Deconvoluting complex tissues for expression quantitative trait locus-based analyses

Ji-Heui Seo¹, Qiyuan Li¹,³,⁴, Aquila Fatima², Aron Eklund⁴, Zoltan Szallasi⁴,⁵,⁶, Kornelia Polyak¹,⁶, Andrea L. Richardson²,⁶,⁷ and Matthew L. Freedman¹,³,⁶

¹Department of Medical Oncology, The Center for Functional Cancer Epigenetics, and ²Department of Cancer Biology, Dana Farber Cancer Institute, Boston, MA 02215, USA
³Program in Medical and Population Genetics, The Broad Institute, Cambridge, MA 02142, USA
⁴Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark
⁵Children’s Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology (CHIP@HST), Boston, MA, USA
⁶Harvard Medical School, Boston, MA, USA
⁷Department of Pathology, Brigham and Women’s Hospital, Boston, MA 02115, USA

Breast cancer genome-wide association studies have pinpointed dozens of variants associated with breast cancer pathogenesis. The majority of risk variants, however, are located outside of known protein-coding regions. Therefore, identifying which genes the risk variants are acting through presents an important challenge. Variants that are associated with mRNA transcript levels are referred to as expression quantitative trait loci (eQTLs). Many studies have demonstrated that eQTL-based strategies provide a direct way to connect a trait-associated locus with its candidate target gene. Performing eQTL-based analyses in human samples is complicated because of the heterogeneous nature of human tissue. We addressed this issue by devising a method to computationally infer the fraction of cell types in normal human breast tissues. We then applied this method to 13 known breast cancer risk loci, which we hypothesized were eQTLs. For each risk locus, we took all known transcripts within a 2 Mb interval and performed an eQTL analysis in 100 reduction mammoplasty cases. A total of 18 significant associations were discovered (eight in the epithelial compartment and 10 in the stromal compartment). This study highlights the ability to perform large-scale eQTL studies in heterogeneous tissues.

1. Introduction

Genome-wide association studies (GWAS) have discovered thousands of variants that are associated with hundreds of traits, including breast cancer. Dozens of common alleles are strongly associated with breast cancer risk [1]. Similar to other common traits, the majority of breast cancer risk loci are outside of known protein-coding regions. Given our rudimentary understanding of the non-protein-coding regions of the genome, a key challenge is to connect risk loci with the genes that they are influencing.

Performing expression quantitative trait locus (eQTL) analysis is a straightforward way to link risk loci with candidate genes. eQTLs are polymorphisms that correlate with mRNA transcript levels. Several studies strongly suggest that GWAS loci are enriched for eQTLs and regulatory elements [2,3]. Empirical data confirm that this strategy can prioritize candidate genes for downstream functional analysis [2,4–9].

Tissue selection is an important factor to consider for eQTL analysis. The effect of a risk locus on a transcript may only be revealed in a tissue-specific manner. For example, a risk locus may affect a gene product in stromal tissue, which is secreted and acts upon the target epithelial cell resulting in transformation.
Studying cancer risk loci presents an additional challenge—investigations can be performed in normal and/or tumour tissue. Investigations have been performed in both normal and tumour tissue [4,6,10,11]. An advantage of studying normal tissue is that the risk loci, by definition, are acting at an early stage. A disadvantage of using normal tissue is that most solid tissues are complex mixtures of various cell types. Because cell types differ in their proportions and expression patterns, it is important to account for this source of variability. A number of techniques are available to purify heterogeneous tissue, including laser capture microdissection and flow sorting. Unfortunately, these techniques make it difficult to process hundreds to thousands of samples. While it is difficult to accurately estimate power for eQTL-based studies, it is clear that large sample sizes are critically important [12]. Thus, a tension exists between power and purity.

In this study, we performed an eQTL analysis using known breast cancer risk loci and measured gene expression in normal breast tissue. Because normal breast tissue is highly heterogeneous, we developed a model to computationally estimate the proportion of cell types in human breast tissue. We then applied this model to our sample set to evaluate eQTL associations between breast cancer risk loci and gene-expression levels.

2. Results
To address tissue heterogeneity, we modelled normal breast tissue as comprising four cell types: adipocytes, epithelial, inflammatory and stromal. In order to estimate the proportion of each cell type, cell-type-specific markers were selected (§4). Both the markers and all of the transcripts selected for this study were measured in a set of highly purified populations representing the four cell types (§4). Using this information, we were able to estimate: (i) the proportion of each cell type for each sample, and (ii) the expression level of a given gene for each cell type (figure 1 and §4). As expected, the ranges of the estimated proportions of cells varied dramatically (figure 2).

To evaluate the model, we compared our results with an independent dataset in which serial analysis of gene expression (SAGE) was performed on eight purified epithelial cell populations and six purified stromal populations [13]. We selected 59 genes that were common between the SAGE data and our data. In the SAGE data, these 59 genes showed significant fold changes (FDR < 0.05) between epithelial and stromal cells. In the SAGE data, we classified 38 genes as being upregulated (i.e. demonstrated a greater expression in epithelial cells compared with stromal) and 21 as downregulated sets (i.e. demonstrated a greater expression in stromal cells compared with epithelial). In our samples, the fold change for each of the 59 genes was calculated based on the inferred expression levels in the epithelial and stromal compartments. Using the SAGE data as a benchmark, we compared the concordance of the direction of the fold changes between the SAGE data and our data. Our inferred data correctly predicted 31 (83.8%) upregulated genes and 13 (61.9%) downregulated genes yielding an overall accuracy of 74.6 per cent. The concordance between the SAGE and our inferred datasets was statistically significant ($p = 0.0013$).

The eQTL strategy consisted of evaluating 13 breast cancer risk single nucleotide polymorphisms (SNPs) for association with RNA expression levels from 101 protein-coding genes. For each risk locus, we tested for the correlation between the number of risk alleles carried and the mRNA expression level of any annotated protein-coding transcript within a 2 Mb window centred on the risk variant (see the electronic supplementary material, table S1). The ability to computationally estimate the proportion of each cell type made it possible to perform eQTL-based analyses in a cell-type-specific manner. We performed analyses for both epithelial and stromal cell types as most samples contained these cell types.

Out of the 13 tested loci, nine demonstrated a significant association with at least one transcript that was within a 2 Mb window ($p < 0.05$). Six variants/eQTLs were significant in the epithelial tissue compartment, and four variants/eQTLs were significant in the stromal compartment (see table 1 and electronic supplementary material, figures S1 and S2). Three eQTL–target gene pairs (representing one risk locus) were significant across both epithelial and stromal cell types.

3. Discussion
The observation that transcripts are under genetic control presents a direct method to identify candidate genes of non-protein-coding trait-associated alleles discovered through GWAS [14]. For cancer risk, investigators can examine eQTL relationships in normal and/or tumour tissues. eQTLs have been discovered using both tissue types, however, more data are necessary before being able to definitively recommend studying one tissue over the other [4,6,10]. For now, we suggest conducting studies in both normal and tumour tissues. Pursuing eQTLs in normal tissue is intuitively appealing because risk variants act at the earliest stages of tumour pathogenesis. Obtaining large sample sizes of purified tissue, however, remains a major bottleneck. In this study, we developed a method to estimate the proportion of cell types and gene levels computationally within each cell type for normal breast tissue samples.

Two types of methods have been previously described to deconvolute cell-type-specific gene expression from heterogeneous tissue samples. Data-driven methods rely solely on matrix decomposition algorithms and/or classification based on expression signatures is sensitive to the noise in the data [15,16]. On the other hand, tissue purification (e.g. laser capture microdissection and flow sorting) is costly and time consuming, and, therefore, less practical for large-scale studies [17]. We described an alternative strategy in which we used a set of highly purified cell populations to represent the cell types in normal human breast tissue. A set of marker genes ($n = 11$) and all of the genes under study ($n = 101$) were measured in each of these purified populations. The same sets of genes were also measured in each of the normal breast tissue samples. Using the information from the purified samples, we were able to predict the cell composition and to estimate the expression profile of each tissue type represented in the normal breast sample.

Only a few studies have performed eQTL-based analyses for breast cancer risk loci. Perhaps the most widely studied eQTL relationship in breast cancer is between the 10q risk locus and FGFR2. Studies have been performed in both normal and tumour tissues. Only one study conducted in
tumours revealed an association [18]. Notably, two prior studies demonstrated an association in normal tissue with one of the studies showing in association in fibroblasts and the other in normal unpurified breast tissue [19,20]. Our data did not show an association with \textit{FGFR2}. Larger studies evaluating all \textit{FGFR2} isoforms in the different cell types will help to further clarify this complex association.

Other eQTL–target gene associations are interesting, but will require independent validation. The imprinted \textit{H19} non-protein-coding gene showed a significant association with the chromosome 11 risk locus in epithelial cells. Up- and down-regulation of \textit{H19} has been shown to play a role in breast cell based tumour assays, including proliferation and anchorage independence [21–23]. Given the fact that \textit{H19} is an imprinted gene, it is interesting that \textit{H19} demonstrated an association because we did not specifically test for parent-of-origin effects. Testing for parent-of-origin effects requires identification of the maternal and paternal alleles, which was not possible using this study design [24]. The associations with AKR1C1 and AKR1C3 in stromal tissue are interesting. These genes encode for members of the aldo/keto reductase superfamily and play a role in hormonal metabolism, including progesterone metabolism [25–27]. Again, future studies will be necessary to replicate these results.
While the ideal samples for eQTL-based analyses are purified populations of cells, obtaining hundreds to thousands of highly purified samples is not currently feasible. Our method computationally ‘purifies’ samples allowing increased throughput. A key limitation of the method, however, is the need for a deep understanding of the tissue composition. In this study, we modelled normal breast tissue as four cell types, however, normal breast tissue is known to be even more complex. For example, epithelial cells could be even further stratified into luminal, ductal and myoepithelial cells [13]. Capturing more of the cell types that comprise a given tissue will obviously improve model performance.

Another important aspect of the model is to accurately estimate the distribution of gene-expression levels in purified populations of the cell types that comprise the tissue. This information is critical to estimate both the cell-type proportion and transcript levels within each cell type for each sample. Future work will focus on the number of purified samples that are necessary in order to optimally calibrate these model parameters.

Using eQTL-based data to annotate loci discovered through GWAS will certainly continue. In fact, large-scale efforts such as the genotype-tissue expression and the multiple tissue human expression resource are specifically designed to connect inherited variation with transcript levels (e.g. http://commonfund.nih.gov/GTEx/ and http://www.muther.ac.uk/). We anticipate that our method can be applied to these types of datasets in order to more accurately evaluate eQTL–expression relationships.

4. Material and methods

(a) Patients
All the DNA and RNA for this project was extracted from frozen tissue derived from patients undergoing reduction mammoplasties. Informed consent was obtained and all IRB regulations were followed.

(b) Gene-expression analysis
To measure gene-expression level, nCounter Nanostring Technology assay was used. Total RNA from fresh frozen samples (100 ng per sample) was subjected to the expression analysis, and the assay was performed according to the manufacturer’s instruction manual with nCounter Nanostring platform. RNA was isolated using Trizol/chloroform isolation and extraction. A total of 122 probes were evaluated (101 probes for candidate target genes, 11 marker genes and 10 housekeeping genes—electronic supplementary material, table S1). In order to normalize the expression level, 10 housekeeping genes representing high, intermediate and low expressing genes were also measured by nCounter Nanostring. The raw data of gene expression generated by Nanostring is normalized and then the background is subtracted from each transcript that is being analysed. For each well, a set of positive control spiked-in transcripts is measured. A normalization factor is calculated for these knowns and then each transcript under investigation in the well is normalized to this reference. Background is determined by averaging counts produced by negative hybridization controls and is subtracted from normalized values.

(c) Genotyping
DNA was extracted from fresh frozen samples using the DNeasy kit from Qiagen was used and the manufacturer’s instructions were followed. Genotypes for each risk locus were determined using the sequenom platform—matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) mass spectrometry. Briefly, 10 ng of genomic DNA per sample was amplified by PCR followed by shrimp alkaline phosphatase treatment and single base extension using primers designed by Sequenom software. Samples were desalted (with 6 mg of clean resin) and dispensed to a microarray chip (SpectroCHIP; Sequenom). Chips were scanned using the MALDI-TOF MS system. Spectra were then analysed with the

### Table 1. Significant eQTL—target gene associations.

<table>
<thead>
<tr>
<th>SNP</th>
<th>genes</th>
<th>epithelial cells p-values &lt; 0.05 (n = 78)</th>
<th>stroma p-values &lt; 0.05 (n = 31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs809944</td>
<td>ZFYVE26</td>
<td>0.0073</td>
<td>0.0248</td>
</tr>
<tr>
<td></td>
<td>DCAF5</td>
<td>0.0127</td>
<td>0.0466</td>
</tr>
<tr>
<td></td>
<td>EXD2</td>
<td>0.0303</td>
<td>0.0263</td>
</tr>
<tr>
<td></td>
<td>C14orf181</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>rs13387042</td>
<td>SMARCAL1</td>
<td>0.0106</td>
<td></td>
</tr>
<tr>
<td>rs4973768</td>
<td>CMUC</td>
<td>0.0162</td>
<td></td>
</tr>
<tr>
<td>rs6861560</td>
<td>MRPS30</td>
<td>0.0361</td>
<td></td>
</tr>
<tr>
<td>rs909116</td>
<td>H19</td>
<td>0.0386</td>
<td></td>
</tr>
<tr>
<td>rs1051545</td>
<td>SLC4A7</td>
<td>0.0424</td>
<td></td>
</tr>
<tr>
<td>rs1011970</td>
<td>CDKN2A_iso5</td>
<td></td>
<td>0.0374</td>
</tr>
<tr>
<td>rs2380205</td>
<td>AKR1C3</td>
<td></td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>AKR1C1</td>
<td></td>
<td>0.0035</td>
</tr>
<tr>
<td></td>
<td>IL15RA</td>
<td></td>
<td>0.0060</td>
</tr>
<tr>
<td>rs909116</td>
<td>MOR2</td>
<td></td>
<td>0.0466</td>
</tr>
<tr>
<td></td>
<td>CTSO</td>
<td></td>
<td>0.0295</td>
</tr>
</tbody>
</table>
SPECTRUM v. 4.0A (Sequenom). Any variant with a call rate less than 85 per cent was excluded.

(d) Overview of tissue modelling
We modelled normal breast tissue as comprising four cell types, including epithelium, stroma, lymph node and adipocytes. In order to distinguish the four different cell types, we chose 11 marker genes, which are specifically expressed in these cell types as a tissue-specific gene-expression signature (see the electronic supplementary material, table S1). Then we measured the expression levels of the 11 marker genes in purified populations derived from clinical samples.

(e) Selection of marker genes
The estimation of the proportions of the different tissue types requires a set of marker genes, which are ideally exclusively present in each cell type. We hypothesized that normal breast tissue is comprised of four cell types: epithelial, stromal, adipose and immune. To represent the epithelial compartment, we used cytokeratins 5 and 19 to mark epithelial and basal epithelial compartments, respectively. To represent the adipose, stromal and inflammatory cell types, we chose three genes which we call ‘modules’. Inspection of these modules from the sample weights in the rotation matrix, we identified a set of representative genes, those mostly strongly correlated with the sample weights in the rotation matrix, which we call ‘modules’. Inspection of these modules from the most significant components revealed that most were biologically coherent, and we were able to manually assign three of these modules to adipose, stromal and inflammatory cell types. Three genes within each of these three modules were then selected as marker genes (see the electronic supplementary material, table S1).

(f) Inferring the fraction of different cell types
The characteristic gene-expression signature of the four cell types are represented by four k-dimensional, normal-distributed random variables, \(X_i\) (epithelium), \(X_j\) (stroma), \(X_k\) (lymphnode) and \(X_l\) (stroma), where \(X_i \sim N(\mu_i, \Sigma_i),\) with \(i = 1, 2, 3; k = 11.\)

The mean \(\mu_i\), and the covariance matrix \(\Sigma_i\) of the multivariate normal distribution can be estimated based on the expression levels of the 11 marker genes in the purified samples of each cell type, respectively. The corresponding density function of \(X_i\) is given as

\[
F_i(X) = \frac{1}{(2\pi)^{k/2} |\Sigma_i|^{1/2}} e^{-1/2(X-\mu_i)^T \Sigma_i^{-1} (X-\mu_i)},
\]

(4.1)

For a normal breast sample, the observed expression profile of the 11 marker genes \(X_i\) is determined by the expression profiles of the four cell types present in the tissue, as well as the fraction of each cell type. Therefore, we further hypothesized that \(X_i\) is a combination of four components, \(X_1\) to \(X_4\). Thus its density function \(G\) can be written as a mixture of \(F_i\):

\[
G = \sum_i \lambda_i F_i,
\]

(4.2)

Here, \(\lambda_i\) is the weight of component \(F_i\) and the estimate of the fraction of the corresponding cell type, if we mandate:

\[
\sum_i \lambda_i = 1.
\]

To infer \(\lambda_i\) for each observed \(X\), the likelihood function is given by

\[
L(\lambda_i) = G(X; \lambda_i) = \sum_i \lambda_i F_i(X).
\]

(4.4)

By assuming an equal prior of all four cell types, \(\lambda_i \sim N(0.25, 1)\), we can infer the posterior probability of \(\lambda_i\) for each observed \(X\).

(g) Inferring the cell-type-specific gene expression
For each observed expression profile of gene set \(Y\) and estimated fraction of the four cell types, we have

\[
Y = \lambda_1 Y_1 + \lambda_2 Y_2 + \lambda_3 Y_3 + \lambda_4 Y_4 + e.
\]

(4.5)

where \(Y_1, \ldots, Y_4\) are expression profiles of the same gene set measured from purified reference cell types \(Y_i \sim N(\omega_i, \Psi_i)\) again, \(\omega_i\) and \(\Psi_i\) are the mean expression and covariance matrix of cell type \(i\) and \(\lambda_1, \ldots, \lambda_4\) the fraction of the cell types in the sample which are inferred based on the 11 marker genes.

To estimate the cell-type-specific gene expression \(Y_i\), we rewrote the same likelihood function (4.4) as

\[
L(Y) = G(Y; \lambda) = \sum_i \lambda_i F(Y_i).
\]

(4.6)

Thus, \(Y_i\) can be obtained by simply maximizing the likelihood function.

(h) Expression quantitative trait locus analysis
For a given cell type \(i\), the estimated cell-type-specific gene expression of gene \(j\) is regarded as a function of genotype of SNP locus \(k\). We have the following linear regression to evaluate the association between SNP \(k\) and gene \(j\) in cell type \(i\)

\[
\bar{Y}_i = \beta_0 g_k + e.
\]

(4.7)

We assessed significance by permutation testing. The permutation test was performed as follows: for each round of permutation, we performed random sampling of the expression profile to generate a matrix of exactly the same dimensions (cases by genes), but all of the values are now randomized. We then used this random matrix of expression as traits and perform the eQTL analysis and retrieve all test statistics values. Then, we repeated the process 1000 times to generate a null distribution of the test statistics. The p-value for the empirically derived data was calculated by comparing the test statistics obtained from the actual samples with the null distribution generated by the permutation scheme. A two-sided permutation test p-value (0.05) was used to declare significance.

References
3. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010 Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 6, e1000888. (doi:10.1371/journal.pgen.1000888)
5. Small KS et al. 2011 Identification of an imprinted master tran regulator at the KLF14 locus related to

multiple metabolic phenotypes. Nat Genet. 43, 561 – 564. (doi:10.1038/ng.833)


7. Musunuru K et al. 2010 From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714 – 719. (doi:10.1038/nature09266)


