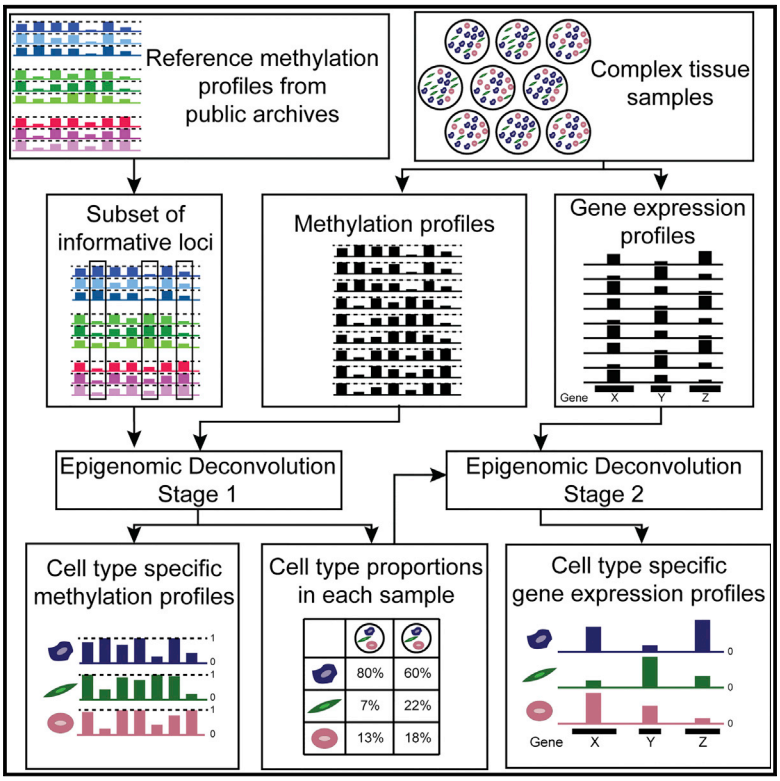
Resource

Cell Reports

Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types

Graphical Abstract



Highlights

- EDec infers cell types within tissues and molecular profiles of constituent cells
- EDec deconvolutes molecular profiles of breast tumors within the TCGA collection
- EDec-estimated immune infiltration predicts prognosis for basal-like breast tumors
- Switch from adipose to fibrous stroma enhances oxidative metabolism of cancer cells

Authors

Vitor Onuchic, Ryan J. Hartmaier, David N. Boone, ..., Matt E. Roth, Adrian V. Lee, Aleksandar Milosavljevic

Correspondence

onuchic@bcm.edu (V.O.), amilosav@bcm.edu (A.M.)

In Brief

Onuchic et al. develop an in silico deconvolution technique (EDec) that can accurately estimate cell type composition and molecular profiles of constituent cell types in the context of breast tumors. Application to breast cancers from TCGA data reveals association between stromal composition and the metabolic phonotype of breast tumors. Explore consortium data at the Cell Press IHEC webportal at www.cell.com/consortium/IHEC.

Accession Numbers

GSE87297

Onuchic et al., 2016, Cell Reports *17*, 2075–2086
November 15, 2016
http://dx.doi.org/10.1016/j.celrep.2016.10.057

CellPress

Cell Reports
Resource

Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types

Vitor Onuchic,[1,6,*] Ryan J. Hartmaier,[2,7] David N. Boone,[2] Michael L. Samuels,[3] Ronak Y. Patel,[1] Wendy M. White,[4] Vesna D. Garovic,[5] Steffi Oesterreich,[2] Matt E. Roth,[1] Adrian V. Lee,[2] and Aleksandar Milosavljevic[1,6,8,*]
[1]Molecular and Human Genetics Department, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA
[2]Department of Pharmacology and Chemical Biology, Magee Womens Research Institute, University of Pittsburgh Cancer Institute, 204 Craft Avenue, B705, Pittsburgh, PA 15213, USA
[3]RainDance Technologies, Inc., 749 Middlesex Turnpike, Billerica, MA 01821, USA
[4]Department of Obstetrics and Gynecology, Mayo Clinic College of Medicine, 200 1st Street SW, Rochester, MN 55905, USA
[5]Division of Nephrology and Hypertension, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, USA
[6]Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA
[7]Present address: Foundation Medicine, Inc., 150 Second Street, Cambridge, MA 02141, USA
[8]Lead Contact
*Correspondence: onuchic@bcm.edu (V.O.), amilosav@bcm.edu (A.M.)
http://dx.doi.org/10.1016/j.celrep.2016.10.057

SUMMARY

Cancer progression depends on both cell-intrinsic processes and interactions between different cell types. However, large-scale assessment of cell type composition and molecular profiles of individual cell types within tumors remains challenging. To address this, we developed epigenomic deconvolution (EDec), an in silico method that infers cell type composition of complex tissues as well as DNA methylation and gene transcription profiles of constituent cell types. By applying EDec to The Cancer Genome Atlas (TCGA) breast tumors, we detect changes in immune cell infiltration related to patient prognosis, and a striking change in stromal fibroblast-to-adipocyte ratio across breast cancer subtypes. Furthermore, we show that a less adipose stroma tends to display lower levels of mitochondrial activity and to be associated with cancerous cells with higher levels of oxidative metabolism. These findings highlight the role of stromal composition in the metabolic coupling between distinct cell types within tumors.

INTRODUCTION

Molecular profiling of breast tumors has led to their categorization into different subtypes with distinct risks and underlying biology. Of particular interest is the classification into five intrinsic subtypes, which can be performed using the prediction analysis of microarray 50 (PAM50) classifier (Parker et al., 2009). However, most molecular-profiling studies to date have been performed on bulk tissue samples, ignoring the complexity of the breast tissue, with its multiple cell types and the interactions between them. Valuable evidence for the significance of heterotypic interactions comes from the study of cell type composition of tumors, as exemplified by the prognostic value of immune cell infiltration (Coussens et al., 2013; Liu et al., 2014) and of epigenomic (Hu et al., 2005) and transcriptomic (Finak et al., 2008) perturbations within stromal cells (Tlsty and Coussens, 2006). Laser capture microdissection (LCM), cell sorting, and other physical methods to isolate cell types from solid tumors for molecular profiling are technically challenging, and severely limit throughput (Debey et al., 2004). A number of methods for in silico deconvolution have been developed to address this problem using as input gene expression profiles (Aran et al., 2015; Gentles et al., 2015; Houseman and Ince, 2014; Kuhn et al., 2011; Li and Xie, 2013; Newman et al., 2015; Shen-Orr et al., 2010; Venet et al., 2001; Yoshihara et al., 2013; Zhong et al., 2013) and, more recently, DNA methylation profiles (Houseman et al., 2012, 2014, 2016; Zheng et al., 2014; Zou et al., 2014; Rahmani et al., 2016) of tissue homogenates. However, the ability of these methods to infer cell type composition of solid tumors and interpret the states of constituent cell types is limited, thus hampering the study of cellular states and cellular interactions within the tumor microenvironment.

To address this gap, we developed epigenomic deconvolution (EDec), a deconvolution method based on a heuristic for constrained matrix factorization using quadratic programming. The deconvolution is based on cell-type-specific patterns of DNA methylation. Such patterns are acquired during normal cellular differentiation, maintained through cell division, and serve as chemically stable cellular markers. We reasoned that methylation profiles would be more amenable to deconvolution than gene expression due to their linearity, measurement within the complete (0–1) dynamic range, and technology independence (including both bisulfite sequencing and array platforms).

Previous methylation-based deconvolution methods either make direct use of reference methylation profiles of constituent cell types (Houseman et al., 2012) or ignore such references

(Houseman et al., 2014, 2016; Rahmani et al., 2016; Zou et al., 2014). Highly accurate reference methylation profiles, essential for reference-based deconvolution approaches, are unavailable for many solid tissues, arguing for a reference-free approach. However, reference methylation profiles from representative cell lines are available and can provide valuable information if used to improve inference while minimizing bias. Toward this goal, EDec uses relevant reference information in indirect ways to minimize bias. First, it uses references to identify sets of loci that are likely to exhibit variation in methylation levels across constituent cell types of a given tissue (feature selection), while taking a reference-free approach to the deconvolution problem itself. Second, it identifies constituent cell types by comparing their deconvoluted molecular profiles to reference profiles.

EDec consists of three stages (0, 1, and 2; Figure 1). Starting with methylation profiles of tumor homogenates over loci selected based on reference methylation profiles (Figure 1A, stage 0), EDec estimates both cell type proportions and methylation profiles of constituent cell types using an reference-free approach (Figure 1A, stage 1) similar to previous reference-free techniques (Gaujoux and Seoighe, 2012; Houseman et al., 2016). The proportion estimates are then used as a "key" to deconvolute gene expression profiles of constituent cell types (Figure 1A, stage 2).

EDec proof-of-concept experiments were performed using both Illumina methylation arrays and RainDance Technologies' ThunderStorm bisulfite sequencing (BS-seq) (Komori et al., 2011; Paul et al., 2014) targeted bisulfite sequencing. The method is validated using both computer simulations and profiling experiments on prepared cell mixtures. By applying EDec to the breast cancer datasets generated by The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Network, 2012), we predict cellular proportions and methylation states of constituent cell types within breast tumors as well as infer changes in gene expression within each constituent cell type. Such predictions were largely confirmed by comparisons with cell type composition estimates based on H&E staining, and by comparison against gene expression profiles of specific cell types isolated through LCM. We show that cancerous epithelial cells exhibit methylomes distinct from those of normal epithelium. EDec also replicates the previously reported association between increased immune cell infiltration in triple-negative breast cancer and better prognosis (Adams et al., 2014). We further detect expression changes that are highly consistent with known hallmarks of cancer, and with known roles of specific cell types within breast cancer. Last, we observe that the degree of stromal adiposity across breast cancer subtypes predicts the pattern of metabolic coupling observed between cancer epithelium and stroma.

RESULTS

Epigenomic Deconvolution Method

The first stage of EDec (Figure 1A, stage 1) performs constrained matrix factorization to find cell type-specific methylation profiles and constituent cell type proportions that minimize the Euclidian distance between their linear combination and the original matrix of tissue methylation profiles (Figure 1B). The minimization algorithm involves an iterative procedure that, in each round, alternates between estimating constituent cell type proportions and methylation profiles by solving constrained least-squares problems through quadratic programming. The minimization problem is made tractable by the constraints that methylation measurements (beta values) and cell type proportions are numbers in the [0, 1] interval, and that cell type proportions within a sample add up to 1. These constraints restrict the space of possible solutions, thus making it possible for the local iterative search to reproducibly find a global minimum and an accurate solution. One key requirement for EDec is that cell type proportions vary across samples. A second requirement is that there must be significant differences across constituent cell type methylation profiles. The latter requirement can be met by providing EDec with loci expected to vary in methylation levels across constituent cell types (Figure 1A, stage 0).

Similar to how tissue methylation profiles are modeled, tissue gene expression profiles can also be modeled by the linear combination of the expression profiles of its constituent cell types. However, due to the less constrained nature of gene expression measurements ($[0, \infty]$) versus methylation measurements ([0, 1]), the same reference-free approach used in stage 1 is not as effective for gene expression deconvolution. Therefore, instead of using that approach, when both DNA methylation and gene expression profiles are available for the same set of samples (e.g., from the same tissue homogenate), EDec-stage 2 uses the cell proportions estimated in stage 1 as a fixed input when estimating the average gene expression profiles of constituent cell types through a constrained least-squares fit using quadratic programming with solutions constrained to $[0, \infty]$ (Figure 1A, stage 2, and Figure 1C).

Validation using in Silico Mixtures of Methylation Profiles Derived from Breast Cancer-Related Cell Lines

We first validated the core EDec algorithm (stage 1) on simulated mixtures of experimentally derived DNA methylation profiles (nine cell lines: six breast cancer, one normal breast epithelial, one immune, and one cancer-associated fibroblast [CAF]). Among the 1,000 target genomic regions included in this breast cancer methylation-focused panel (Table S2), 149 exhibited particularly distinct methylation patterns across different breast cell types (based on reference epigenomes) (Kundaje et al., 2015) and were used in EDec-stage 1. The simulation dataset consisted of 100 mixtures, each composed of four cell types (one breast cancer cell line, one normal mammary epithelial cell type, one stromal cell type, and one immune cell type). About one-half of the simulated mixtures contained on average higher levels of breast cancer (60%) and immune cell types (20%), representing distributions observed in tumor samples such as those in the TCGA dataset. To simulate the presence of different breast cancer subtypes, different simulated mixtures had a different cancerous epithelium constituent. Specifically, the breast cancer cell type for each mixture was chosen randomly from the set of six breast cancer cell lines. Simulated normal breast contained higher than average levels of normal epithelial (60%) and stromal cell types (30%). To better represent real samples, random noise was introduced into the methylation

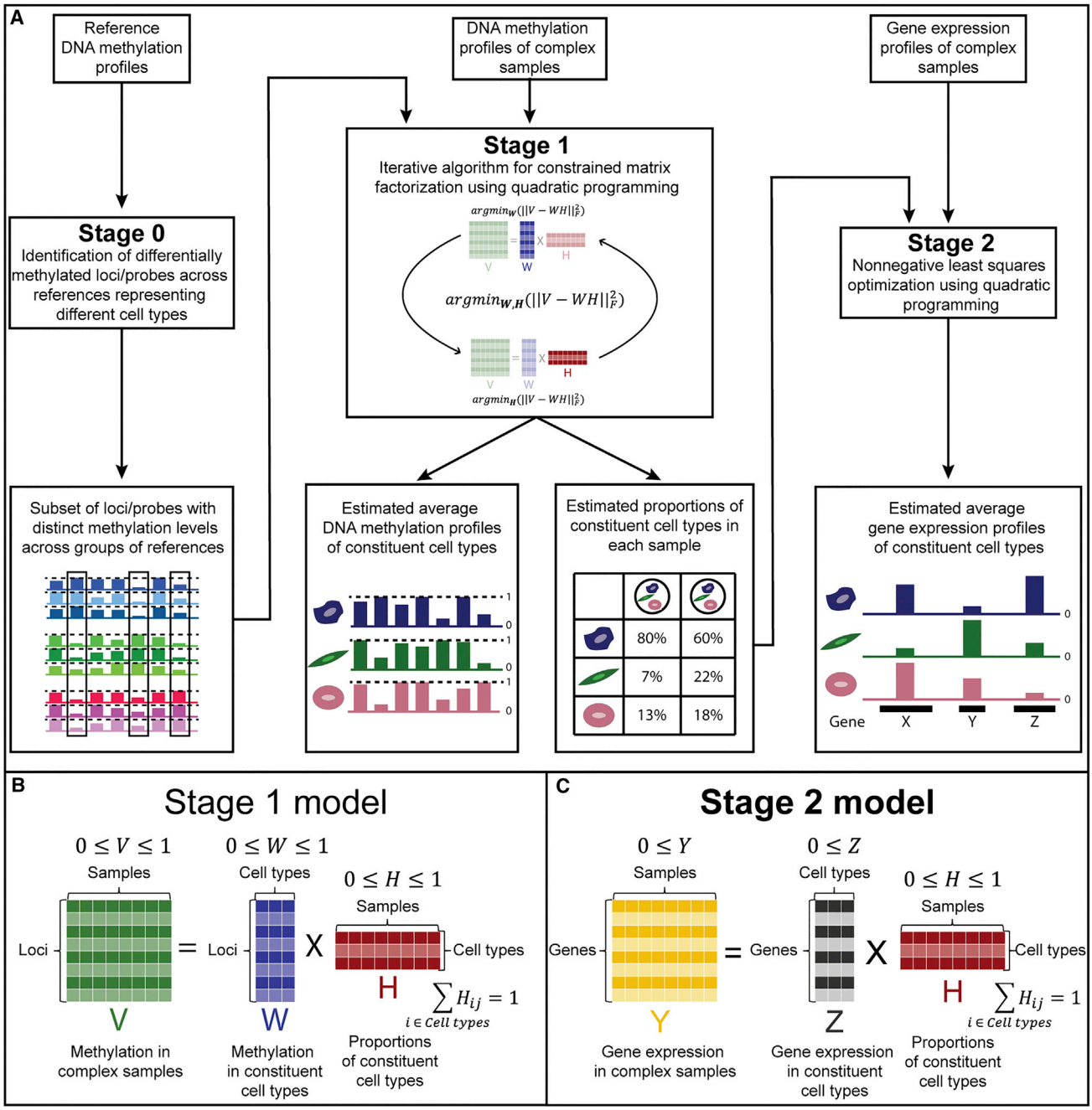


Figure 1. Description of the EDec Method

(A) The EDec method has two main stages (stages 1 and 2), preceded by a preparation stage (stage 0). In stage 0, a set of reference methylation profiles is used to select a set of genomic loci or array probes with distinct methylation levels across groups of references representing different constituent cell types. Methylation profiles of complex tissue samples over the set of loci/probes selected in stage 0 are used as the input for the stage 1 of the EDec method. In stage 1, EDec estimates both the average methylation profiles of constituent cell types and the proportions of constituent cell types in each input sample using an iterative algorithm for constrained matrix factorization using quadratic programming. Stage 2 of EDec takes as input the gene expression profiles of the same tissue samples profiled for DNA methylation, as well as the proportions of constituent cell types for those samples, estimated in stage 1, and outputs the gene expression profiles of constituent cell types.

(B) Representation of the model associated with stage 1 of EDec method.

(C) Representation of the model used for gene expression deconvolution in stage 2 of the EDec method.

profiles across all samples (Supplemental Experimental Procedures). We applied EDec to this dataset assuming nine different cell types in the model (six possible breast cancer cell lines, one normal epithelial, one stromal, and one immune). EDec accurately estimated DNA methylation profiles (r = 0.982; Figure 2A) and proportions (r = 0.983) for all constituent cell types (Figure 2B).

Validation on Cell Line Mixtures Profiled by Targeted Bisulfite Sequencing

We next validated EDec on cellular mixtures prepared in vitro. Specifically, we profiled 10 samples using targeted bisulfite sequencing and applied EDec using the set of 149 loci selected in EDec-stage 0. Four of the 10 samples were pure cells lines, including the following: MCF-7, HMEC (human mammary epithelial cells), a CAF cell line, and $CD8^+$ cytotoxic T cells. The other six samples consisted of three pairwise combinations (MCF-7/HMEC, MCF-7/T cells, and MCF-7/CAF), each in two proportions (75%:25% and 95%:5%). There was a strong concordance between the EDec estimated and the true proportions (r = 0.996; Figure 2C). In addition, the estimated methylation profiles for the four different cellular fractions closely matched the methylation profiles of cells used to create the mixtures (r = 0.998; Figure 2D).

Validation on Breast Tumor Samples Profiled by Targeted Bisulfite Sequencing

We next generated DNA methylation profiles for 31 breast tumors and 8 normal breast samples using targeted bisulfite sequencing. We applied EDec, assuming six constituent cell types (Supplemental Experimental Procedures), and asked how similar the estimated methylation profiles were to a set of external reference methylation profiles (Figure 2E). Three of the six estimated methylation profiles were most similar to one of the reference breast cancer cell lines. The three remaining profiles had particularly high correlation with the methylation profiles of either $CD8^+$ cytotoxic T cells, CAF cell line, or the HMEC cell line. This indicates that EDec identifies three components that explain the diversity of cancerous epithelial cells in those samples, whereas the other three components correspond to an immune fraction, a fibroblast/stromal fraction, and a normal epithelial fraction.

To further validate EDec, clinical pathologist evaluations of cell type composition were obtained for 29 of the 39 samples based on H&E staining. The pathologist estimated proportions for cancerous epithelial, normal epithelial, stromal, and immune fractions. Since the EDec method had proportion estimates for three different cancer epithelial fractions, we combined the proportions for those three fractions to make the two techniques comparable. Despite observing good consistency for the cancer epithelial and immune fractions, we observed low correlation for the normal epithelial and stromal fractions. We reasoned that the low correlation may be explained by extensive epithelial-mesenchymal transitions that may blur the boundary between epithelial and stromal cells. We therefore modified the analysis by combining proportion estimates of normal epithelial and stromal components and examined concordance of EDec and H&E proportion estimates for three fractions (cancerous epithelial, normal epithelial/stromal, and immune). The estimates were highly concordant for all three cell type fractions (r = 0.74, p value < 10^{-15}; Figure 2F). The highest correlation was for the immune fraction (0.78) and the lowest for cancerous epithelial fraction (0.67). The concordance between these two techniques indicates that EDec's estimates of proportions and methylation profiles correspond to real cell types and are not just general components that explain variability in the methylation dataset.

Deconvolution of Breast Tumors from the TCGA Collection Confirms the Role of Immune Response in Tumor Progression

We next applied EDec to deconvolute DNA methylation profiles of 1,061 breast tumors and 123 adjacent normal breast samples generated using Infinium HumanMethylation arrays by TCGA (Cancer Genome Atlas Network, 2012). We selected 391 informative loci (EDec-stage 0) from 45 reference DNA methylation profiles gathered from the NCBI GEO archive for the following four relevant cell types: cancer epithelial (25), normal epithelial (3), stromal (9), and immune (8) (Figure 3A) (Supplemental Experimental Procedures).

EDec-stage 1 (Figure 1A) was then applied to the TCGA DNA methylation data over the 391 probes, assuming 4–15 constituent cell types. Reference methylation profiles (20) were added to the TCGA dataset to improve stability of convergence (Supplemental Experimental Procedures and Figure S5). Based on model reproducibility and goodness of fit (Supplemental Experimental Procedures), we chose the model with eight cell types for all further analyses. We generated heat maps of correlations between the eight EDec-estimated methylation profiles and each GEO reference methylation profile (Figure 3B). The correlations suggest that EDec identified methylation profiles corresponding to one immune, one stromal, one normal epithelial, and five different cancerous epithelial components.

DNA methylation profiles were also generated for nine of the TCGA samples using targeted bisulfite sequencing. This allowed us to compare EDec-estimated proportions for those samples based on sequencing data, in the context of 39 breast tissue samples profiled by bisulfite sequencing, versus those estimates based on arrays in the context of 1,184 TCGA samples (Figure 3C). Estimated proportions were highly correlated (r = 0.88), suggesting that EDec operates independently of the methylation profiling method. EDec and pathologist (H&E staining) proportion estimates were also consistent (r = 0.90) (Figure 3D).

Consistent with expectations, EDec predicts normal breast samples to have negligible proportions of cancerous epithelial cells, whereas in breast tumors those cell types are generally the ones with highest proportions (Figure 3E). We also observe that the cancerous cell fraction of the different breast cancer samples is explained by a different combination of the five cancerous epithelial components, with one of them typically being dominant. Grouping tumor samples based on the dominant cancer epithelial component showed some concordance with their PAM50 classification (Parker et al., 2009). In particular, basal-like samples were nearly all in the same EDec-defined group (Figure 3E, red box). We further investigated methylation heterogeneity of the epithelial fraction over the 391 chosen probes within and between tumor subtypes (Supplemental

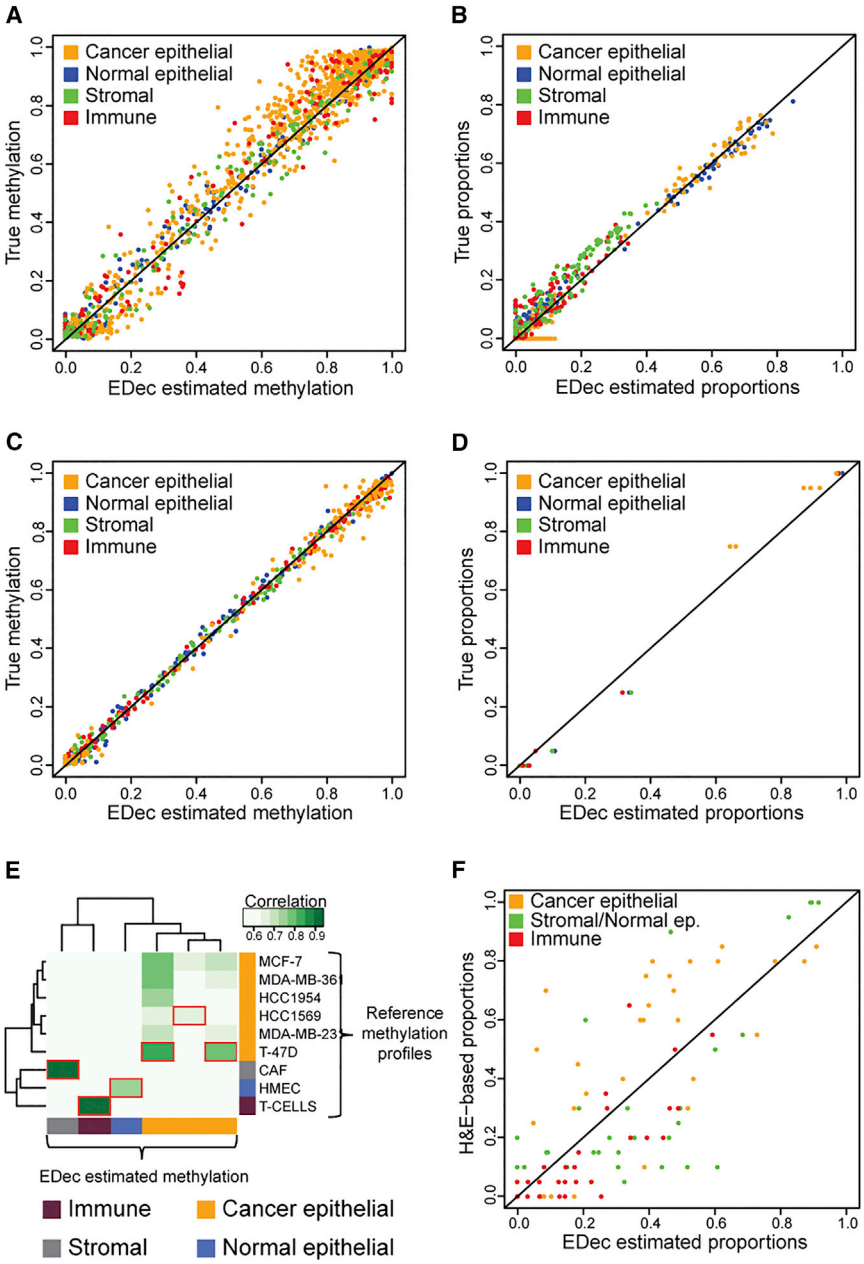


Figure 2. EDec Validation on Simulated Mixtures, Experimental Mixtures, and Solid Tumors

(A) Estimated versus true methylation levels for each constituent cell type and locus involved in the simulated mixtures dataset.

(B) Estimated versus true proportions for each constituent cell type in each of the samples involved in the simulated mixtures dataset.

(C) Estimated versus true methylation levels for each constituent cell type and locus profiled in the experimental mixtures dataset.

(D) Estimated versus true proportions for each constituent cell type in each of the samples profiled in the experimental mixtures dataset.

(E) Heat map representing the level of correlation between the estimated methylation profiles from the application of EDec to the targeted bisulfite-sequencing dataset and the reference methylation profiles. Red boxes indicate the highest level of correlation for each estimated methylation profile. The estimated methylation profiles were labeled as cancer epithelial, normal epithelial, immune, or stromal based on what reference methylation profile was most correlated to each of them.

(F) Proportion of constituent cell types estimated by EDec for samples in the targeted bisulfite-sequencing dataset versus pathologist-estimated proportions (H&E staining). Color key for all panels: orange (MCF-7), blue (HMEC), green (CAF), and red (T cell).

Experimental Procedures and Figure S1). Luminal B tumors had the most heterogeneous profiles, whereas normal breast samples had the most homogeneous epithelial profile. Despite having an intermediary level of heterogeneity, basal-like tumors exhibited epithelial methylation profiles highly distinct from the other breast tumor subtypes.

We also found that tumor subtypes differ significantly in the degree of infiltration by either immune or stromal cells (Figure S2). Normal-like samples contained the highest median stromal proportion (18%), and Luminal B tumors, the lowest (4%). Basal-like tumors displayed the highest median degree of immune cell infiltration (21%), whereas Luminal B tumors again had the lowest (7%). Normal breast tissue samples displayed a much higher median proportion of stromal cells (37%) than breast tumors (8%).

We next investigated whether the predicted immune proportion was associated with survival of basal-like breast cancer patients. Indeed, patients with greater than 20% immune cell infiltration survived significantly longer ($p < 0.01$) than those with less than 20% (Figure 3F), consistent with previous microscopy-based evaluation of immune cell infiltration (Adams et al., 2014). We also investigated whether immune infiltration levels in adjacent normal tissue were related to immune infiltration levels in the matched tumor sample. No such correlation was observed, indicating that immune infiltration of tumors is not dependent on the amount of immune cells in the surrounding normal tissue (Figure S2).

Deconvolution of RNA-Sequencing Profiles of Breast Tumors from the TCGA Collection Reveals Cell Type-Specific Tumorigenic Perturbations with the Tumor Microenvironment

Given the availability of both mRNA-sequencing and DNA methylation profiles for the TCGA breast samples, we applied EDec-stage 2 to estimate gene expression profiles of constituent cell types. EDec-stage 2 was independently applied to six subsets of the 1,114 TCGA expression profiles, corresponding to

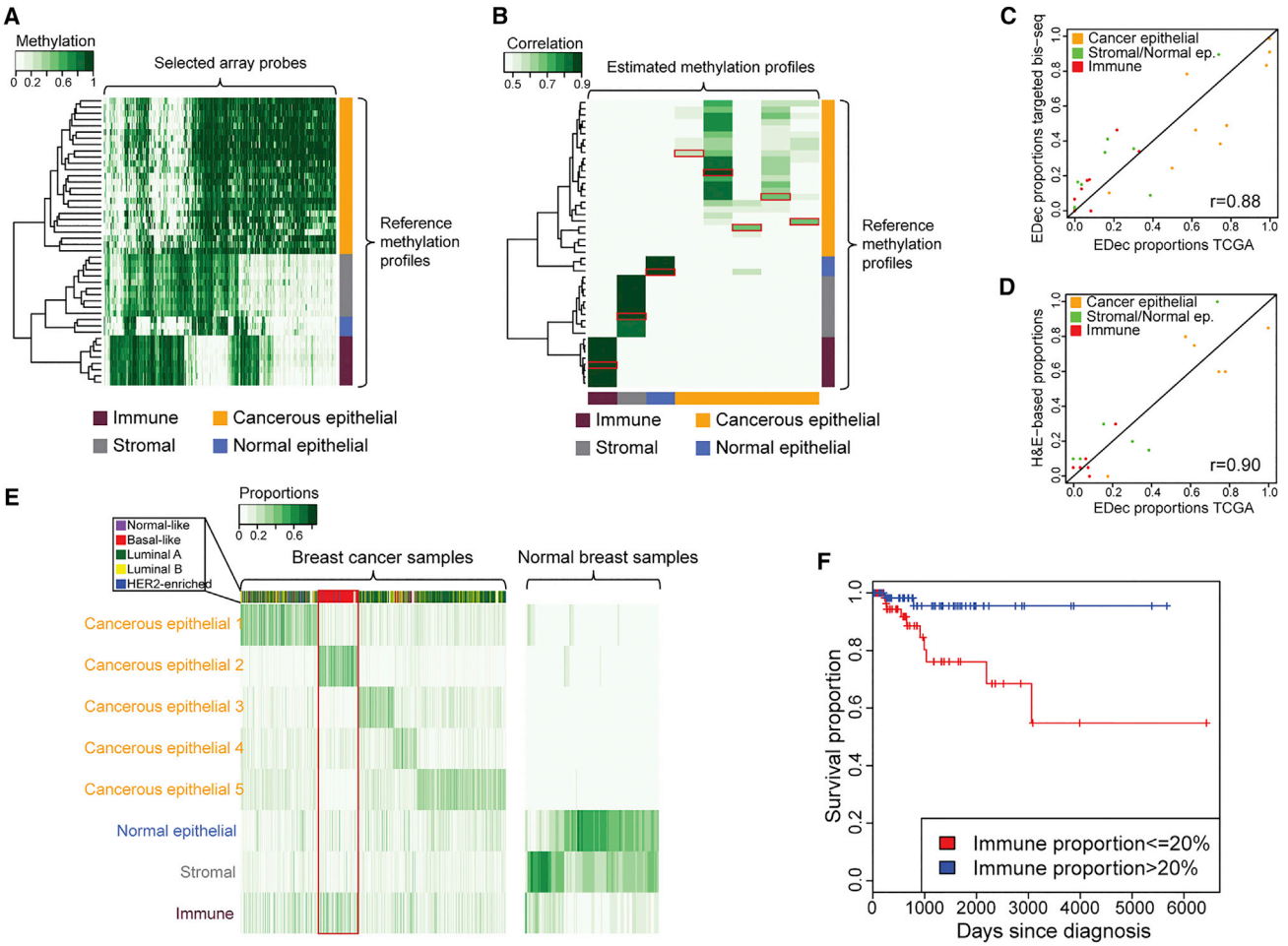


Figure 3. Analysis of DNA Methylation Profiles of Breast Tumors Samples from the TCGA Collection using EDec

(A) Heat map representing the methylation levels over the chosen set of array probes for the reference methylation profiles.

(B) Heat map representing the correlation between the methylation profiles estimated by EDec and the reference methylation profiles. Red boxes indicate the highest correlation for each estimated methylation profile.

(C) Scatterplot of EDec cell type proportion estimates for nine TCGA samples based on targeted bisulfite sequencing (y axis) and microarray (x axis).

(D) Scatterplot between EDec and pathologist (H&E) estimates of proportions of constituent cell types for a subset (six samples) of the TCGA dataset for which H&E staining-based estimates were available.

(E) EDec-estimated proportions of constituent cell types for samples in the TCGA dataset. Side bar represents separation of TCGA cancers samples into PAM50 expression subtypes. The red box highlights the samples best explained by the cancerous epithelial 2 profile, which are almost exclusively classified as basal-like.

(F) Kaplan-Meier plot indicating the significant difference in prognosis (p value < 0.01) for patients within the group of samples best explained by the cancer epithelial 2 profile (red box in [F]; basal-like) with high versus low estimated immune cell type proportion. See also Figures S1 and S2.

the five PAM50 subtypes (Luminal A [523 samples], Luminal B [207], HER2-enriched [78], basal-like [173], normal-like [33]) (Parker et al., 2009), plus normal breast tissue samples (100). We combined the eight EDec-stage 1-estimated proportions (Figure 3E) into the following three cell type fractions: epithelial (including five cancer epithelial and one normal epithelial), stromal, and immune. Proportion estimates for those three cell types were then used in EDec-stage 2 to estimate expression profiles of epithelial, stromal, and immune cell types for each PAM50 subtype and normal breast.

EDec predicts epithelial-specific expression of *ESR1*, *PGR*, and *FOXA1* in Luminal A and Luminal B subtypes (Figure 4A), consistent with previous reports (Toss and Cristofanilli, 2015). Due to poor model fit, as indicated by large error bars, cell-type-specific expression could not be established for a number of genes, *ERBB2* within HER2-enriched tumors being the most conspicuous example. The poor fit of the model for that gene is due to its exceedingly high variance in expression within epithelial cells of this tumor type (Figure S3). We can show through simulations (Supplemental Experimental Procedures) that this effect is mitigated by increasing the number of input breast cancer samples. We note that the large estimated standard error provides a clear signal that cell type-specific expression cannot be established for specific genes, thus preventing erroneous conclusion suggested by high mean values.

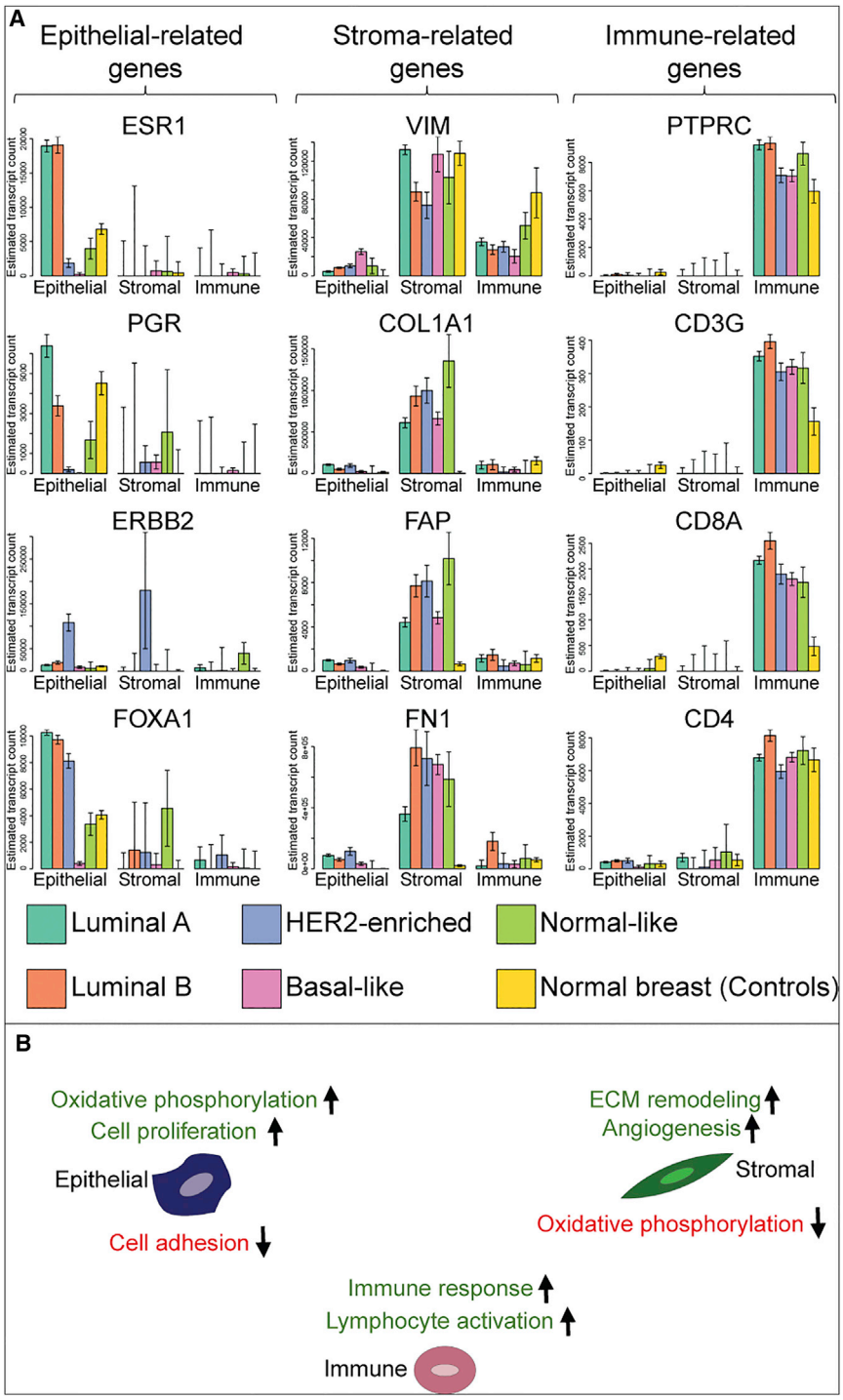


Figure 4. Cell Type-Specific Gene Expression

(A) Bar plots represent the estimated expression profiles of 12 different genes within constituent cell types for each of the breast cancer intrinsic subtypes, as well as for the set of normal breast (control) samples. Error bars represent estimated standard error associated with each cell type specific gene expression estimate.

(B) Summary of main enriched gene sets among upregulated or downregulated genes between cancer and normal breast in each cell type. See also Figures S3 and S4.

EDec predicts stroma-specific expression of vimentin (*VIM*), a general mesenchymal cell marker (Kalluri and Zeisberg, 2006), in normal breast and in all tumor subtypes. Conversely, the stroma-specific expression of *COL1A1*, *FAP*, and *FN1* is observed in tumors, but not in normal breast (Figure 4A). That observation is consistent with the activation of such genes in CAFs, the main constituent of the tumor stroma (Kalluri and Zeisberg, 2006).

EDec correctly predicts immune-specific expression of immune cell markers (PTPRC, CD3G, CD8A, and CD4) in every group of samples (Figure 4A). Note that the $CD8^+$ T cell marker *CD8A* shows significantly higher expression in breast cancers than in normal breasts, consistent with observations that the immune components of breast tumors contains a larger proportion of $CD8^+$ T cells compared to the immune component of normal breasts.

We next compared gene expression profiles of the three tumor-constituent cell types against the profiles of their normal control counterparts. A gene set enrichment analysis (Huang et al., 2009a, 2009b) was then performed on the resulting sets of differentially expressed genes. Figure 4B summarizes the top gene set enrichments for genes upregulated or downregulated in tumor cells compared the normal controls (for full set of gene set enrichments, see Table S1). The terms found to be enriched in each of the sets of differentially expressed genes are consistent with known hallmarks of cancer (sustaining proliferative signaling, activating invasion and metastasis, inducing angiogenesis, deregulating cellular energetics, avoiding immune destruction, etc.) and with the known roles of each cell type within breast tumors (e.g., "extracellular matrix remodeling" genes upregulated specifically in stromal cells and "sustaining inflammation in tumor" category in immune cells) (Hanahan and Weinberg, 2011).

We next sought to further validate EDec-stage 2 predictions of differentially expressed genes against a previously published dataset, in which gene expression profiling was performed on epithelial and stromal components of matched invasive carcinomas and adjacent normal tissue, after LCM (Ma et al., 2009). Despite the fact that the study did not separate out the immune component, focused on the fibrous portion of the stroma (both in

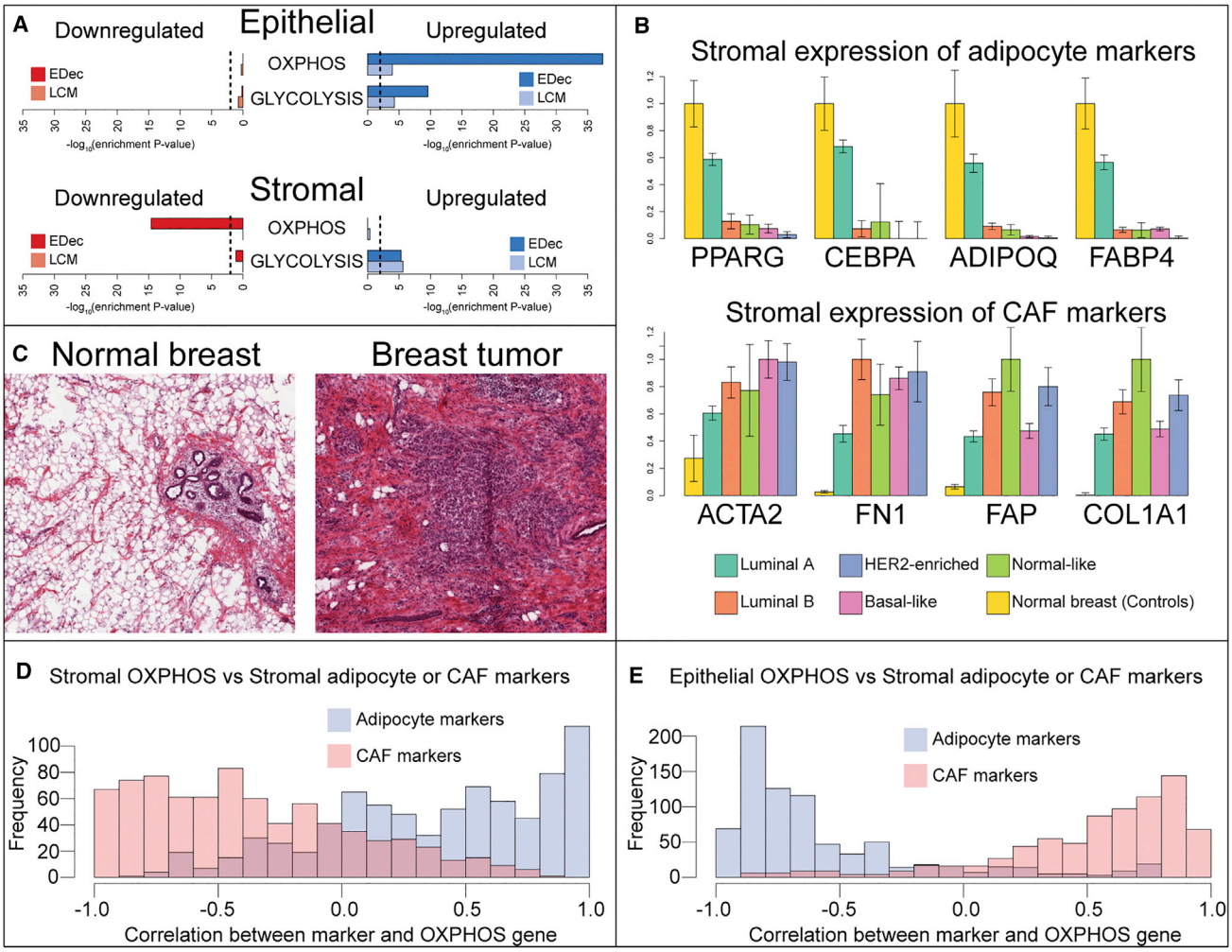


Figure 5. Switch from Adipose to Fibrous Stroma Influences the Metabolic Phenotype of the Tumor

(A) Enrichment of either OXPHOS or GLYCOLYSIS gene sets (hallmark gene sets MSigDB [Liberzon et al., 2015]) among those upregulated or downregulated in epithelial or stromal cells of breast cancer. Cell-type-specific differential expression analysis was performed with either by applying EDec to TCGA dataset, or in the LCM dataset. Dashed lines represent a p value of 0.01.

(B) Estimated stromal expression of either adipocyte or CAF markers across breast cancer subtypes.

(C) Representative H&E staining images of matched tumor and normal breast samples from TCGA (TCGA-BH-A0B2).

(D) Histogram of correlations between stromal expression of OXPHOS genes and stromal expression of marker genes of either adipocyte or CAF across breast cancer subtypes.

(E) Histogram of correlations between epithelial expression of OXPHOS genes and stromal expression of marker genes of either adipocyte or CAF across breast cancer subtypes.

normal breast and breast cancer), and used microarrays to profile expression, we still observe significant overlaps between the differentially expressed genes predicted by EDec and those observed in the LCM dataset (Figure S4). Consistency is observed both for expression differences in epithelial and stromal components.

Switch from Adipose to Fibrous Stroma Supports Oxidative Metabolism in Cancerous Cells

Tumor cells are often more glycolytic than their normal counterparts even in the presence of oxygen. This phenomenon is known as the Warburg effect (Wallace, 2005) and is thought to occur due to the higher anabolic needs of highly proliferative tumor cells (Vander Heiden et al., 2009). Consistent with this phenomenon, we observe enrichment for glycolysis genes among those upregulated in cancer epithelium compared to normal epithelium (Figure 5A). However, contrary to the reduction in mitochondrial activity in cancerous cells predicted by the Warburg effect, we observe strong enrichment for genes involved in oxidative phosphorylation (OXPHOS) among those upregulated in cancer epithelium compared to normal epithelium (Figure 5A). Furthermore, upregulation both glycolysis and OXPHOS genes can be confirmed in comparisons of gene expression profiles of tumor versus normal breast epithelium after LCM (Figure 5A).

The upregulation of both glycolytic and oxidative pathways in cancer cells comes with a demand for nutrients and oxygen, which can be met both by increased angiogenesis and potentially by the support of other cells in the microenvironment. The previously proposed reverse Warburg effect model (Martinez-Outschoorn et al., 2014, 2015; Pavlides et al., 2009) postulates that tumor cells can induce shutdown of oxidative metabolism in the surrounding stromal cells, causing them to reduce oxygen consumption and to secrete high-energy metabolites produced through glycolysis. Those metabolites may then be taken up by cancerous cells to fuel their own oxidative metabolism. Consistent with that model, we observe enrichment for OXPHOS genes among those downregulated in tumor stroma, and for glycolysis genes among those upregulated in the tumor stroma (Figure 5A).

Given that adipocytes have higher rates of mitochondrial activity than fibroblasts (Hofmann et al., 2012; Wilson-Fritch et al., 2003), the observed downregulation of OXPHOS genes in the tumor stroma may reflect the change in stromal composition, from a more adipose (oxidative) stroma in normal breast to a more fibrous (glycolytic) stroma in breast tumors. To determine whether such change indeed occurs, we examined expression levels of adipocyte (PPARG, CEBPA, ADIPOQ, and FABP4) or CAF (ACTA2, FN1, FAP, and COL1A1) markers in the stroma of normal breast and different breast tumor subtypes (Figure 5B). Adipocyte markers are highly expressed in the stroma of normal breast and Luminal A tumors, with negligible expression in other tumor subtypes. CAF markers, in contrast, seem to display the opposite pattern of expression. Such observations are consistent with fibrosis in breast tumors, and with the higher incidence of tumors with adipose stroma among those of the Luminal A subtype (Jung et al., 2015). The change in stromal adipocyte content between normal breast and breast tumor is also apparent in H&E staining slides gathered from matched tumor/normal samples from TCGA (Figure 5C). In the LCM dataset, only the fibrous portion of the stroma was selected for analysis both in normal breast and in breast tumors. Therefore, consistent with the idea that the observed changes in stromal OXPHOS gene expression result from a change from adipose to fibrous stroma, no change in expression of those genes is observed in the LCM dataset (Figure 5A).

We next asked whether the change from adipose to fibrous stroma was associated with a change from oxidative to glycolytic stroma. To examine this, we analyzed the correlation between the expression of either adipocyte or CAF markers in the stroma and the stromal expression of OXPHOS genes across breast cancer subtypes. We observed that, as expected, the stromal expression of most OXPHOS genes had a strong positive correlation with the stromal expression of adipocyte markers, whereas the expression of CAF markers in the stroma was negatively correlated with OXPHOS genes (Figure 5D).

The reverse Warburg effect model predicts that a glycolytic stroma associates with oxidative cancerous epithelial cells, whereas an oxidative stroma would be associated with more glycolytic tumor cells. Given that a fibrous stroma seems to be more glycolytic than an adipose one, we hypothesized that a change from adipose to fibrous stroma would associate with a change from glycolytic to oxidative cancerous epithelium. We therefore analyzed the degree of correlation between the expression of either adipocyte or CAF markers in the stroma and the expression of OXPHOS genes in the epithelial fraction across breast cancer subtypes. Stromal expression of CAF markers was indeed positively correlated with epithelial OXPHOS gene expression, whereas adipocyte marker expression in the stroma was negatively correlated with OXPHOS gene expression in the epithelial fraction (Figure 5E). Interestingly, the stromal expression of CAV1, a gene whose low expression in breast cancer stroma is known to associate with negative prognosis and with tumors with reverse Warburg metabolism (Martinez-Outschoorn et al., 2014, 2015; Pavlides et al., 2009), is strongly correlated with the expression of adipocyte markers in the stroma (mean $r = 0.97$), providing further support for the hypothesis that stromal adiposity associates negatively and the stromal fibroblast content associates positively with the reverse Warburg pattern of metabolism.

DISCUSSION

The EDec method provides accurate platform-independent estimation of cell type proportions, DNA methylation profiles, and gene expression profiles of constituent cell types. By significantly relaxing the dependence on reference methylation profiles of constituent cell types compared to previous methods (Houseman et al., 2012), EDec enables deconvolution of complex tumor tissues where highly accurate references are unavailable. In contrast to reference-free methods (Houseman et al., 2014, 2016; Rahmani et al., 2016; Zheng et al., 2014; Zou et al., 2014), EDec's indirect use of surrogate references greatly assisted in the interpretation of deconvolution results, allowing us to uncover more complex biological patterns than possible by applying other reference-free deconvolution techniques. Furthermore, unlike previous methylation-based deconvolution methods, EDec does not require that each cell type be explained by a single component (e.g., cancerous epithelial fraction in the TCGA dataset was modeled by five different components), thus making it possible to model the full diversity of cancerous epithelial cells. Despite such methodological advances, we note that the current tissue models obtained by EDec still only approximate the full complexity of breast tumors. For example, more detailed deconvolution of individual components of the stromal and immune fractions would likely yield additional biological insights.

By addressing the confounding issue of tissue heterogeneity, EDec enables the comparison of tumors of various cell type compositions based on inferred molecular profiles of constitutive cancer epithelial cells and also the comparisons between cancer cell fractions of tumors and experimentally more tractable cell line models. EDec reveals that methylome profiles of breast cancer cells are distinct from those of normal epithelial cell types, and that they can be mapped to specific groups of cancer cell lines. We also observe that cancerous cells of basal-like tumors have particularly distinct cellular identity as indicated by their distinct methylation profiles.

By providing information about the epigenomic and transcriptomic states of both cancerous epithelial and non-epithelial tumor cells, the method enables the study of heterotypic interactions driving tumor progression. The most striking pattern that emerged from our analyses is metabolic coupling between

epithelium and stroma that seems to be related to the degree of adiposity of the stroma. Specifically, upregulation of both glycolysis and OXPHOS in cancerous epithelial cells supports the idea that, despite the long-postulated Warburg effect, cancer cells in breast tumors still upregulate their energy production through OXPHOS in comparison to normal cells (Zu and Guppy, 2004). Furthermore, the switch from adipose to fibrous stroma leads to lower stromal mitochondrial activity, which in turn seems to support upregulation of OXPHOS in cancerous epithelial cells. Our findings therefore refine the reverse Warburg effect model (Martinez-Outschoorn et al., 2014, 2015; Pavlides et al., 2009) by showing that it may be mediated by changes in cell type composition of tumor stroma. It is tempting to speculate that the differences in stroma composition across tumor subtypes may be related to a different capacity of distinct tumor types to induce the conversion of adipocytes into fibroblasts (Bochet et al., 2013; Dirat et al., 2011), which would be more supportive of reverse Warburg metabolism. Despite these encouraging results, which are largely confirmed by expression profiling of microdissected tumors, further experiments focusing on protein and metabolite levels in different tumor cell types will be needed to conclusively confirm this model.

In conclusion, EDec reveals layers of biological information about distinct cell types within solid tumors and about their heterotypic interactions that were previously inaccessible at such large scale due to tissue heterogeneity. EDec improves on previous methods by employing a data-driven approach that makes indirect use of reference profiles of constituent cell types and adequately models the variability of methylation profiles across different cancer cells. We note that EDec is a general technique and could potentially be applied to different types of tumors and other complex non-tumor tissues. However, such applications would involve new feature selection with a set of references appropriate for that tissue, and would need to be validated. In addition to the method itself, we have also developed a "deconvoluted breast cancer" data resource for breast tumors and normal breast tissues within the TCGA collection (http://genboree.org/theCommons/projects/edec). This resource can now be further explored by the community to derive or test new hypotheses.

EXPERIMENTAL PROCEDURES

ThunderStorm BS-Seq Assay and Breast Cancer Target Panel

A set of 1,000 target regions of around 300 bp in length were preselected for targeted bisulfite sequencing based on previous reports of their involvement in breast cancer biology (Table S2). Of the 1,000 genomic regions, 149 were selected based on cell type-specific methylation based on Roadmap Epigenomics reference DNA methylation profiles (Kundaje et al., 2015).

Primer pairs designed to specifically amplify each selected target region were designed by RainDance Technologies. The ThunderStorm BS-seq assay using that set of primer pairs was performed at RainDance Technologies according to the manufacturer's specification. In summary, that assay uses a microfluidic chip to perform multiplex amplification of bisulfite-treated DNA using the set of primers designed to amplify the selected set of genomic regions. This step is followed by sequencing of PCR product. Read mapping and methylation level calling was performed using Bismark (Krueger and Andrews, 2011). Target regions were sequenced on average to 200× coverage. For all subsequent analyses, DNA methylation levels for all CpGs overlapping each of the target regions were averaged, giving an average methylation value for each region of interest. For eight of the breast cancer samples profiled using this assay, 450k arrays were also performed by the TCGA group. We observed over 0.9 correlation between methylation levels measured by both platforms over the 614 regions overlapping 450k array probes for all samples analyzed.

TCGA Data Processing

Methylation Array Data

The breast cancer TCGA DNA methylation data were generated using either the Infinium HumanMethylation450 BeadChip (450k array) or the Infinium HumanMethylation27 BeadChip (27k array). We used the TCGA Assembler (Zhu et al., 2014) to download level 3 data (fully processed) for all 27k and 450k profiles. Because most 27k probes are present in the 450k array, we merged the two datasets and included only overlapping probes in our analysis. We also removed any probe with a detection p value less than 0.05 in at least one sample, those that overlapped known SNPs, and those that were previously reported as cross reactive (Chen et al., 2013). The final number of probes passing these criteria was 17,907. We also corrected for platform biases using an Empirical Bayes-based approach (ComBat) (Johnson et al., 2007), implemented in the SVA package in R (Leek et al., 2015).

RNA-Sequencing Data

TCGA Assembler (Zhu et al., 2014) was used to download normalized (RNA sequencing [RNA-seq] v2 – RNA-seq by expectation maximization) gene transcript abundance measurements from the TCGA database. PAM50 classification (Parker et al., 2009) based on RNA-seq for 1,030 breast cancer samples generated by the TCGA Analysis Working Group were obtained from the University of California, Santa Cruz (UCSC), Cancer Genomics Browser (Goldman et al., 2013). Of the TCGA breast cancer samples that had RNA-sequencing data and associated PAM50 classification, 1,005 also had DNA methylation data. For normal breast samples, 100 had both DNA methylation and RNA-sequencing data. Therefore, the final set of RNA-sequencing samples contained 1,105 samples.

Code and Dataset Availability

The EDec software is available as an R package. It can be downloaded from https://github.com/BRL-BCM/EDec. Documentation and usage examples are also available on that same page. All datasets associated with this publication can be found at http://genboree.org/theCommons/projects/edec. 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.

ACCESSION NUMBERS

The accession number for the targeted bisulfite sequencing data reported in this paper is GEO: GSE87297.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.10.057.

AUTHOR CONTRIBUTIONS

Conceptualization, V.O., R.J.H., D.N.B., M.E.R., M.L.S., S.O., W.M.W., V.D.G., A.V.L., A.M.; Methodology, V.O., R.J.H., D.N.B, M.E.R., M.L.S., A.V.L., A.M.; Software, V.O.; Validation, V.O., R.J.H., D.N.B.; Formal Analysis, V.O.; Investigation, V.O., R.J.H., D.N.B., A.V.L., A.M.; Resources, A.V.L., M.L.S., A.M.; Data Curation, V.O., D.N.B., R.J.H., R.Y.P.; Writing – Original Draft, V.O.; Writing – Review & Editing, V.O., R.J.H., D.N.B., M.E.R., M.L.S., A.V.L., A.M.; Visualization, V.O.; Supervision, A.V.L., A.M.; Project Administration, V.O., M.E.R. A.V.L., A.M.; Funding Acquisition, A.V.L., M.L.S., A.M.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the Common Fund of the NIH (Roadmap Epigenomics Program, grant U01 DA025956) (to A.M.) and the

Scientific Advisory Council award from Susan G. Komen for the Cure and the Hillman Foundation Fellow award (to A.V.L.). This study used University of Pittsburgh Cancer Institute (UPCI) Cancer Tissue and Research Pathology services supported by NIH Award P30CA047904. We also acknowledge the support of the Health Sciences Tissue Bank (HSTB) at the University of Pittsburgh. We thank Christina Kline and the other HSTB team members. We would also like to thank the UPCI and the clinicians, staff, and patients at UPMC for making this study possible.

Received: January 26, 2016
Revised: July 28, 2016
Accepted: September 26, 2016
Published: November 15, 2016

REFERENCES

Adams, S., Gray, R.J., Demaria, S., Goldstein, L., Perez, E.A., Shulman, L.N., Martino, S., Wang, M., Jones, V.E., Saphner, T.J., et al. (2014). Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. J. Clin. Oncol. *32*, 2959–2966.

Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity. Nat. Commun. *6*, 8971.

Bochet, L., Lehuédé, C., Dauvillier, S., Wang, Y.Y., Dirat, B., Laurent, V., Dray, C., Guiet, R., Maridonneau-Parini, I., Le Gonidec, S., et al. (2013). Adipocyte-derived fibroblasts promote tumor progression and contribute to the desmoplastic reaction in breast cancer. Cancer Res. *73*, 5657–5668.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Chen, Y.A., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., and Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics *8*, 203–209.

Coussens, L.M., Zitvogel, L., and Palucka, A.K. (2013). Neutralizing tumor-promoting chronic inflammation: a magic bullet? Science *339*, 286–291.

Debey, S., Schoenbeck, U., Hellmich, M., Gathof, B.S., Pillai, R., Zander, T., and Schultze, J.L. (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. Pharmacogenomics J. *4*, 193–207.

Dirat, B., Bochet, L., Dabek, M., Daviaud, D., Dauvillier, S., Majed, B., Wang, Y.Y., Meulle, A., Salles, B., Le Gonidec, S., et al. (2011). Cancer-associated adipocytes exhibit an activated phenotype and contribute to breast cancer invasion. Cancer Res. *71*, 2455–2465.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., et al. (2008). Stromal gene expression predicts clinical outcome in breast cancer. Nat. Med. *14*, 518–527.

Gaujoux, R., and Seoighe, C. (2012). Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. Infect. Genet. Evol. *12*, 913–921.

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

Goldman, M., Craft, B., Swatloski, T., Ellrott, K., Cline, M., Diekhans, M., Ma, S., Wilks, C., Stuart, J., Haussler, D., and Zhu, J. (2013). The UCSC Cancer Genomics Browser: update 2013. Nucleic Acids Res. *41*, D949–D954.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hofmann, A.D., Beyer, M., Krause-Buchholz, U., Wobus, M., Bornhäuser, M., and Rödel, G. (2012). OXPHOS supercomplexes as a hallmark of the mitochondrial phenotype of adipogenic differentiated human MSCs. PLoS One *7*, e35160.

Houseman, E.A., and Ince, T.A. (2014). Normal cell-type epigenetics and breast cancer classification: a case study of cell mixture-adjusted analysis of DNA methylation data from tumors. Cancer Inform. *13* (*Suppl 4*), 53–64.

Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86.

Houseman, E.A., Molitor, J., and Marsit, C.J. (2014). Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics *30*, 1431–1439.

Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., and Marsit, C.J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics *17*, 259.

Hu, M., Yao, J., Cai, L., Bachman, K.E., van den Brûle, F., Velculescu, V., and Polyak, K. (2005). Distinct epigenetic changes in the stromal cells of breast cancers. Nat. Genet. *37*, 899–905.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Jung, Y.Y., Lee, Y.K., and Koo, J.S. (2015). Expression of cancer-associated fibroblast-related proteins in adipose stroma of breast cancer. Tumour Biol. *36*, 8685–8695.

Kalluri, R., and Zeisberg, M. (2006). Fibroblasts in cancer. Nat. Rev. Cancer *6*, 392–401.

Komori, H.K., LaMere, S.A., Torkamani, A., Hart, G.T., Kotsopoulos, S., Warner, J., Samuels, M.L., Olson, J., Head, S.R., Ordoukhanian, P., et al. (2011). Application of microdroplet PCR for large-scale targeted bisulfite sequencing. Genome Res. *21*, 1738–1745.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L., and Luthi-Carter, R. (2011). Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat. Methods *8*, 945–947.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Leek, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., and Storey, J.D. (2015). sva: Surrogate Variable Analysis. R package, version 3.19.0, https://www.bioconductor.org/packages/release/bioc/html/sva.html.

Li, Y., and Xie, X. (2013). A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. BMC Bioinformatics *14* (*Suppl 5*), S11.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Liu, S., Foulkes, W.D., Leung, S., Gao, D., Lau, S., Kos, Z., and Nielsen, T.O. (2014). Prognostic significance of FOXP3$^+$ tumor-infiltrating lymphocytes in breast cancer depends on estrogen receptor and human epidermal growth factor receptor-2 expression status and concurrent cytotoxic T-cell infiltration. Breast Cancer Res. *16*, 432.

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

Martinez-Outschoorn, U.E., Lisanti, M.P., and Sotgia, F. (2014). Catabolic cancer-associated fibroblasts transfer energy and biomass to anabolic cancer cells, fueling tumor growth. Semin. Cancer Biol. *25*, 47–60.

Martinez-Outschoorn, U.E., Sotgia, F., and Lisanti, M.P. (2015). Caveolae and signalling in cancer. Nat. Rev. Cancer *15*, 225–237.

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

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Paul, D.S., Guilhamon, P., Karpathakis, A., Butcher, L.M., Thirlwell, C., Feber, A., and Beck, S. (2014). Assessment of RainDrop BS-seq as a method for large-scale, targeted bisulfite sequencing. Epigenetics *9*, 678–684.

Pavlides, S., Whitaker-Menezes, D., Castello-Cros, R., Flomenberg, N., Witkiewicz, A.K., Frank, P.G., Casimiro, M.C., Wang, C., Fortina, P., Addya, S., et al. (2009). The reverse Warburg effect: aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. Cell Cycle *8*, 3984–4001.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G., Eskin, E., Zou, J., and Halperin, E. (2016). Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat. Methods *13*, 443–445.

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

Tlsty, T.D., and Coussens, L.M. (2006). Tumor stroma and regulation of cancer development. Annu. Rev. Pathol. *1*, 119–150.

Toss, A., and Cristofanilli, M. (2015). Molecular characterization and targeted therapeutic approaches in breast cancer. Breast Cancer Res. *17*, 60.

Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science *324*, 1029–1033.

Venet, D., Pecasse, F., Maenhaut, C., and Bersini, H. (2001). Separation of samples into their constituents using gene expression data. Bioinformatics *17* (*Suppl 1*), S279–S287.

Wallace, D.C. (2005). Mitochondria and cancer: Warburg addressed. Cold Spring Harb. Symp. Quant. Biol. *70*, 363–374.

Wilson-Fritch, L., Burkart, A., Bell, G., Mendelson, K., Leszyk, J., Nicoloro, S., Czech, M., and Corvera, S. (2003). Mitochondrial biogenesis and remodeling during adipogenesis and in response to the insulin sensitizer rosiglitazone. Mol. Cell. Biol. *23*, 1085–1094.

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

Zheng, X., Zhao, Q., Wu, H.J., Li, W., Wang, H., Meyer, C.A., Qin, Q.A., Xu, H., Zang, C., Jiang, P., et al. (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biol. *15*, 419.

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

Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat. Methods *11*, 599–600.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. Nat. Methods *11*, 309–311.

Zu, X.L., and Guppy, M. (2004). Cancer metabolism: facts, fantasy, and fiction. Biochem. Biophys. Res. Commun. *313*, 459–465.

Cell Reports, Volume *17*

Supplemental Information

Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types

Vitor Onuchic, Ryan J. Hartmaier, David N. Boone, Michael L. Samuels, Ronak Y. Patel, Wendy M. White, Vesna D. Garovic, Steffi Oesterreich, Matt E. Roth, Adrian V. Lee, and Aleksandar Milosavljevic

Supplemental Figures

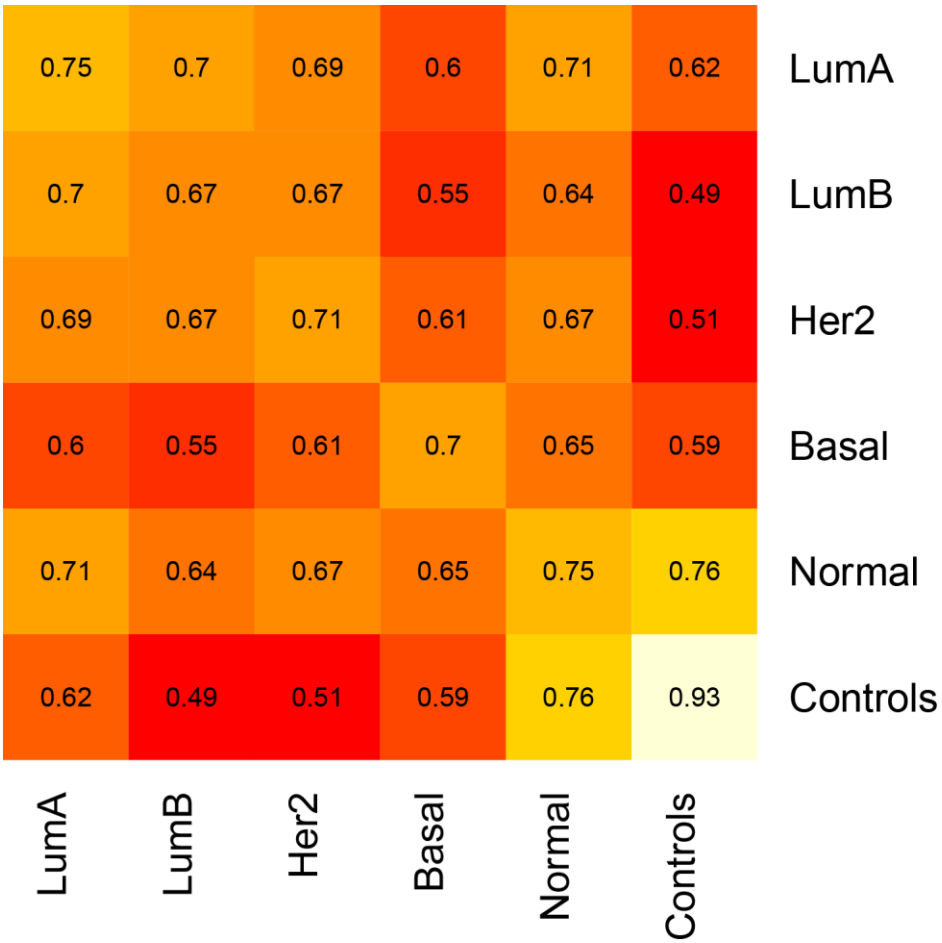


Figure S1. Related to Figure 3. Heterogeneity of methylation profiles of epithelial fraction within and between breast cancer subtypes. Values represent average spearman correlation between methylation profiles of epithelial fraction of all pairs of samples from the subtypes represented in the corresponding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

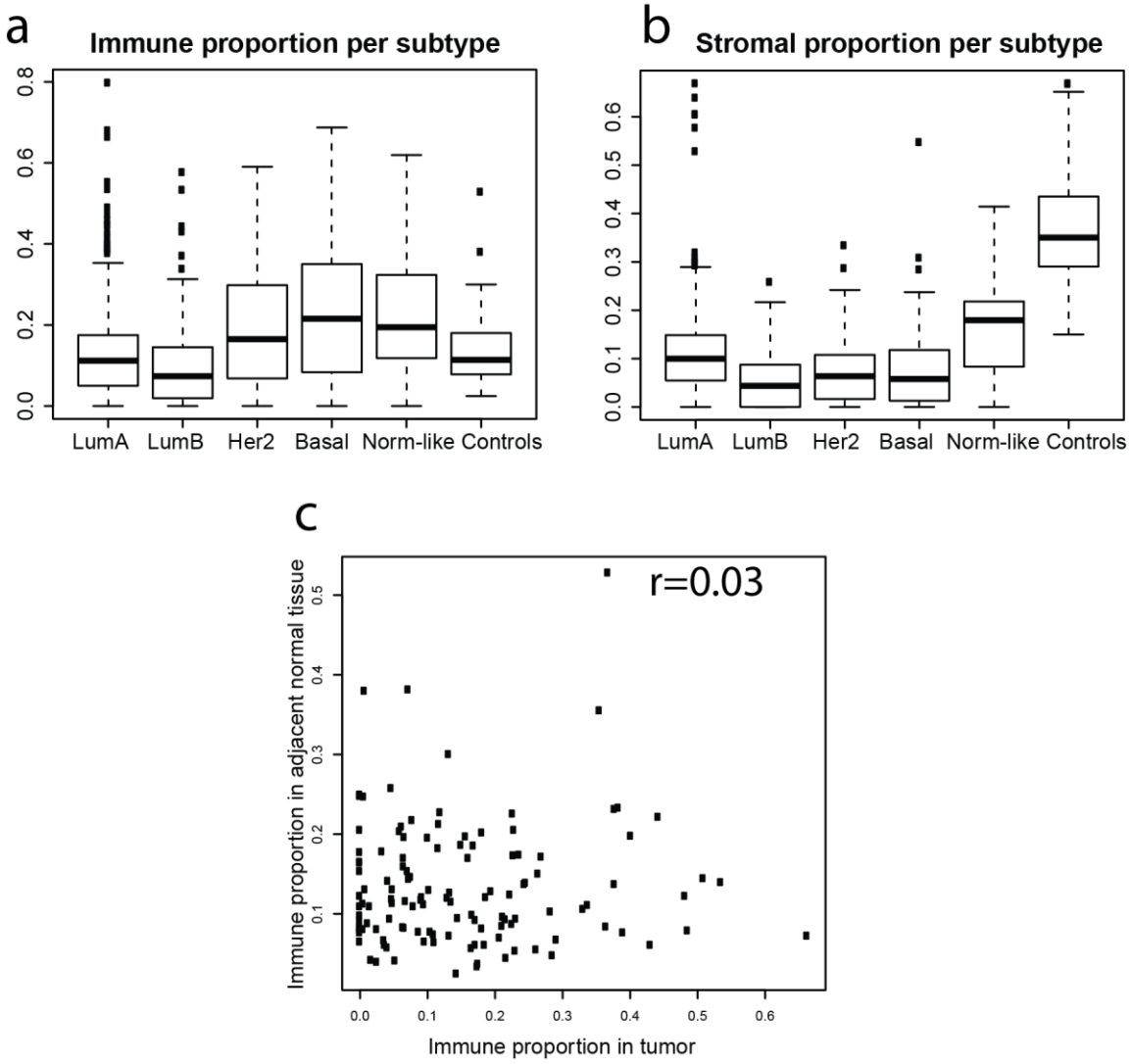


Figure S2. Related to Figure 3. Distribution of proportions of immune and stromal cell types across TCGA samples. a. Box plots of proportions of immune cells across cancer subtypes. **b.** Box plots of proportions of stromal cells across cancer subtypes. **c.** Scatterplot comparing the level of immune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

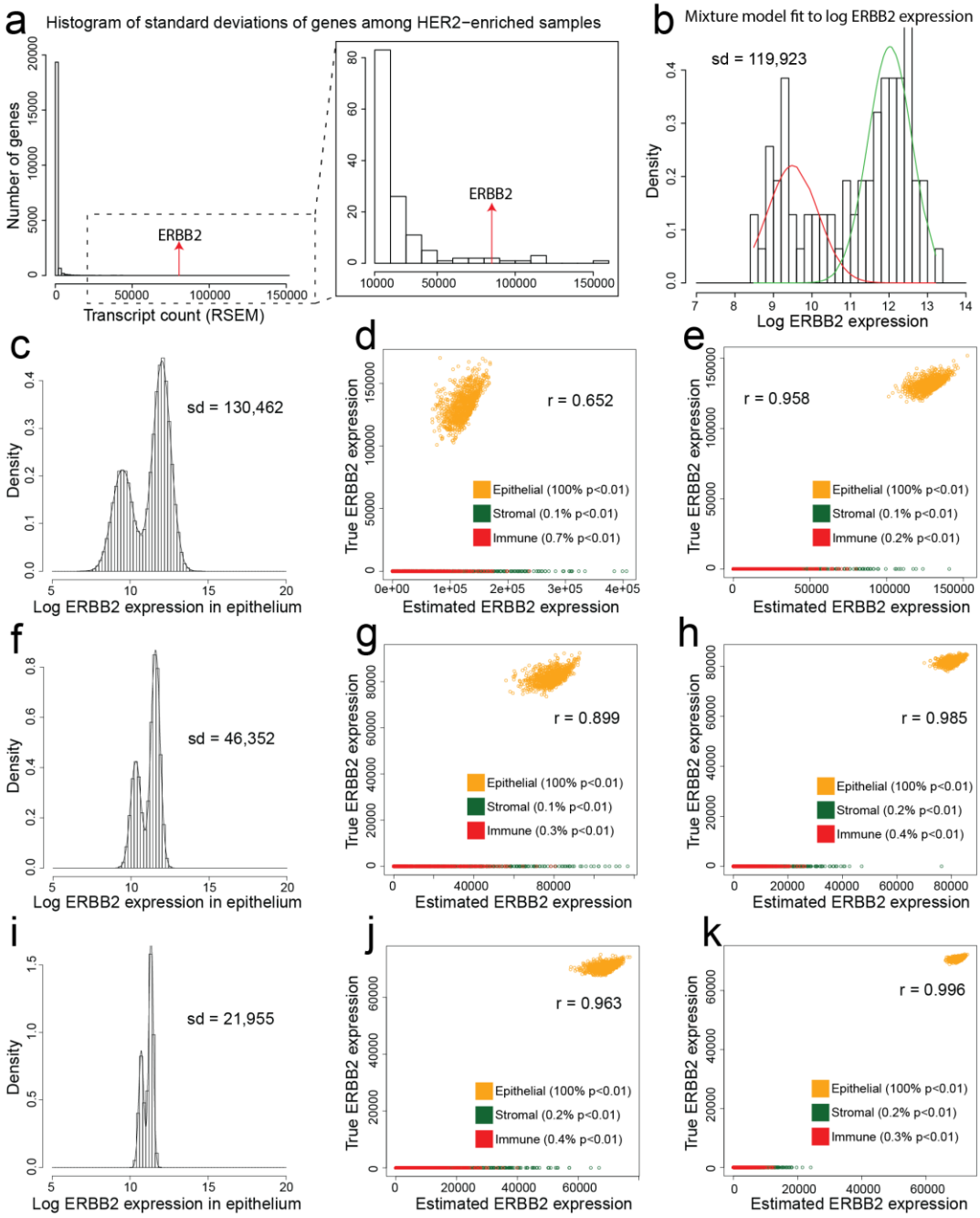


Figure S3. Related to Figure 4. Analysis of ERBB2 expression in HER2-enriched tumors. a. Histogram of standard deviation of expression of each gene among HER2-enriched tumor samples. **b.** Gaussian mixture model fit to log of ERBB2 expression values in epithelial cells of HER2-enriched tumors. **c,f,i.** 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**.**

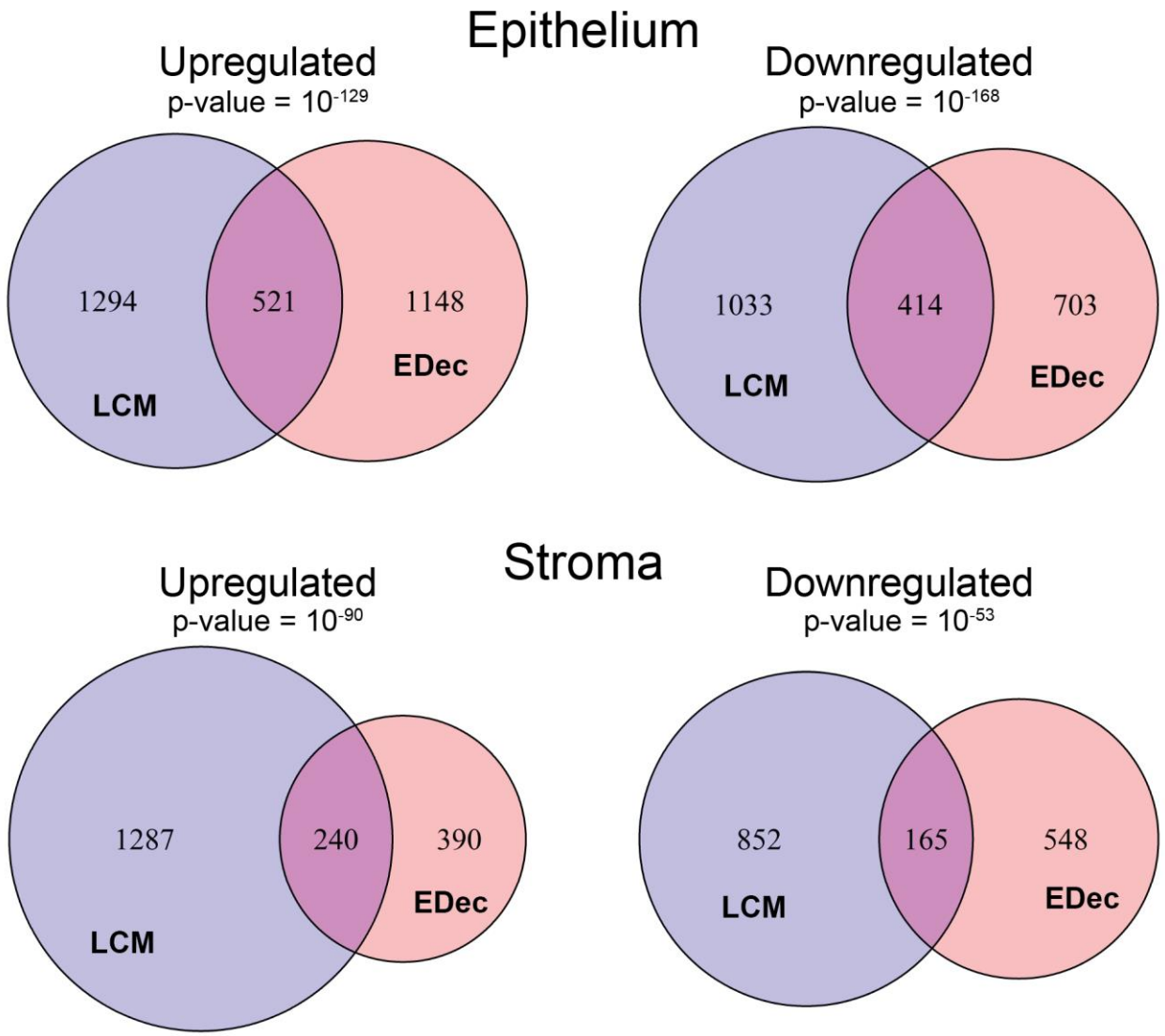


Figure S4. Related to Figure 4. Comparison of differentially expressed genes identified by EDec and LCM. Venn diagrams representing the overlap between differentially expressed genes in laser capture microdissection dataset and EDec analysis of TCGA dataset. P-values were computed using the hypergeometric 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).

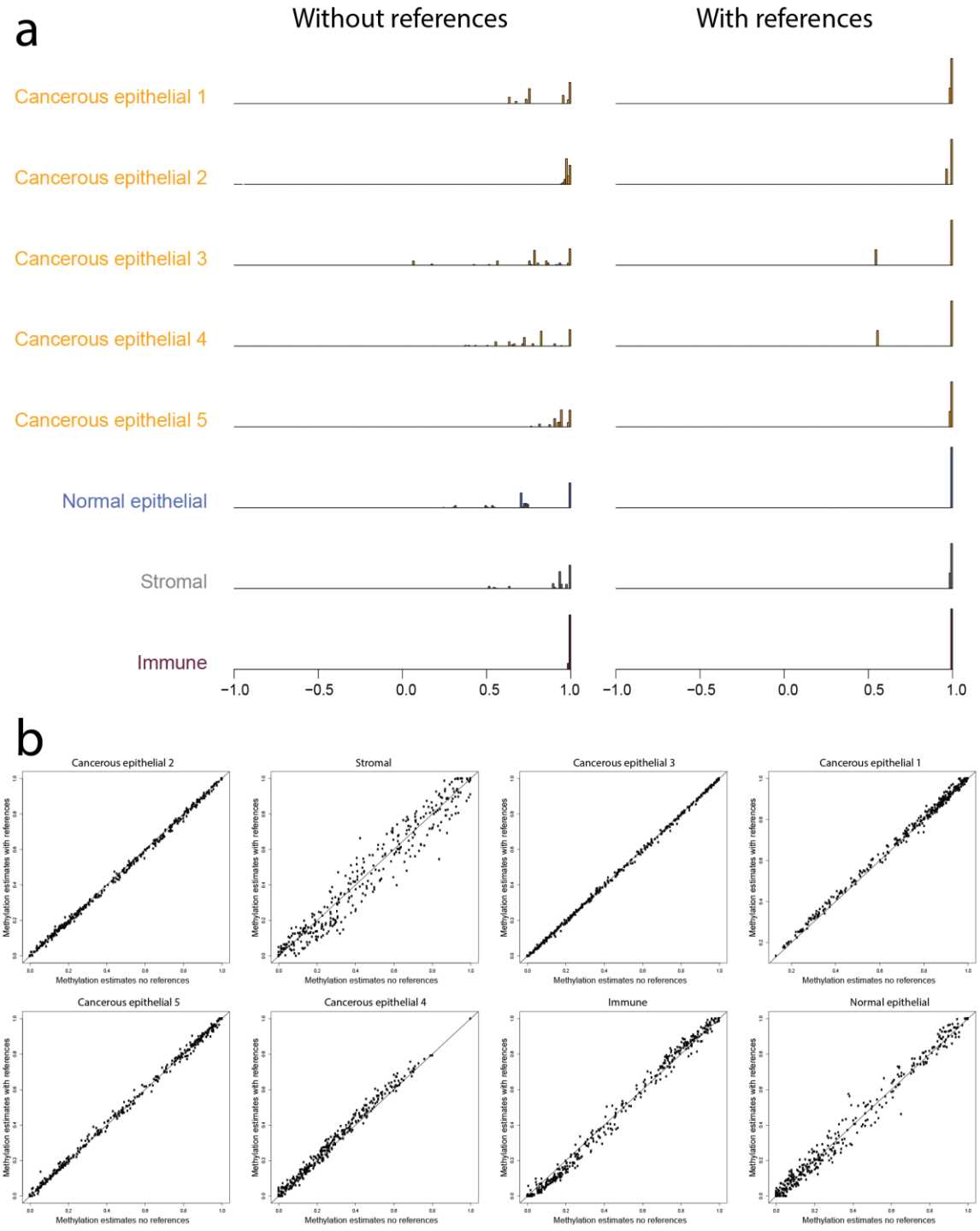


Figure S5. Related to Figure 3. Stability of deconvolution with or without the addition of reference methylation profiles to the TCGA dataset. a. Histograms of correlations between pairs of methylation profiles estimated for each of the 8 cell types in the 20 runs of EDec with references and in the 20 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.