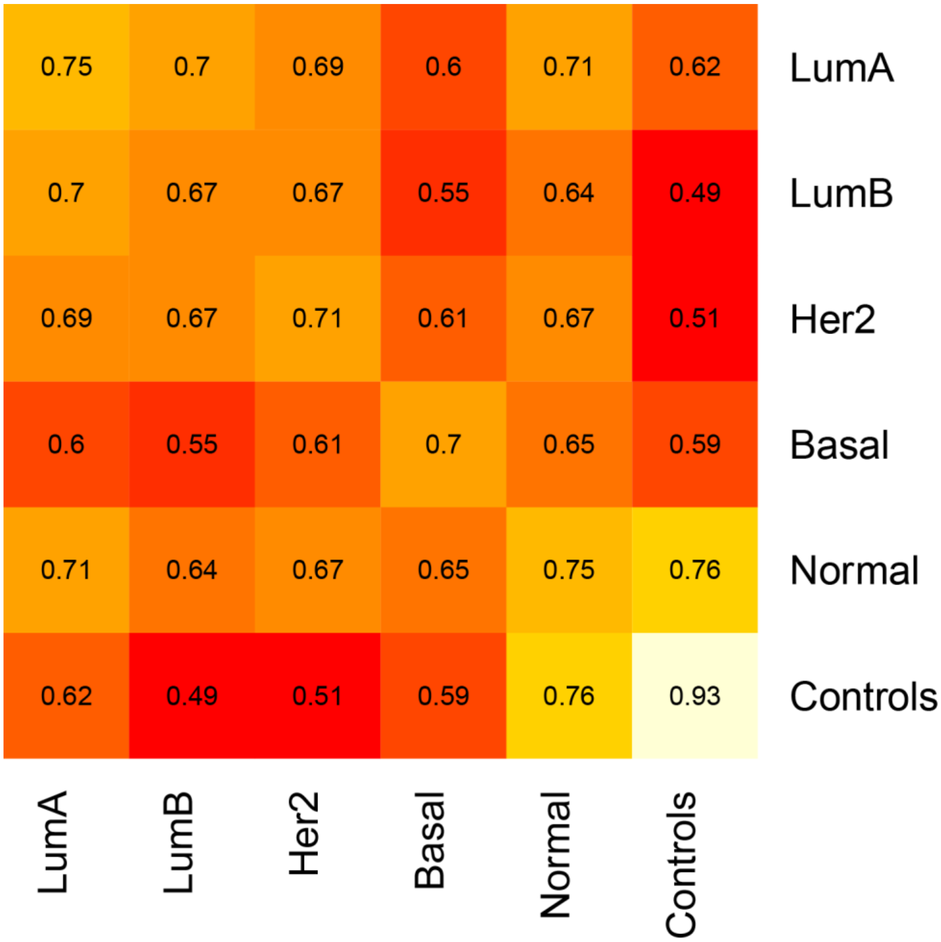
Supplemental Figures



	LumA	LumB	Her2	Basal	Normal	Controls
LumA	0.75	0.7	0.69	0.6	0.71	0.62
LumB	0.7	0.67	0.67	0.55	0.64	0.49
Her2	0.69	0.67	0.71	0.61	0.67	0.51
Basal	0.6	0.55	0.61	0.7	0.65	0.59
Normal	0.71	0.64	0.67	0.65	0.75	0.76
Controls	0.62	0.49	0.51	0.59	0.76	0.93

Figure S1. Related to Figure 3. Heterogeneity of methylation profiles of epithelial fraction within and between breast cancer subtypes. Values represent average spearman correlation between methylation profiles of epithelial fraction of all pairs of samples from the subtypes represented in the corresponding row and column.

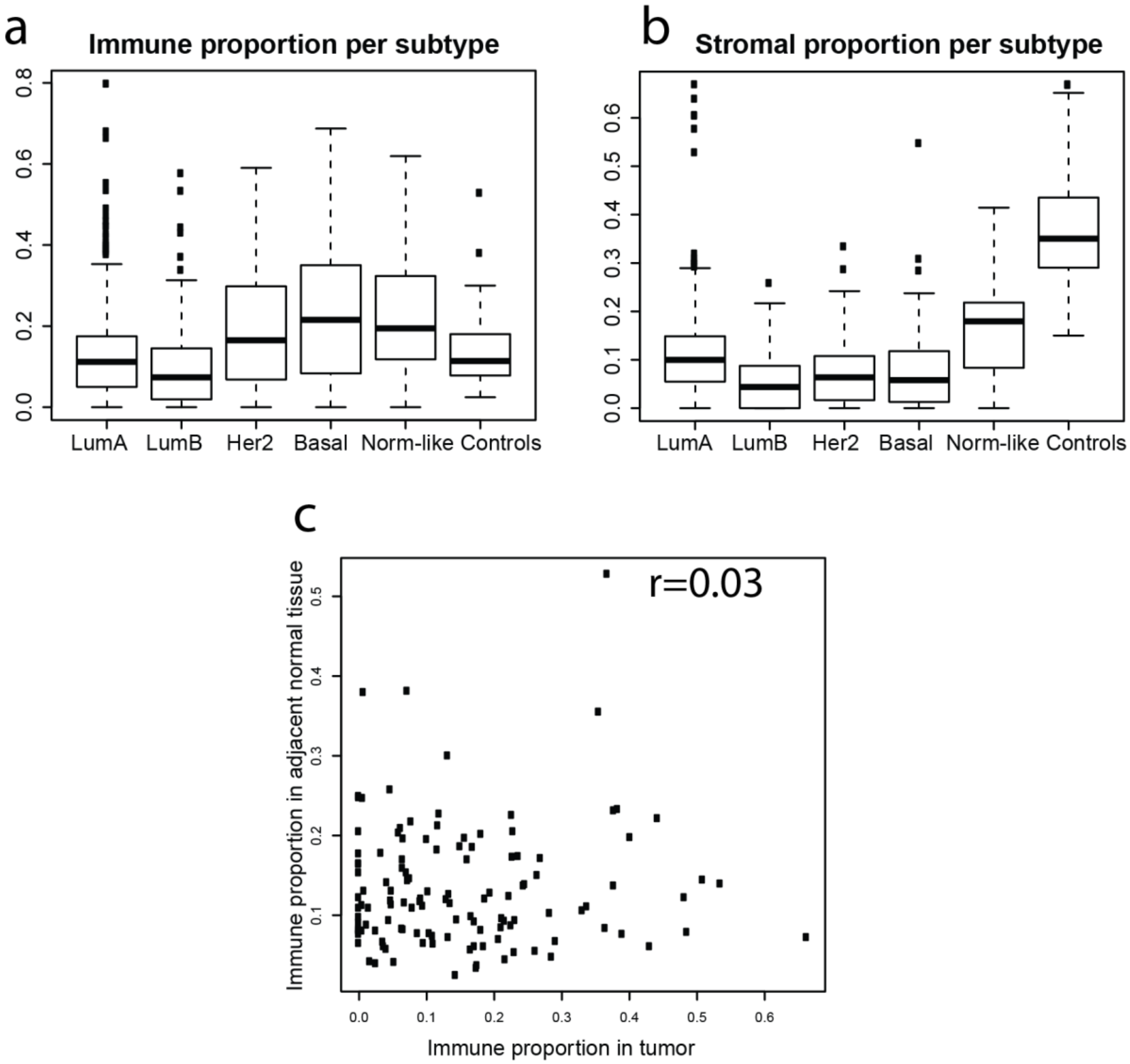


Figure S2. Related to Figure 3. Distribution of proportions of immune and stromal cell types across TCGA samples. a. Box plots of proportions of immune cells across cancer subtypes. **b.** Box plots of proportions of stromal cells across cancer subtypes. **c.** Scatterplot comparing the level of immune infiltration in tumor versus matched adjacent normal tissue samples.

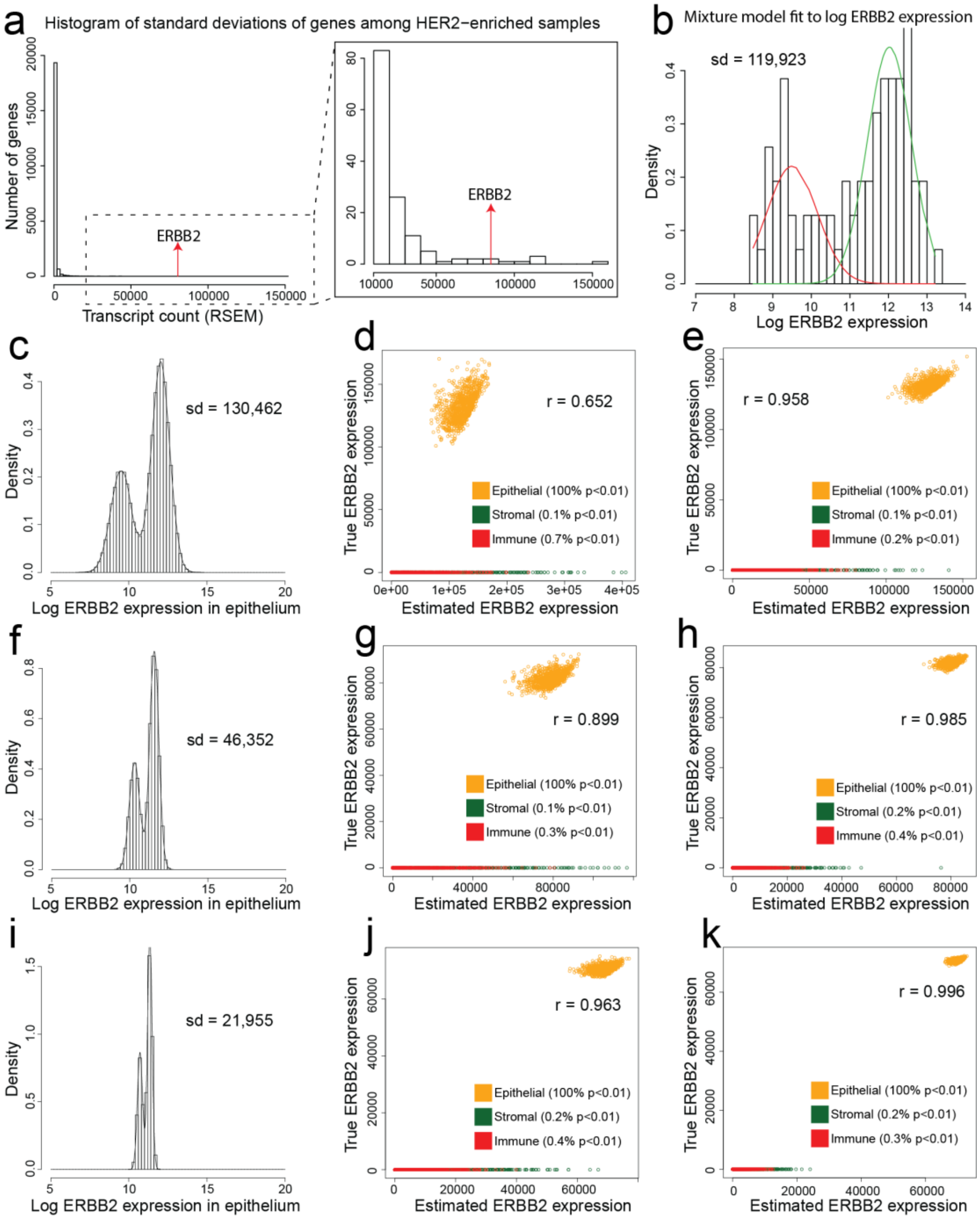


Figure S3. Related to Figure 4. Analysis of ERBB2 expression in HER2-enriched tumors. a. Histogram of standard deviation of expression of each gene among HER2-enriched tumor samples. **b.** Gaussian mixture model fit to log of ERBB2 expression values in epithelial cells of HER2-enriched tumors. **c,f,i.** Histogram/density of simulated expression values of ERBB2 in epithelial cells. **d,g,j.** EDec performance in 1000 simulated datasets containing 78 samples each generated using the distributions represented in **c**, **f** and **i** respectively**. e,h,k.** EDec performance in 1000 simulated datasets containing 500 samples each generated using the distributions represented in **c**, **f** and **i** respectively**.**

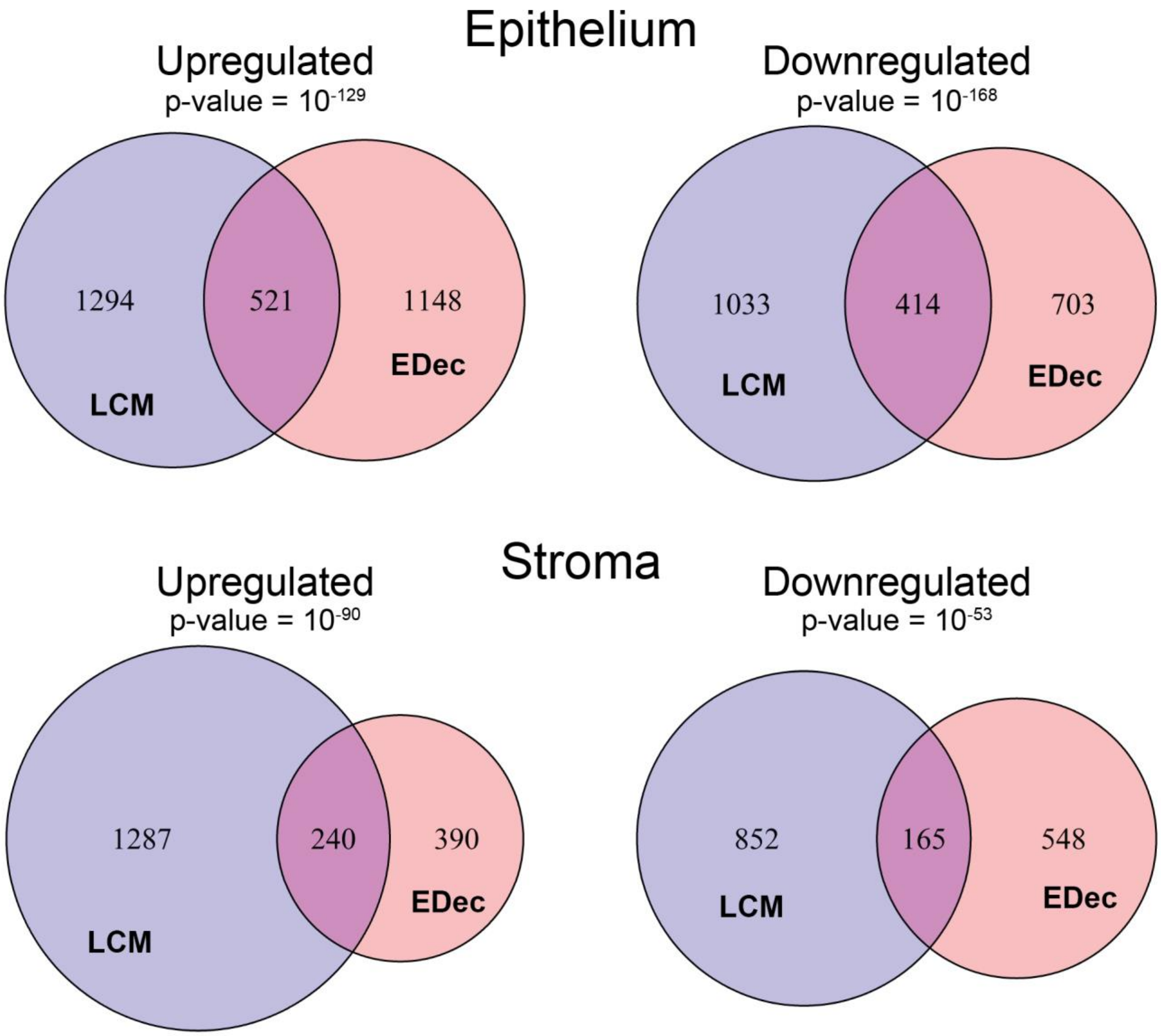


Figure S4. Related to Figure 4. Comparison of differentially expressed genes identified by EDec and LCM. Venn diagrams representing the overlap between differentially expressed genes in laser capture microdissection dataset and EDec analysis of TCGA dataset. P-values were computed using the hypergeometric test, and assuming that the full set of genes contained 16,708 genes (genes covered both by TCGA RNA-seq assay, and microarrays used to profile expression in LCM dataset).

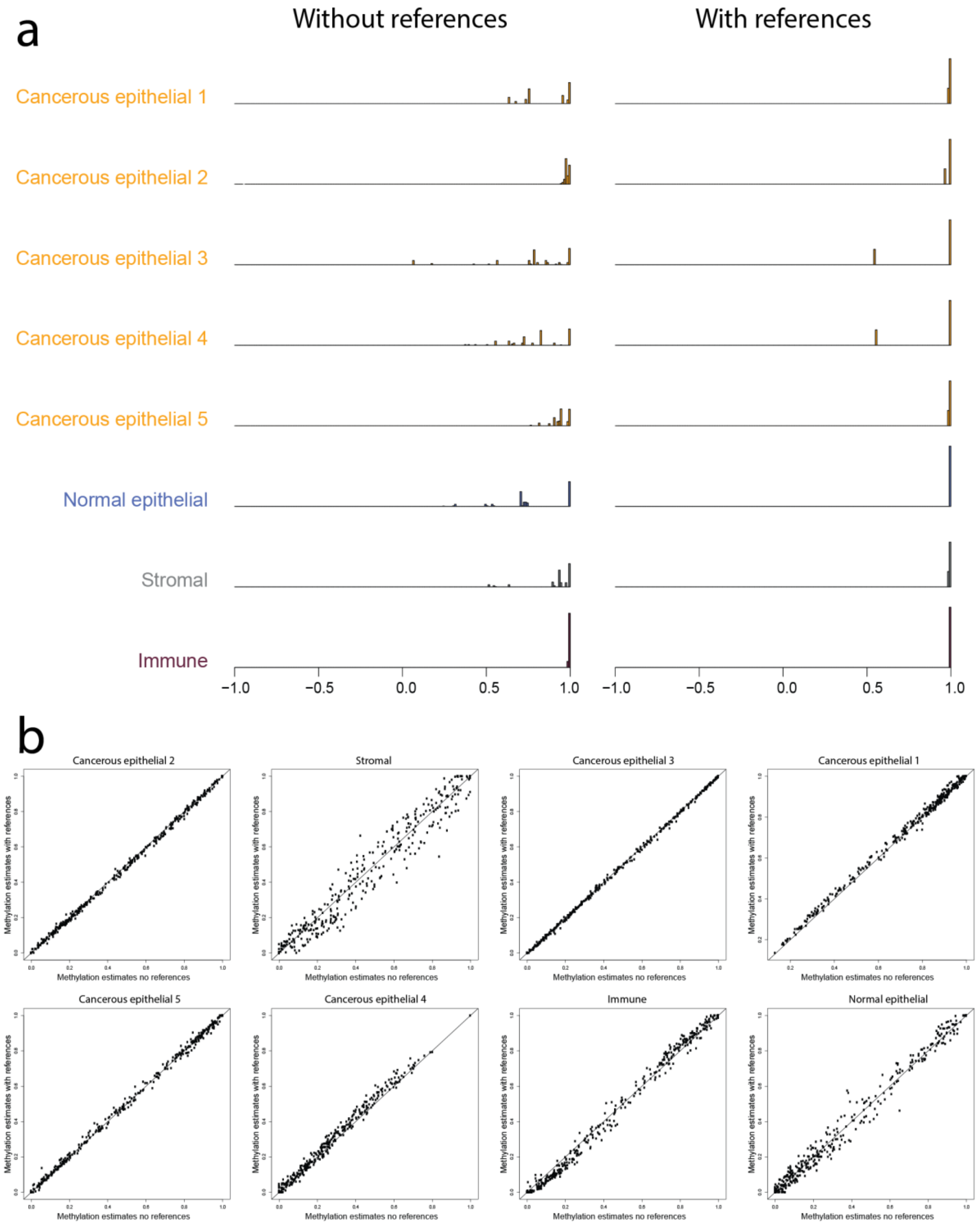


Figure S5. Related to Figure 3. Stability of deconvolution with or without the addition of reference methylation profiles to the TCGA dataset. a. Histograms of correlations between pairs of methylation profiles estimated for each of the 8 cell types in the 20 runs of EDec with references and in the 20 runs of EDec without references. **b.** Scatterplots comparing the methylation profile estimates for each cell type in the best solution with references added against the best solution without added references.

Supplemental Tables

Table S1. Functional annotation clustering of differentially expressed genes in each constituent cell type. See TableS1.xlsx.

Table S2. Description of loci amplified in targeted bisulfite sequencing experiments. See TableS2.xlsx

Extended Experimental Procedures

EDec Stage 1 formal description

Suppose we have an experiment in which N complex tissue samples have been profiled for methylation over M loci. We represent the result of this experiment as a matrix V_{MxN} of beta values, in which each column corresponds to one of the N samples, and each row corresponds to one of the M loci profiled for methylation. Further, suppose that the complex tissues profiled in our experiment are constituted by K cell types, and that the matrix W_{MxK} of beta values contains in each of its columns the methylation profiles (over the same M loci), of each of those K cell types. Lastly, suppose that the matrix H_{KxN} contains in its nth column the proportions of each of the K cell types in the nth sample of the complex tissue. Note that every value in V, W, and H is a number between 0 and 1. Further, note that the columns of H must sum up to 1, since the sum of proportions for all the cell types that constitute a particular sample will comprise 100% of the cells in that sample.

We assume that the methylation level in each probed locus for a sample of a complex tissue is linearly related to the proportion of each constituent cell type in that sample and to the level of methylation in that same locus in each of those cell types. This assumption leads to the following model:

$$V = WH$$

In the method proposed here we assume that both W and H matrices are unknown and must be estimated based on V alone. We assume that the best estimates of W and H are those that satisfy:

$$argmin_{W,H}||V - WH||_F^2$$

with $0 \leq W \leq 1$, $0 \leq H \leq 1$, and $\sum_{i=1}^{K} H_{ij} = 1$ for every j between 1 and N.

The problem of simultaneously identifying the W and H matrices that satisfy the condition above, but substituting the boundary conditions for $0 \leq W$and $0 \leq H$, is one of the formulations of the non-negative matrix factorization problem. It has been shown that finding the globally optimal solution to that problem is NP-hard. Therefore, many heuristics have been proposed that attempt to identify locally minimal solutions to the problem. One class of such heuristics, previously referred to as block coordinate descent (Kim et al., 2013) , has been adapted by us to identify the solution to the same problem with the new boundary conditions.

The idea behind the block coordinate descent heuristic is to perform an iterative procedure in which we alternatively assume a fixed H and estimate W, then assume a fixed W (previously estimated) and estimate H (to be used in next iteration). Identifying the W matrix that satisfies $argmin_W||V - WH||_F^2$ with $0 \leq W \leq 1$ given V and H, can be done through quadratic programming algorithms (Zhong et al., 2012) . Similarly, identifying the H matrix that satisfies

$argmin_H ||V - WH||_F^2$ with $0 \leq H \leq 1$ and $\sum_{i=1}^{K} H_{ij} = 1$, given V and W can also be done through quadratic programming.

We initialize our algorithm with random guesses of proportions of cell types in each sample (randomized H matrix). Such proportions are generated from a Dirichlet distribution to guarantee that the boundary conditions on H are satisfied. The iterative procedure then goes on until it reaches a specified maximum number of iterations, or until:

$$||V - W^i H^i||_F^2 - ||V - W^{i-1} H^{i-1}||_F^2 \leq \varepsilon$$

with i being the number of iterations. In the experiments performed in this article, the maximum number of iterations was either 800 or 2000, and the chosen ε was either 10^{-8} or 10^{-10}.

The method presented here was implemented in R. The quadratic programming approximations to W and H matrices are performed using the quadprog library (Turlach, B.A et al., 2013).

EDec Stage 2 formal description

The levels of expression of a set of P genes for a set of N complex tissue samples are represented here as a matrix Y_{PxN} of non-negative real numbers. We assume that the gene expression profiles of complex tissue samples are linearly related to the gene expression profiles of each of its K constituent cell types (Z_{PxK}), and to the proportions of each constituent cell type in each sample (H_{KxN}). Those assumptions lead us to the model:

$$Y=ZH$$

The Stage 2 of the EDec method assumes that the proportions of constituent cell types in the aliquot of a complex tissue sample used for DNA methylation profiling and the one used for gene expression profiling are the same. Further we assume that the matrix Y of gene expression profiles of complex tissue samples is known, and that the proportions of constituent cell types for that set of complex tissue samples has already been estimated from the DNA methylation data through the Stage 1 of the EDec method. With those assumption the Stage 2 of EDec method attempts to solve the following problem:

$$argmin_Z ||Y - ZH||_F^2$$

with $0 \leq Z$. This problem can be solved with the nonnegativity constraint on Z by quadratic programming. We use the quadprog package (Turlach, B.A et al., 2013) to complete that task.

Estimating standard error for cell type specific gene expression

As illustrated in Figure 1c, the estimation of mean gene expression values for each constituent cell type is performed by assuming a linear model in which the level of expression of a gene in the tissue sample is a linear combination of the levels of expression of that gene in each constituent cell type. In such model, the explanatory variables are the proportions of constituent cell types, which we assume were accurately estimated in the first stage of EDec using DNA methylation. Similarly to a multiple linear regression problem, using this model the average gene expression values in each constituent cell type are estimated through least squares optimization (solved with the nonnegativity constraint through quadratic programming). Again in the same way as for a multiple linear regression problem, the standard error (S_e) for each of the coefficients (levels of expression in each gene (i) in each constituent cell type (j)) in our linear model can be estimated using the formula:

$$S_e\{Z_{i,j}\} = \sqrt{[MSE_i\,(HH^t)^{-1}]_{j,j}}$$

where MSE$_i$ (mean squared error for gene i) can be computed by:

$$MSE_i = \frac{\sum_{j=1}^{K} {R_{i,j}}^2}{N-K}$$

with N being the number of tissue samples, K being the number of constituent cell types, and being the matrix of residuals (R=Y-ZH).

Details on cell culture and human breast tissue sample preparation

Normal Human Mammary Epithelial Cells (Clonetics), primary human fibroblasts (Asterand), human CD8+ cytotoxic T-cells (Sanguine), and breast cancer cell lines – MCF7, MDA-MB-231, MDA-MB-361, HCC1954, HCC1569, MCF10A (ATCC)—were cultured according to the respective manufacturer protocol. Logarithmically growing cultures were harvested at ~75-90% confluency. Frozen, primary human breast tumor tissue and adjacent normal tissue were obtained with local Institutional Review Board (IRB# PRO11090404) from the University of Pittsburgh's Health Science Tissue Bank. Frozen samples were pulverized with mortar and pestle under liquid nitrogen conditions. DNA from cell culture, tumor tissue, normal tissue, and buffy coat was isolated using Qiagen's DNeasy Blood and Tissue kit and bisulfite converted with the EpiTect Bisulfite Kit (Qiagen). Bisulfite converted DNA was quantified by Nanodrop, mixed in the indicated proportions, and sent to RainDance Technologies for assessment of quality, amplification of regions of interest, construction of libraries, and sequencing.

Simulating cell type mixtures

A set of 9 methylation profiles of cell lines generated using the targeted bisulfite sequencing assay was used to build simulated mixtures. The methylation profiles in this set corresponded to 6 different breast cancer cell lines (MCF-7, T47D, MDA-MB-231, MDA-MB-361, HCC1954, HCC1569), a normal breast cell line (HMEC), a CAF cell line, and purified T-cells. Each simulated mixture sample was constituted of 4 different cell types, including one of the 6 breast cancer cell lines, and the other 3 normal cell types (normal breast epithelial, stromal, and immune). For each mixed sample, the breast cancer cell line that was used was chosen randomly.

Once the cell types that would constitute a particular mixture were chosen, we used independent beta random variables to generate noisy versions of each of their methylation profiles. For each locus of a particular methylation profile, we would use the original methylation level for that locus as the mean of the beta random variable. Since the beta random variable used here has values restricted to the [0,1] interval, its variance cannot go beyond a certain limit (variance < mean*(1-mean)). The variance for the beta random variables used in our simulations was chosen as 10% of the maximum variance allowed for a beta variable with the given mean when dealing with a breast cancer profile, and 5% of the maximum variance allowed for a beta with the given mean when dealing with normal cell types. A random value was then generated from that distribution and was used as the methylation value for that locus on the noisy version of the methylation profile of that particular cell type. Once this procedure was performed for every locus of the four methylation profiles that would be used to build that particular mixture sample, a linear combination of the noisy methylation profiles with a given set of proportions gave us the methylation profile of that mixture.

The proportions associated with each cell type were generated for each mixture from one of two dirichlet distributions. Those distributions both generated vectors $[x_1, x_2, x_3, x_4]$ where $x_i \in [0,1]$ and $\sum_{i=1}^{4} x_i = 1$. The moments of the dirichlet distributions can be represented as:

$$E[X_i] = \frac{\alpha_i}{\alpha_0} \text{ and } Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \text{ where } \alpha_0 = \sum_{i=1}^{4} \alpha_i$$

The first dirichlet distribution had parameters $\alpha = [12,2,2,4]$, generating sets of proportions with averages 0.6 for the breast cancer fraction, 0.1 for the normal breast fraction, 0.2 for the immune fraction, and 0.1 for the stromal fraction. The second dirichlet distribution had parameters $\alpha = [1,12,6,1]$, generating proportions with averages 0.05 for the breast cancer fraction, 0.6 for the normal breast fraction, 0.05 for the immune fraction, and 0.3 for the stromal fraction. These average proportions should emulate the expected proportions of cell types found in breast cancer samples or on normal breast samples respectively. In fact, such numbers were very similar to the average proportions found in pathologist estimates for each of these cellular fractions in breast cancer or normal breast samples.

EDec application to targeted bisulfite sequencing dataset

Selecting appropriate numbers of cell types In the targeted bisulfite sequencing dataset, we applied the EDec method assuming between 3 and 10 constituent cell types. For the models with each of the possible numbers of cell types, we then computed the Akaike Information Criterion (AIC) metric to determine which of those models was the most appropriate (Burnham, 2004) . That metric is dependent on the goodness of fit of the model, but also accounts for the possible overfitting that can occur when the number of components of the model increases. According to that criterion, the ideal number of cell types to be used in this deconvolution was six. It is worth noting, that the major components of the model tend to remain similar as the number of cell types changes.

EDec application to TCGA breast cancer DNA methylation dataset

Selecting cell type specific probes A set of reference DNA methylation profiles was compiled from a series of previously published datasets. It contained: 450k array profiles of 25 different breast cancer cell lines (GSE44837), 9 different fibroblast cell lines (GSE40699) generated by the ENCODE project (ENCODE Project Consortium, 2012) , 8 different types of purified immune cells profiled with either 450k or 27k arrays with multiple replicates for each of them (GSE35069,GSE39981), 450k array profile of HMEC cell line (GSE40699), and whole genome bisulfite sequencing (WGBS) profiles for purified normal breast luminal epithelial cells and purified normal breast myoepithelial cells (GSE16368) generated by the Roadmap Epigenomics Project (Kundaje et al., 2015) . Final beta values were gathered directly from GEO for each of those datasets. In order to combine the 27k and 450k array data, we included only those probes from the 27k array that were also present in the 450k array. We also removed all probes known to overlap common SNPs, as well as those previously reported to show cross-reactivity. We also removed from our reference dataset all probes that had low detection p-values (< 0.05) for any of the reference or TCGA breast cancer samples. For the samples profiled using WGBS, we calculated the average level of methylation for all CpGs overlapping 100bp windows around each of the 450k array probes. If any 450k array probe did not have coverage in one of the WGBS profiles, that probe was also removed from the analysis. The final number of methylation probes included in our reference methylation profiles was 12,021. For cell lines, or immune cell types that had more than one replicate, we computed the average methylation profile for all replicates, and used that in all later analyses.

We then divided our reference methylation profile set into four groups: cancer epithelial cells, normal epithelial cells, stromal cells, and immune cells. The cancer epithelial group included 25 different breast cancer cell lines. The normal epithelial group included three cell types: HMEC, purified breast luminal epithelial cells, and purified breast myoepithelial cells. The stromal cell type group included nine different fibroblast cell lines. Lastly, the immune cell type group included 8 different types of purified immune cells: CD8+ T-cells, CD4+ T-cells, T regulatory cells, Mixed T-cells, Granulocytes, Monocytes, NK cells, and B-cells.

Once the cell type groups were defined, we performed t-tests comparing the methylation levels over each probe between each group of references against the rest of the reference methylation profiles. From among the probes that showed significant differences (p-value < 0.0001) in the comparison of each group against the rest of the reference samples we selected the 50 most hypermethylated and the 50 most hypomethylated probes. Due to the greater similarity between normal epithelial and cancerous epithelial cell types, we performed a specific t-test comparing only the samples in those two groups, and included in our final set of probes those that had a significant difference (p-value < 0.00001) and had the 100 highest or 100 lowest differences in methylation between those two groups. Due to some overlap between probes selected in each comparison, the final set contained 391 probes.

Addition of reference methylation profiles to TCGA dataset Given the heuristic nature of the EDec method, it is possible that for some runs of the method the iterative procedure in Stage 1 will get stuck in local minima that do not correspond to the true proportions and methylation profiles of constituent cell types. The selection of probes that are highly variable across cell types is an attempt to attenuate this issue, by making the convergence landscape smoother. In order to further mitigate this issue, we have included in the TCGA DNA methylation dataset 20 of the reference methylation profiles that we compiled from the public domain. Among the methylation profiles that were included in the dataset were the 9 fibroblast cell lines, 8 immune cell types, and 3 normal epithelial cell types. By running the method 20 times with the references and 20 times without, we show that the addition of such references indeed improved the stability of the solution, while having minimal impact in the proportions and methylation profiles found as the globally optimal solution (Figure S5). Therefore, the addition of such references to the dataset helped guide the method to identifying components that corresponded to the real cell types that constituted the breast cancer samples, while introducing minimal bias due to the relatively small number of samples that were included.

Selecting appropriate numbers of cell types Similarly to what was done for the targeted bisulfite sequencing dataset, we've also attempted to apply the AIC metric as a guide for picking an appropriate number of cell types in the TCGA dataset. However, the number of cell types found to give the model with best AIC was 23. Due to the difficulty in interpretation of such model, we've focused instead on looking for a number of cell types that led to models with good fit while at the same time giving highly reproducible and interpretable models. In order to select an appropriate number of cell types for the deconvolution, and simultaneously investigate the reproducibility of our methylation and proportion estimates, we created three datasets containing a random subsampling of 80% of the samples in the TCGA methylation dataset plus the 20 reference profiles. We then applied the EDec method to each of these datasets with the number of cell types varying between 4 and 15. With that, we were able to compare the estimated methylation profiles and proportions across these three partially overlapping dataset and analyze the level of reproducibility of the deconvolution with each of the chosen number of cell types. As expected, we observed that lower numbers of cell types tend to give higher degree of reproducibility, but the goodness of fit of the final model is better with higher number of cell types. Due to its near perfect reproducibility across replicates, and high level of explained variance we decided to select the model with 8 cell types for all further analyses. It is worth noting, however, that the major components of the model tend to remain similar as the number of cell types changes.

Cell type specific comparative analyses of gene expression Given the estimated means and standard error for the level of expression of each gene in each constituent cell type in any two groups of samples, we determined whether a gene was significantly differentially expressed between those two groups using a t-test. This test assumes that for each gene the residuals are independent, have mean zero, have constant variance, and are normally distributed. Given the nonnegativity constraint included in our model, the assumption of normality is likely violated. Also, we do not account here for the possible errors in the estimation of proportions of constituent cell types. Therefore, even though the computed p-values can be helpful in identifying genes with large differences in expression between different groups of samples, it is likely that differences identified with borderline levels of significance may be unreliable. In our comparison between breast cancer and normal breast, we required very strict significance thresholds in order to claim that a gene was indeed differentially expressed (FDR less than 0.001 and fold change in expression greater than or equal to 2).

Processing and analysis of laser capture microdissection dataset

The processed dataset containing gene expression profiles of laser capture microdissected tumors (Ma et al., 2009) was downloaded from NCBI GEO (accession number GSE14548). We used the 9 available sets of matched epithelium and stromal samples from both invasive breast carcinoma and adjacent normal breast. Epithelial and stromal samples from in situ carcinoma and breast tumor samples that did not have all four types of cells (invasive carcinoma epithelium and stoma and normal breast epithelium and stroma) were discarded. The Limma R package was used to perform paired differential expression analysis between invasive carcinoma epithelium and normal breast epithelium as well as between invasive carcinoma stroma and normal breast stroma. Genes were considered as differentially expressed if at least one probe associated with that gene had FDR of less than 0.05.

Heterogeneity of epithelial methylation profiles within and between tumor types

We assume that EDec-estimated TCGA cell type proportions, and EDec-estimated methylation profiles of stromal and immune cell types are correct and do not vary between tumor samples (all unexplained variability is due to differences in cancerous epithelial cell type). Under those assumptions the methylation profile of epithelial cells for a particular tumor sample can be written as:

$$W_e = \frac{V - H_s W_s - H_i W_i}{H_e}$$

Where V is the known methylation profile of the bulk tumor sample; H_e, H_s, and H_i, represent the known proportions of epithelial, stromal and immune cell types in that sample respectively; W_s and W_i represent the known methylation profiles of stromal and immune cell types in that sample respectively; and W_e represents the methylation profile of the epithelial fraction in that sample, the only unknown variable in that equation under the above assumptions.

Once the methylation profiles of the epithelial fraction of each tumor sample were estimated in that fashion, we were able to compute the spearman correlation between epithelial methylation profiles for pairs of samples either from the same or from different breast tumor subtypes over the set of 391 loci used in the TCGA deconvolution. A heat map with the average values of spearman correlation between methylation profiles of epithelial fraction of all sample pairs in the same or different subtypes is shown in Figure S1.

Analysis of ERBB2 expression in HER2-enriched tumors

In Figure 4a one can observe that EDec assigns a high expression for the ERBB2 gene in the stroma of HER2 enriched tumors. Due to the fact that ERBB2 protein expression in breast tumors is routinely assayed by immunohistochemistry, and stromal expression of ERBB2 is not reported, we assume that such assignment is an artifact. Further, the high standard error associated with that particular estimate strengthens that assumption.

In order to determine what caused EDec to make such false statement, we looked at the distribution of the expression of ERBB2 across HER2-enriched tumors. We noticed that the expression of that gene has a particularly high degree of variability (Figure S3a), being among the 10 genes with highest standard deviation among HER2-enriched tumors. We also observed that the distribution of ERBB2 expression in epithelial cells (ERBB2 expression divided by the proportion of epithelial cells in each sample), instead of following a log-normal distribution with a single mode like other genes, had a bimodal distribution in log scale (Figure S3b). With the intent of simulating ERBB2 expression values, we fit a Gaussian mixture model on the logarithm of ERBB2 expression, inferring therefore the mean and standard deviation for two Gaussian distributions. When combined, those distributions fit the logarithm of expression values of ERBB2 quite well (Figure S3b). Using the parameters of those two Gaussian distributions, as well as the number of HER2-enriched samples best explained by each of them, we were able to simulate ERBB2 expression values in epithelial cells by sampling from those two log-normal distributions. We also used a Dirichlet distribution to simulate proportions of epithelial, stromal, and immune cell types. The parameters of that distribution were adjusted to match the mean and standard deviation of proportions of each cell type across HER2-enriched tumors (as estimated by EDec). Finally, to simulate the expression of ERBB2 in bulk tumor samples, we multiplied each simulated expression of ERBB2 in epithelial cells by the simulated proportion of epithelial cells. This procedure leads to expression values that assume no expression of ERBB2 in either stromal or immune cells.

Given a set of simulated expression values of ERBB2 in bulk tumor samples and the set of proportion of epithelial, stromal, and immune cells in each sample, we were able to verify the performance of the Stage 2 of EDec in a situation very similar to the real HER2-enriched tumor set. However, with this simulation framework, we were also able to control the variance in ERBB2 expression as well as the number of samples, allowing us to investigate how those parameters affect the deconvolution performance. Figures S4c, S4f, and S4i display the distribution of simulated values of ERBB2 expression in epithelial cells in three different simulations. Note that in Figure S3c we used the same distribution parameters as the mixed Gaussian models fit in the real ERBB2 expression data, leading to a similar level of standard deviation as the original expression values. In the simulations represented in Figures S4f and S4i, the means of the two Gaussians were brought closer to the center, and the standard deviations of each of them were also reduced, leading to decreased variance in ERBB2 expression. 1000 simulated sets of the same size (78 samples) as the original set of HER2-enriched samples were generated for each set of parameters. In figures S4d, S4g and S4j, we display a scatterplot of the true expression values of ERBB2 in each cell type versus those estimated by EDec in the simulation datasets represented in figures S4c, S4f, and S4i respectively. Notice that as the variance of ERBB2 expression within epithelial cells is reduced, the performance of EDec significantly improves. In particular, notice that the expression values assigned to immune or stromal fractions, which should be all zero, are significantly reduced as that variance decreases. Further note that even in the situation with highest level of variance (Figure S3d), even though high mean values are often incorrectly assigned to immune and stromal cells, a t-test using the mean and standard deviation values estimated by EDec fails to reject the hypothesis that the stromal and immune expression are different from zero in over 99% of the cases. In contrast, the expression of ERBB2 in epithelial cells can always be found to be significantly different from zero.

In order to investigate whether an increase in the number of tumor samples can improve EDec performance for genes with extremely high variance levels, we generated 1000 simulated ERBB2 expression datasets containing 500 samples each, instead of the 78 samples per dataset used before. In figures S4e, S4h and S4k, we display scatterplots of the true expression values of ERBB2 in each cell type versus those estimated by EDec in the simulation datasets with 500 samples each generated according to the distributions represented in figures S4c, S4f, and S4i respectively. Notice that the performance of EDec when the sample size increases improves significantly compared to the datasets with smaller sample sizes.

From these analyses we conclude that the incorrect assignment of ERBB2 expression to the stromal fraction of HER2-enriched tumors is indeed caused by the extreme level of variability of ERBB2 expression within the epithelial fraction. We further show that, similarly to what happens in the application of EDec to TCGA, the standard error assigned to the expression of ERBB2 in stromal or immune cell types also tends to be high in simulated datasets, leading to failure to reject the hypothesis that those expression values are significantly different from zero. Lastly, we show that increasing sample size can indeed mitigate the issues in estimation of cell type specific expression for genes with exceedingly high variance.

Supplemental References

Burnham, K.P., and Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33, 261–304.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Kim, J., He, Y., and Park, H. (2013). Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J. Glob. Optim. 58, 285319.

Kundaje, A., et al., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–30.

Ma, X.-J.J., Dahiya, S., Richardson, E., Erlander, M., Sgroi, D.C., 2009. Gene expression profiling of the tumor microenvironment during breast cancer progression. Breast Cancer Res. 11, R7.

Turlach, B.A., and Weingessel, A. (2013). quadprog: Functions to solve Quadratic Programming Problems. R package version 1.5-5, http://cran.r-project.org/web/packages/quadprog/index.html.

Zhong, Y., Wan, Y.-W., Pang, K., Chow, L., Liu, Z., 2013. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC bioinformatics 14, 89.