Germline expression quantitative trait loci shared between invasive breast and ductal carcinoma *in situ* may determine tumor progression.

Chaitanya R. Acharya[1] & Herbert K. Lyerly[1†]

[1] Department of Surgery, Duke University Medical Center, Durham, NC, USA

[†] Corresponding author

Importance

Ductal carcinoma *in situ* (DCIS) tumors, a non-invasive form of breast cancer, are currently treated as if they were obligate precursors to invasive breast cancer (IBC). Nonetheless, the genomic drivers of DCIS progression to IBC are unclear.

Objective

To identify germline variation associated with the gene expression (expression quantitative trait loci, eQTL) that determines tumor progression, and validate these eQTLs in an external DCIS dataset.

Executive summary

Ductal carcinoma *in situ* (DCIS) is a non-invasive form of breast cancer. Even though DCIS is widely considered to be a precursor to IBC, it is unclear which DCIS lesions progress to IBC. Any previous and current attempt to understand anti-tumor immune response relied on the intrinsic properties of the tumor, which entails identifying gene expression-based markers that are potential drug targets for immunomodulation. However, genetic germline aberrations can also influence the T cell response by altering expression levels of immune checkpoint molecules at the molecular level. Given the surge of interest in utilizing immunomodulatory drugs for the treatment of cancer patients, it is critical to not just understand the underlying tumor characteristics that dictate the inter-tumor heterogeneity in immune landscapes but also genetic control of gene expression that defines the likelihood that a given person's tumor will adopt a more inflamed or non-inflamed phenotype. This will eventually enable us make rational decisions in the clinical use of immunomodulatory strategies, supporting a path towards personalized immunotherapy.

To this effect, we hypothesize that germline variants such as single nucleotide polymorphisms (SNPs) contribute to the progression of DCIS to invasive disease.

Methods overview

Please see Figure 1 for our detailed strategy.

1. A total of 1411 RNA-seq libraries were generated from individual breast cancer lesions using the Illumina sequencing platform. Following quality control measures, raw sequence reads were aligned to human transcriptome in order extract gene expression for each individual transcript as a count matrix.

2. Using diffusion mapping, we identify DCIS tumors that are transitioning into invasive tumors by identifying various "states" that describe transforming gene expression patterns from DCIS to IBC.

3. We then map state-specific *cis*- and *trans*-eQTLs that drive the transition of DCIS to IBC using a novel statistical approach (see below for method description).

4. Finally, we also assess the functional effect of other genomic determinants such as copy number variation, somatic and epigenetic variation on these state-specific eQTLs by evaluating gene expression data obtained from the Cancer Genome Atlas (TCGA) Research Network. These eQTLs were then assessed for their association with tumor progression-free survival in invasive tumors (survival analysis).

Statistical method to perform state-specific eQTL analysis

For a given gene-SNP pair, our approach models gene expression across states using a linear mixed model in which both fixed and random effects are used to capture the effect of a variant on gene expression. Briefly, for each state t and individual i we model the potential genetic association between a target SNP and the expression levels of a target gene j at a single locus by using the following vectorized form of the linear-mixed model (the t-variate normal law with mean $\mu \in \mathbb{R}^t$ and variance $\Sigma \in \mathbb{R}^{t\times t}$ will be denoted as $N_t(\mu, \Sigma)$) –

$$y_{ij} = \alpha_j + \mathbf{1}\beta_j g_i + \mathbf{1}u_i + g_i v_j + \xi_{ij} \qquad \xi_{ij} \overset{i.i.d.}{\sim} N_t(0, \epsilon\mathbb{I}) \tag{1}$$

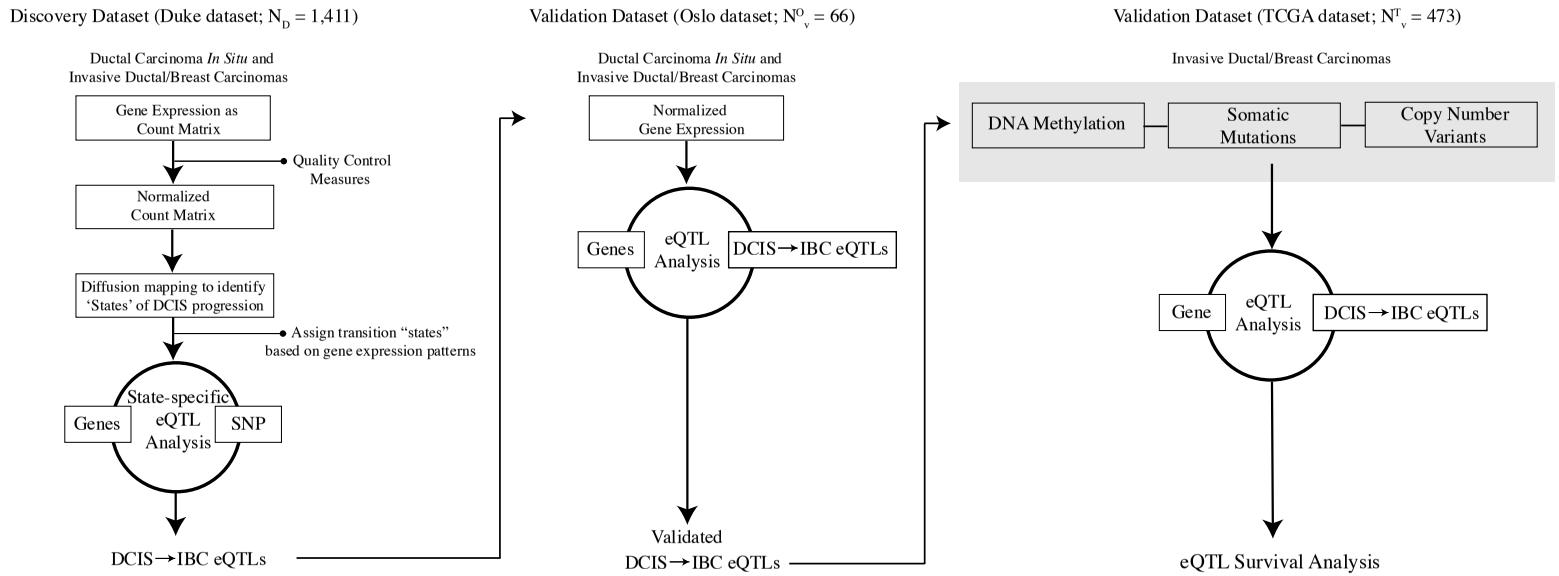

Figure 1 Our strategy to map and validate eQTLs that determine progression of DCIS to IBC. **Left panel:** We start with gene expression data (N = 1,411) as count matrix and appropriately normalize it. Following dimensionality reduction, we use diffusion mapping to identify groups of DCIS samples transitioning into IBC. Leveraging gene expression patterns between these states, we map both *cis*- and *trans*-eQTLs. **Middle panel:** We validate these state-specific eQTLs in an independent dataset (N = 66). **Right panel:** We further assess the effects of other genomic determinants such as epigenetic, somatic and copy number variation on state-specific eQTLs. A subsequent survival analysis established whether any of these eQTLs are associated with tumor progression-free survival.

where y_{ij} is a $t \times 1$ vector of gene expression data, $\mathbb{I}$ denotes the corresponding $t \times t$ diagonal matrix, $\alpha \in \mathbb{R}^t$ is the fixed effect for the mRNA level for t states, β_j is the fixed effect for the SNP ($\beta_j \in \mathbb{R}^1$), g_i is the value of a bi-allelic genotype such that $g_i \in (0, 1, 2)$, which represents the number of copies of the minor allele. $\mathbf{1}$ denotes a column vector of t ones. The random effect $v_j \in \mathbb{R}^t$ represents state-specific interaction with the genotype and $u_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that the random effects are independent and that $v_j \sim N_t(0, \gamma\mathbb{I})$ and $u_i \sim N_1(0, \tau)$.

Since state-specific effects are modeled as random effects, a test of whether there are state-specific effects is equivalent to testing whether the variance of the random effect (γ) is zero. Thus our approach involves testing only two scalar parameters (β and γ), regardless of the number of states being considered. We develop a score test of the null hypothesis that both of these parameters are zero, i.e., that the variant does not affect gene expression across any of the states.

We begin with a linear mixed effects model that models expression patterns across states as a function of genotype. In a matrix notation, for a given gene-SNP pair

$$Y = J\alpha + G\beta + Zu + Xv + \xi \tag{2}$$

where Y is a nt-dimensional matrix of expression levels in t states and n individuals, α is a fixed effect representing the state-specific intercepts, G is a nt-dimensional matrix of genotypes, β is a fixed effect of genotype across state, $u \sim N\left(0, \tau ZZ^T\right)$ is a nt-dimensional matrix of subject-specific random effect, $v \sim N\left(0, \gamma XX^T\right)$ is a nt-dimensional matrix of state-specific random effects, and $\xi \sim N(0, \epsilon I_{nt})$ and I is the identity matrix . The matrices J, Z and X are design matrices with X being a function of genotype. J is $nt \times t$ dimensional matrix denoting the design matrix for the state-specific intercepts. Z is $nt \times nt$ design matrix for the subject-specific intercepts. X is a $nt \times t$ design matrix of stacked genotypes. The parameters of interest are β and γ; α, τ and ϵ are nuisance parameters.

We test the null hypothesis that $H_0 : \beta = \gamma = 0$, i.e. the variant does not affect gene expression across any of the states. To do so, we compute the efficient scores for β and γ by projecting off components correlated with the nuisance parameters. From equation 1, the log-likelihood function of Y conditional on the genotype is –

$$\ell(\beta, \theta) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - J\alpha - G\beta)^T \Sigma^{-1}(Y - J\alpha - G\beta)$$

where θ represents the vector of all the variance components involved in Σ and c is a constant. Alternatively, under equation 1 and normality, we have

$$Y \sim N(J\alpha + G\beta, \Sigma) \quad with \quad \Sigma = \epsilon I + \tau ZZ^T + \gamma XX^T$$

The efficient scores evaluated under the null are given by –

$$U_\beta = \left(G - \bar{G}\right)^T \hat{\Sigma}_n^{-1}(Y - J\hat{\alpha}) \tag{3}$$

and

$$U_\gamma = \frac{1}{2}(Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} XX^T \hat{\Sigma}_n^{-1}(Y - J\hat{\alpha}) \tag{4}$$

where $\hat{\Sigma} = \hat{\tau}ZZ^T + \hat{\epsilon}I$ and $\hat{\tau}$ along with $\hat{\epsilon}$ are the maximum likelihood estimators of τ and ϵ under the null.

We propose a weighted sum of U_β and U_γ to arrive at our joint score test statistic, U_ψ. Since U_β is linear in Y while U_γ is quadratic, we propose the following rule to combine them –

$$\begin{aligned} U_\psi &\equiv a_\beta U_\beta^2 + a_\gamma U_\gamma \\ &= (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} \left[a_\beta \left(G - \bar{G}\right)\left(G - \bar{G}\right)^T + a_\gamma \left(\frac{1}{2} XX^T\right) \right] \hat{\Sigma}_n^{-1}(Y - J\hat{\alpha}), \end{aligned} \tag{5}$$

where a_β and a_γ are scalar constants chosen to minimize the variance of U_ψ. Under the null, U_ψ is distributed as a mixture of chi-square random variables. Several approximation and exact methods were proposed to obtain the distribution of U_ψ. Here, we use the Satterthwaite method to approximate the p values from a scaled χ^2 distribution by matching the first two moments as $U_\psi \sim \kappa\chi^2_\nu$ where $\kappa = \frac{2Var(U_\psi)}{E[U_\psi]}$ and $\nu = \frac{E[U_\psi]^2}{2Var(U_\psi)}$.