 1.79e-122, 4e-158
log2(Deconvoluted Expression)
5
10
15
Immune
Stromal
Tumor