Science
AAAS

Supplementary Materials for

Dynamic genetic regulation of gene expression during cellular differentiation

B.J. Strober†, R. Elorbany†, K. Rhodes†, N. Krishnan, K. Tayeb, A. Battle*, Y. Gilad*

† These authors contributed equally to this work.
*Corresponding author. Email: ajbattle@jhu.edu (A.B.); gilad@uchicago.edu (Y.G.)

This PDF file includes:

Materials and Methods

Samples

We used induced pluripotent stem cell (iPSC) lines from 19 individuals from the Yoruba HapMap population. The iPSC lines were reprogrammed from LCLs and characterized previously (*9*). All 19 individuals are female and unrelated. We chose to use only female individuals to avoid introducing additional variance that is not of interest in this study.

iPSC Maintenance

Feeder-free iPSC cultures were maintained on Matrigel Growth Factor Reduced Matrix (CB-40230, Thermo Fisher Scientific) with Essential 8 Medium (A1517001, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning). Cells were grown in an incubator at 37°C, 5% CO2, and atmospheric O2. Cells were passaged to a new dish every 3-5 days using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded with ROCK inhibitor Y-27632 (ab120129, Abcam).

Cardiomyocyte Differentiation

We differentiated iPSCs using a protocol previously optimized for use with the Yoruba HapMap panel (*9*). This protocol implements slight modifications to the cardiomyocyte differentiation protocols from Lian et al. 2013 and Burridge et al. 2014 (*12, 25*). Feeder-free iPSCs were seeded onto wells of a 6-well plate and grown for 3-5 days prior to differentiation. When most lines were 70%-100% confluent, E8 media was replaced with "heart media" along with 1:100 Matrigel hESC-qualified Matrix (08-774-552, Corning) and 12uM of GSK-3 inhibitor CHIR99021 trihydrochloride (4953, Tocris). "Heart media" is composed of RPMI (15-040-CM, Thermo Fisher Scientific) with B27 Supplement minus insulin (A1895601, Thermo Fisher Scientific), 2mM GlutaMAX (35050-061, Thermo Fisher Scientific), and 100mg/mL Penicillin/Streptomycin (30002Cl, Corning). CHIR99021 is a small molecule that activates WNT signaling and initiates the differentiation on day 0 (after the 'day 0' cell collection) (*12*). "Heart media" was replaced 24 hours later at day 1 of differentiation. 48 hours later, at day 3 of differentiation, cells were fed with new "heart media" containing 2uM of the WNT inhibitor Wnt-C59 (5148, Tocris) (*12*). We cultured cells in Wnt-C59 heart media for 48 hours. At day 5, Wnt-C59 was removed and base "heart media" was added. "Heart media" was refreshed on days 7, 10, 12, and 14 of differentiation. Cells began spontaneous mechanical beating between days 7 and 10 of differentiation (Table S1).

Sample Collection and Processing

We performed cardiomyocyte differentiations in batches of two to five cell lines at a time. Every 24 hours from day 0 (iPSC, before treatment with CHIR99021) to day 15 for every cell line, cells in one well of a 6-well culture dish were harvested using mechanical scraping. Cells were rinsed and suspended in PBS and flash-frozen in liquid nitrogen. On day 15 of cardiomyocyte differentiation for all cell lines, we performed flow cytometry to establish purity using a cardiac-specific marker, cardiac Troponin T (564767, BD Biosciences) (Table S2). Cells were profiled on the BD LSRFortessa Cell Analyzer.

After each time-course was completed, we processed each cell line and balanced our study design with respect to differentiation batch, RNA extraction batch, person who performed the

RNA extraction, library batch, and sequencing lane to mitigate technical batch effects (Table S1). For all experimental steps after cell collection, all time points of a given cell line were processed together to minimize technical variation related to our factor of interest, which is time. We recorded 27 technical and biological covariates and measured their contribution to variation in our data (Fig. S3b).

We extracted RNA from frozen cells using the Qiagen Qiashredder and RNeasy Mini Kit (79656 & 217004, Qiagen). RNA concentration and quality was measured using the Agilent 2100 Bioanalyzer. The average RIN score for all samples was 9.51, with a standard deviation of 1.09.

Library preparation was performed using the Illumina TruSeq RNA Sample Preparation Kit v2 (RS-122-2001 & -2002, Illumina). Libraries in each batch were multiplexed together so that every sequencing lane contained samples from at least two cell lines. Cell lines were randomized such that lines that were processed together in a sequencing batch were not also together in an RNA extraction batch or a differentiation batch. In total, most sequencing lanes contained 23 to 24 multiplexed samples each. Samples were sequenced 50 base pairs, single-end using the Illumina HiSeq4000 according to manufacturer instructions. The same multiplexed library pool was sequenced twice with the goal of achieving at least 15 million reads per sample (Fig. S2).

Genotype data

We used previously collected and imputed genotype data for the 19 Yoruba individuals from the HapMap and 1000 Genomes Project (*26*).

RNA-seq quantification

All RNA-seq samples were aligned to the human genome (GRCh37) using Subread. We counted reads and estimated gene level expression with reads per kilobase million (RPKM) using the `edgeR` R package. We then filtered to genes that were protein-coding, autosomal, and had at least 10 samples such that RPKM >= .1 and raw read counts >= 6. This yielded 16,319 genes. The RPKM distribution in each sample was then quantile normalized and each gene, across all samples, was standardized (mean 0, standard deviation 1).

Biological Replication

We computed replication of day 0 cell lines within previously generated iPSC lines (*9*) and replication of day 15 cell lines within previously generated iPSC-derived cardiomyocyte cell lines (*9*). Notably, the samples from Banovich et al. were also generated in the Gilad lab and use the same panel of iPSCs. Count data from all 4 data sets was re-processed under a uniform pipeline:

1. Count data was $\log_2$(count+1) transformed
2. Each gene was standardized to have mean zero and standard deviation 1
3. Top gene expression PCs (in each data set separately) were regressed out.

We regressed out the top 3 PCs in the day 0 and day 15 data sets, top 10 PCs in the Banovich et al iPSC data set, and top 3 PCs in the Banovich et al. iPSC-derived cardiomyocyte data set. The choice of 3 PCs was selected to match the number of PCs in the non-dynamic eQTL analysis. The choice of 10 PCs in the Banovich et al. iPSC data set was selected to match their analysis.

Cell line clustering model (split-GPM)

We applied a generative model that assumes a joint clustering over the 19 cell lines and 16,319 genes. That is, the model encodes a global assignment of each of G genes to L gene clusters and assignment of each of N cell lines to K cell line clusters. For each cell line cluster, each gene cluster specifies a Gaussian process (GP) representing a latent gene expression trajectory across time. Thus, the model identifies groups of cell lines with globally different behavior, and groups of genes with similar expression trajectories within each cell line cluster.

Let y_{ng} be the observed gene expression trajectory for gene g in cell line n at times t_{ng}. Our observations are generated as follows:

$$\Phi_n \sim Categorical(\pi)$$
$$\Lambda_g \sim Categorical(\psi)$$
$$f^{kl} \sim GP\big(0, K(\theta)\big)$$
$$y_{ng} |\, \Phi_n = k, \Lambda_g = l, f^{kl}, t_{ng} \sim N(f^{kl}\big(t_{ng}\big), \sigma^2 I)$$

$\pi \in R^K \geq 0\ s.t. \sum_{k=1}^{K} \quad \pi_k = 1,\ \ \psi\ \ \in R^L \geq 0\ s.t. \sum_{l=1}^{L} \quad \psi_l = 1$ are cell line cluster mixture weights and gene cluster mixture weights respectively, θ are GP kernel hyperparameters and σ^2 is a global variance parameter. f^{kl} is a function drawn from a gaussian process, while $f^{kl}(t)$ is the function evaluated at points t.

We collect $\{\Phi_n\}_{n=1,\dots N}$ into an NxK binary matrix $\Phi\ s.t. \Phi_{nk} = 1 \iff \Phi_n = k$. Likewise, we collect $\{\Lambda_g\}_{g=1\dots G}$ into a GxL binary matrix $\ s.t.\ \Lambda_{gl} = 1 \iff \lambda_g = l$. The observed data points are conditionally independent given the functions and assignments. Our full likelihood is:

$$p(\{y_{ng}\} |\, \{f^{kl}\}, t_{ng}, \Phi, \Lambda) = \prod_{n,g,k,l}^{N,G,K,L} N(y_{ng} | f^{kl}\big(t_{ng}\big), \sigma^2)^{1(\Phi_{nk})1(\Lambda_{gl})}$$

split-GPM approximate inference

Exact computation of the posterior $p(\{f^{kl}\}, \Phi, \Lambda, |\, \{y_{ng}\}, \{t_{ng}\})$ is intractable so we resort to a variational approximation that factorizes and minimizes the KL-divergence of the true posterior:

$$q(\{f^{kl}\}, \Lambda, \Gamma) = \prod_{k,l}^{K,L} \quad q(f^{kl}) \prod_{n}^{N} \quad q(\Phi_n) \prod_{g}^{G} \quad q(\Lambda_g)$$

$$f^{kl} \sim GP(0, K(\theta))$$
$$\Phi_n \sim Categorical(\widehat{\Phi_n})$$
$$\Lambda_g \sim Categorical\big(\widehat{\Lambda_g}\big)$$

This model bears strong resemblance to the Overlapping Mixture of Gaussian process of Lazaro-Gredilla et.al (*27*) and inference proceeds the same way with the exception that the assignment

matrix is decomposed into Φ and Λ. To update the assignments, we iteratively update Φ and Λ until convergence or until a fixed number of iterations is reached.

$$ELBO(q) = E_q[log\ p(\{\{y_{ng}\}|\{f^{kl}\}, \{t_{ng}\}, \Phi, \Lambda)] + E_q[log\ p(\{f^{kl}\}, \Phi, \Lambda) - E_q[log\ q(\{f^{kl}\}, \Phi, \Lambda)]$$

We iteratively estimate assignment variables and trajectory estimates, then perform gradient based optimization with respect to the kernel parameters. This approximation requires $K \cdot L$ GP regressions, each computed over every data point. To make the problem tractable we further approximate each GP via SVGP (*28*).

In this analysis, we train a model with $K = 2$ cell line clusters, $L = 20$ gene clusters and an RBF kernel with shared length-scale and variance parameters for all $K \cdot L$ clusters.

Non-dynamic cis-eQTL calling per time point

Separately, each time point has a small sample size (maximum of 19 samples). Therefore, we used the WASP combined haplotype test (CHT) (*15*) to increase power, integrating both total expression and allelic imbalance data into the same test, to detect cis-eQTLs in each of the 16 time points, independently. In order to increase accuracy of allele-specific expression estimates, RNA-seq data was re-quantified for eQTL calling by filtering Subread mapped reads using the WASP mapping pipeline under default settings in order to reduce biases in allelic mapping. We tested cis-eQTL association for variants within 50 KB of each gene's transcription start site. Further, we tested the same set of variant-gene pairs in all time points, limiting to variant-gene pairs that passed the following filters in all 16 time points:

1. Variant has minor allele frequency >= .1
2. Gene passes all filters described in "RNA-seq quantification" section
3. Gene has >= 100 reads mapped summed across all cell lines
4. Exon of the gene contains a heterozygous variant in at least 5 cell lines
5. Sum of reads mapping to minor allele across all cell line, heterozygous variant pairs >= 25

These filters yielded 1,009,173 variant-gene pairs (6,362 unique genes) tested in each time point. The same variant-gene pairs were tested in each time point to reduce bias when comparing genetic regulatory effects between time points. We included the first three raw read count expression PCs from samples belonging to the corresponding time point as covariates. The choice to control for three PCs was motivated by maximizing the number of significant non-dynamic eQTLs detected in each time step (Fig. S9B). We ran one permutation of the CHT genome-wide. It is worth noting that the CHT is not well calibrated (Fig. S10**)**. Multiple testing correction was performed using empirical FDR (eFDR) (*29*) to assess genome-wide significance based on a vector of observed p-values and a vector of null (permuted) p-values. An empirical approach to FDR correction should account and control for the lack of calibration observed when the CHT was applied to our data.

Sparse non-negative matrix factorization

We performed sparse, non-negative matrix factorization of eQTL statistics for all time points to identify broad patterns in eQTL effects. Here, we limited to genes with at least one significant

eQTL (eFDR <= .05) across time points. If a gene had more than one significant eQTL, we selected a single variant for that gene with the smallest geometric mean p-value across all 16 time points. We then filled in a matrix, X, where each row represents one gene, each column represents a time point, and each element represents the -$\log_{10}$ p-value corresponding to the row's gene and the column's time point. We then performed sparse non-negative matrix factorization on X (dim NxT) using the python function `sklearn.decomposition.NMF` (*30*). With K latent factors, this will reduce X into the product of a loadings matrix (L; dim NxK) and a factor matrix (F; dim KxT). F captures shared patterns of eQTL effect sizes across time while L reflects which factors are relevant for each eQTL. All default settings were used except we set `l1_ratio=1` to enforce an element-wise L1 penalty. We ran this analysis for a range of number of latent factors and L1 penalties (alpha) (Fig. S11).

Linear dynamic eQTLs

Linear dynamic eQTLs are cis-eQTLs whose effects are linearly modulated by differentiation time. We detected linear dynamic eQTLs with a gaussian linear model that quantified the interaction between genotype and differentiation time on gene expression, while controlling for the linear effects of both genotype and differentiation time. We also controlled for linear effects of the first five cell line collapsed PCs (see below) and, critically, the linear effects of the interaction between the first five cell line collapsed PCs and differentiation time.

We built a separate linear model for each tested variant-gene pair. Specifically, let t denote the time point of the current sample, c denote the cell line of the current sample, T denote the total number of time points, and C denote the total number of samples. $E \in R^{CxT}$ denotes the standardized expression matrix for the current gene, $G \in R^{C}$ denotes the dosage based genotype vector for the current variant, and $PC^{K} \in R^{C}$ denotes the Kth cell line collapsed PC vector. We modeled the expression levels as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 PC_c^1 + \beta_4 PC_c^1 t + \cdots + \beta_{11} PC_c^5 + \beta_{12} PC_c^5 t + \beta_{13} G_c t, \sigma)$$

We used R `lm` to quantify the significance of the interaction between genotype and time (β_{13}). We computed a null distribution by randomly permuting the time point variable that was used for the term capturing the interaction between genotype and time (β_{13}), while keeping the time point variable in all other terms not permuted. An independent permutation was used for every tested variant gene pair. Using this permutation run, we computed significance with eFDR.

We tested the same set of variant-gene pairs that was tested in the non-dynamic eQTL calling analysis. This was done to reduce bias when comparing non-dynamic eQTLs and dynamic eQTLs.

Cell line confounder estimation using cell line collapsed PCA

Different cell lines can display broadly different patterns of expression across the entire time course, including not only consistent shifts upward or downward in expression of subsets of genes, but different slopes and more generally different expression trajectory shapes (Fig. 1B). Variability in slope is of particular concern for detection of dynamic eQTLs – if a subset of cell lines display different slopes over time for many genes, this would lead directly to false positive dynamic eQTLs. Specifically, these cell line subsets reflecting confounders could by chance

correspond to the same grouping as genotype across numerous SNPs given the large number of SNPs compared to cell lines. This would then produce apparently large effect $\beta_{13} G_c t$ terms in the dynamic eQTL linear model, and thus numerous false positives. To combat this problem, we used a PCA-based approach we refer to as "cell line collapsed PCA" to identify broad, cell line specific patterns across the entire time course. To do so, we simply rearranged the gene expression matrix from the standard RNA-seq quantification (RPKM levels across 297 samples by 16,319 genes) such that each row was now expression from one cell line and each column was a gene at a single time point. We excluded time points that were not fully observed (days 2, 4, and 13) to avoid missing entries, yielding a final matrix of size 19 by 212,147 (Fig. S13). After standardizing each column, we applied PCA to this matrix to learn a low dimensional representation. Here, each cell line has a shared loading across all time points, and PCs reflect trajectories across all genes, rather than a standard application of PCA with loadings for each sample (a cell line, time point pair).

To ensure that we effectively controlled for the potential confounding effects of cell lines displaying broad trajectory differences over time, we calculated the frequency at which each pair of cell lines share the same genotype across all significant dynamic eQTLs. As noted above, a confounder would cause subsets of cell line to have the same eQTL SNP genotype more often than expected by chance alone, corresponding to cell line clusters with broad differences. In fact, when we do not include cell line collapsed PC loadings in our model, we do see an abundance of such likely false positives (Table S4). After controlling for 5 cell line collapsed PCs, the cell lines do not share the same genotype across significant dynamic QTLs more often than background (Fig. S16), confirming that cell line PCs help address confounding effects of individual cell line trajectories.

An alternative approach of using pseudo-time, rather than actual time in association testing, does not fully address the problem mentioned here – cell lines don't simply progress faster or slower along the same ultimate trajectory, but seem to deviate in a more complex pattern. Here, this pattern appears to correspond to cell type purity, but more generally, differentiation or any temporal response that follows branching trajectories that can't be captured by a single monotonic pseudo-time term could lead to similar false positives.

We controlled for the first five cell line collapsed PCs and their interaction with differentiation time when detecting both linear and nonlinear dynamic eQTLs. While there does not exist an optimal method to select the number of cell line collapsed PCs, we selected 5 cell line collapsed PCs that: (a) capture most of the variance in gene expression (Fig. S14a), (b) ensure cell lines do not share the same genotype across significant dynamic QTLs more often than background (Fig. S16), and (c) result in consistency between non-dynamic eQTLs and dynamic eQTLs (Fig. S21 and S25).

Simulating expression samples for linear dynamic eQTL power analysis

Using the same notation as defined in the "Linear dynamic eQTLs" section, we define the alternate model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t + \beta_3 (t * G_c), \sigma)$$

And the null model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t, \sigma)$$

For each setting of number of cell lines, t-statistic and minor allele frequency, we simulated 10,000 independent tests (variant-gene pairs) where a specified proportion of those tests follow the null and alternate models. We made the simplifying assumption that each cell line contained 16 time points (T=16). For each test:

1. The genotype vector (G_c) was randomly generated assuming a specified minor allele frequency. Specifically, both alleles of the variant were drawn independently and both alleles were forced to have the specified minor allele frequency
2. β_1 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1
3. β_2 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1
4. β_3 was equal to the t-statistic multiplied by σ. For convenience, σ was fixed to be .1
5. E_{ct} was randomly drawn
6. p-values were computed using the linear model described in the "Linear dynamic eQTLs" section excluding any fixed effects containing cell line collapsed PCs

Significance of simulated tests was assessed at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data).

Nonlinear dynamic eQTLs

To detect dynamic eQTLs whose effect size changes non-linearly with time, we used a second order polynomial basis function over time, which alters the above linear dynamic eQTL model as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 t^2 + \beta_4 PC_c^1 + \beta_5 PC_c^1 t + \beta_6 PC_c^1 t^2 + \cdots + \beta_{16} PC_c^5 + \beta_{17} PC_c^5 t + \beta_{18} PC_c^5 t^2 + \beta_{19} G_c t + \beta_{20} G_c t^2, \sigma)$$

We quantify the joint effect of the two interaction terms between genotype and time (β_{19} and β_{20}) with a likelihood ratio test with two degrees of freedom using the R `lmtest` package. We computed a null distribution by randomly permuting the time point variable that was used for the two terms capturing the interaction between genotype and time (β_{19} and β_{20}), while keeping the time point variable in all other terms not permuted. An independent permutation was used for every tested variant gene pair. It is worth noting that the nonlinear dynamic eQTLs are not well calibrated (Fig. S18**)**. Using this permutation run, we computed significance using eFDR. An empirical approach to FDR correction should account and control for the observed lack of calibration of this test.

Simulating expression samples for nonlinear dynamic eQTL power analysis

Linear dynamic eQTLs allow us to capture dynamic eQTLs whose effect size changes linearly with differentiation time. Nonlinear dynamic eQTLs allow us to capture dynamic eQTLs whose effect size changes as a quadratic function of differentiation time. However, both of these approaches are unable to capture arbitrary nonlinear functions of differentiation time. A statistical test that could capture arbitrary nonlinear functions of differentiation time is an ANOVA analysis where time is fit as a factor with 16 levels (ANOVA eQTLs). Here, we simulate several nonlinear dynamic eQTLs and access detection power using three different dynamic eQTL methods:

1. Linear dynamic eQTLs

2. Nonlinear dynamic eQTLs
3. ANOVA dynamic eQTLs

Using a similar notation as defined in the "Linear dynamic eQTLs" section, we define the alternate model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t_{new} + \beta_3 (t_{new} * G_c), \sigma)$$

And the null model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t_{new}, \sigma)$$

Here, t_{new} is a transformation of t. We used four arbitrary transformations of t:

1. $t_{new} = t(t-10)$
2. $t_{new} = t(t-7)(t-15)$
3. $t_{new} = \sin(pi * \frac{t}{5})$
4. $t_{new} = I[t > 7]$

Transformed differentiation time (t_{new}) was scaled to have the same standard deviation as the original values of differentiation time. For each setting of number of cell lines, t-statistic and time transformation, we simulated 10,000 independent tests (variant-gene pairs) where 30% of those tests follow the alternate model and 70% follow the null model. We made the simplifying assumption that each cell line contained 16 time points (T=16). For each test:

1. The genotype vector (G_c) was randomly generated assuming a minor allele frequency of .4. Specifically, both alleles of the variant were drawn independently and both alleles were forced to have a minor allele frequency of .4.
2. β_1 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1
3. β_3 was equal to the t-statistic multiplied by σ. For convenience, σ was fixed to be .1
4. E_{ct} was randomly drawn
5. p-values were computed using the three statistical models described above

Significance of simulated tests was assessed at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data).

Linear dynamic eQTL classifications

We classified the linear dynamic eQTLs as *early* (when the eQTL effect size decreased over time), *late* (when the eQTL effect size increased over time), or *switch* (when the eQTL effect size changes sign over the time course. To do so, we computed predicted eQTL effect size at day 0 and day 15 according to the fitted linear dynamic eQTL model:

Let $\hat{E}_{vg}(t = x, G = y)$ be the predicted expression (according to the fitted dynamic eQTL model) of gene g at time x for a sample with genotype dosage y for variant v. We defined the eQTL effect size ($\beta_{vg}(t = x)$) of variant v on gene g at time x as:

$$\beta_{vg}(t = x) = \hat{E}_{vg}(t = x, G = 0) - \hat{E}_{vg}(t = x, G = 2)$$

If the sign of $\beta_{vg}(t = 0)$ is equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t = 0)| \geq |\beta_{vg}(t = 15)|$
2. late if $|\beta_{vg}(t = 0)| < |\beta_{vg}(t = 15)|$

If the sign of $\beta_{vg}(t=0)$ is not equal to the sign of $\beta_{vg}(t=15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t=0)| \geq |\beta_{vg}(t=15)|$ and $|\beta_{vg}(t=15)| <$ thresh
2. late if $|\beta_{vg}(t=0)| < |\beta_{vg}(t=15)|$ and $|\beta_{vg}(t=0)| <$ thresh
3. switch if $|\beta_{vg}(t=0)| \geq$ thresh and $|\beta_{vg}(t=15)| \geq$ thresh

We assigned thresh = 1.

Nonlinear dynamic eQTL classifications

We classified the nonlinear dynamic eQTLs as early (when the eQTL effect size decreased over time), late (when the eQTL effect size increased over time), switch (when the eQTL effect size changes sign over the time course, or middle (when the eQTL is strongest in the middle of the time course). To do so, we computed predicted eQTL effect size at t=0, t=7.5, and t=15 according to the fitted nonlinear dynamic eQTL model:

$$\beta_{vg}(t=0) = \hat{E}_{vg}(t=0, G=0) - \hat{E}_{vg}(t=0, G=2)$$
$$\beta_{vg}(t=7.5) = \hat{E}_{vg}(t=7.5, G=0) - \hat{E}_{vg}(t=7.5, G=2)$$
$$\beta_{vg}(t=15) = \hat{E}_{vg}(t=15, G=0) - \hat{E}_{vg}(t=15, G=2)$$

If $\beta_{vg}(t=7.5) \geq \beta_{vg}(t=0)$ and $\beta_{vg}(t=7.5) \geq \beta_{vg}(t=15)$, we assigned the dynamic eQTL to middle.

If the sign of $\beta_{vg}(t=0)$ is equal to the sign of $\beta_{vg}(t=15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t=0)| \geq |\beta_{vg}(t=15)|$
2. late if $|\beta_{vg}(t=0)| < |\beta_{vg}(t=15)|$

If the sign of $\beta_{vg}(t=0)$ is not equal to the sign of $\beta_{vg}(t=15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t=0)| \geq |\beta_{vg}(t=15)|$ and $|\beta_{vg}(t=15)| <$ thresh
2. late if $|\beta_{vg}(t=0)| < |\beta_{vg}(t=15)|$ and $|\beta_{vg}(t=0)| <$ thresh
3. switch if $|\beta_{vg}(t=0)| \geq$ thresh and $|\beta_{vg}(t=15)| \geq$ thresh

We assigned thresh = 1.

ChromHMM enrichment analysis

We computed enrichment of dynamic eQTLs within cell type specific chromHMM (15 state model) enhancer elements relative to 1,000 sets of randomly selected background variants matched for distance to transcription start site and minor allele frequency (*16*). We considered the following four chromHMM states to represent enhancer elements:

1. EnhG (state 6)
2. Enh (state 7)
3. BivFlnk (state 11)
4. EnhBiv (state 12)

We used the following five Roadmap cell types to represent iPSCs (*17*):

1. E018: iPS-15b Cells

2. E019: iPS-18 Cells
3. E020: iPS-20b Cells
4. E021: iPS DF 6.9 Cells
5. E022: iPSC DF 19.11 Cells

And the following five Roadmap cell types to represent heart-related cells (*17*):

1. E065: Aorta
2. E083: Fetal heart
3. E095: Left ventricle
4. E104: Right atrium
5. E105: Right Ventricle

To compute enrichment within iPSC specific enhancer elements, we limited to enhancer elements found in at least one of the 5 iPSC cell types and none of the heart-related cell types. Likewise, for enrichment with heart specific enhancer elements, we limited to enhancer elements found in at least one of the 5 heart-related cell types and none of the iPSC related cell types. Odds ratios were smoothed by adding smoothing constant of 1 to each overlap count.

Dilated cardiomyopathy gene set enrichment analysis

We define the dilated cardiomyopathy gene set as the union of all genes in Supplementary Table 3 of Burke et al. (*18*). Enrichment was computed via Fisher's exact test.

Supplementary Figures

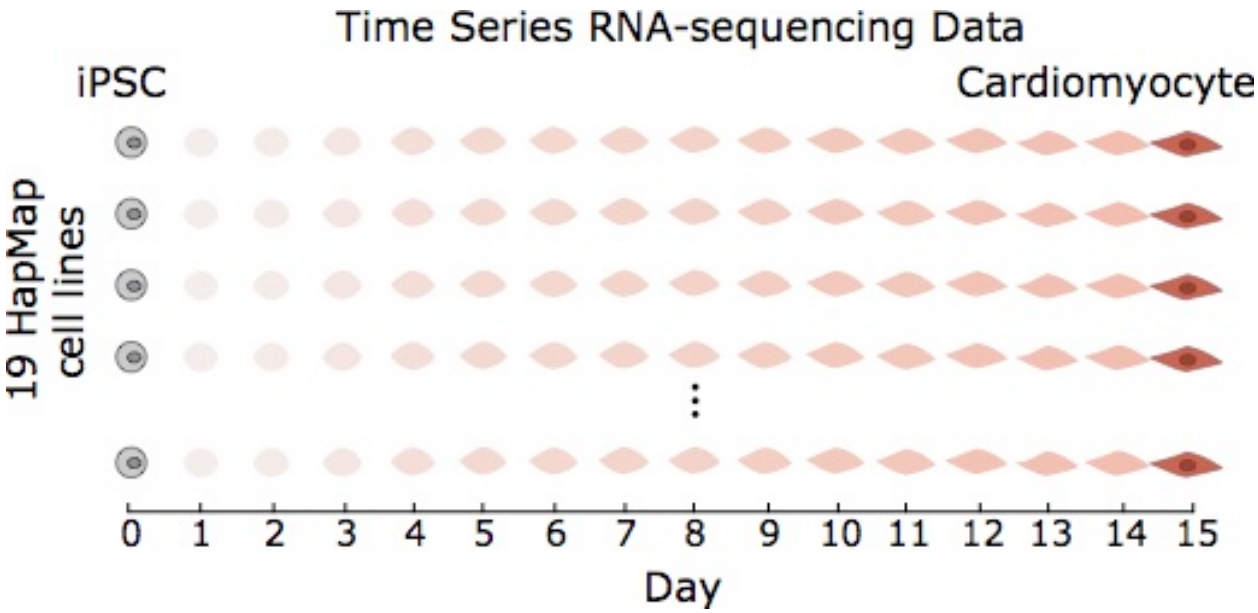

Figure S1: RNA-seq sample collection: Overview of RNA-seq sample collection. In 19 Yoruba HapMap cell lines, RNA was extracted and sequenced every 24 hours at 16 time points, generating 297 RNA-seq samples.

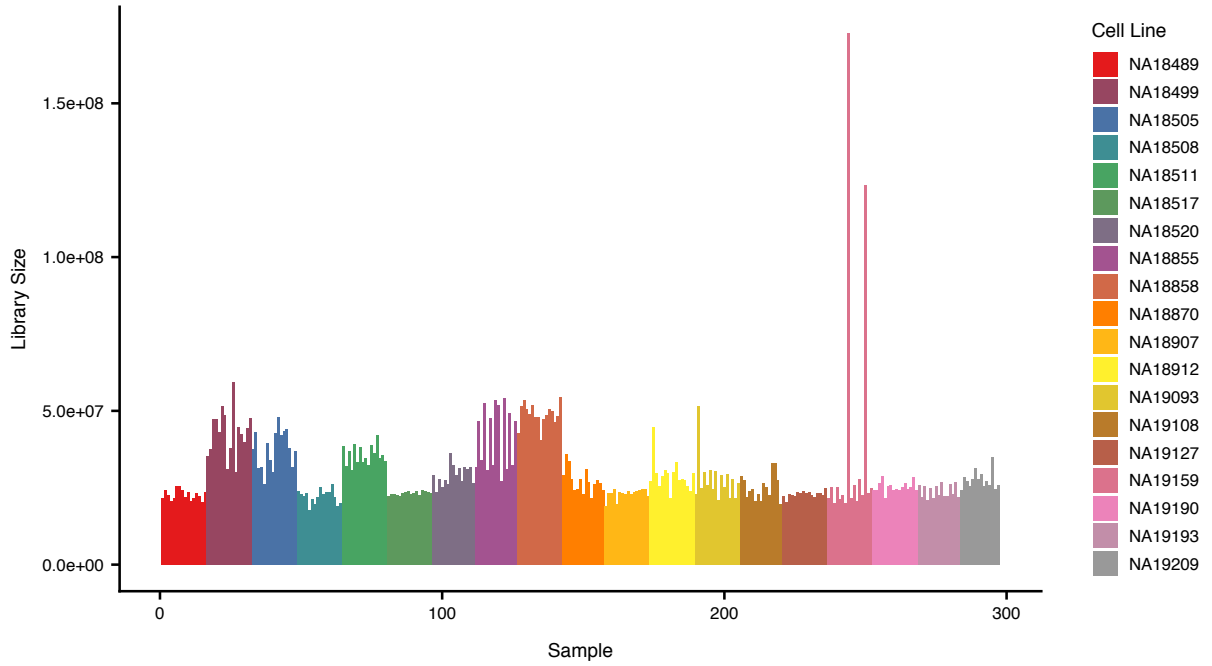

Figure S2: Library size of RNA-seq samples. The library sizes of 297 RNA-seq samples colored by their cell line identity. Within each cell line, samples are ordered along the x-axis by their differentiation time point from day 0 to 15.

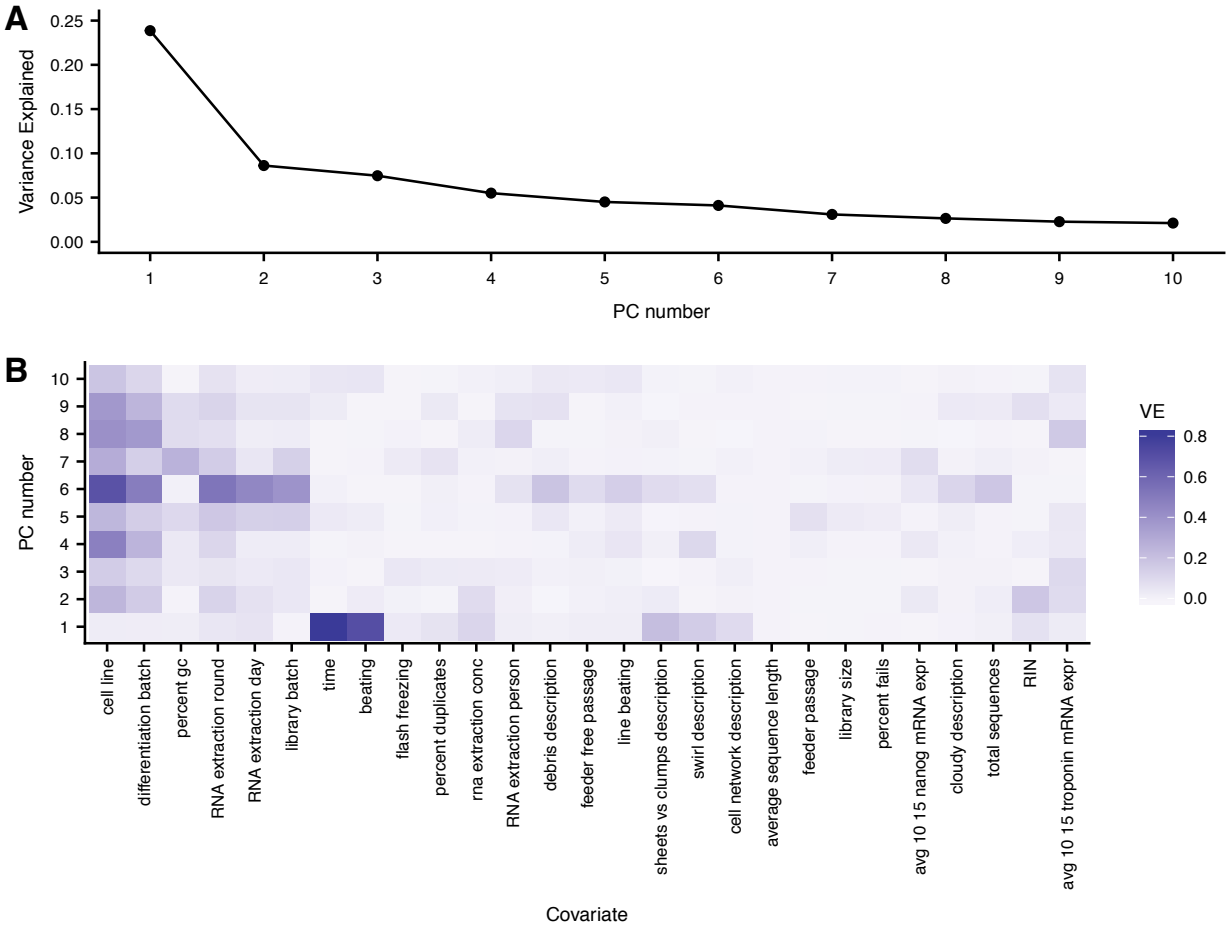

Figure S3: Explaining principal components with sample covariates. (A) Variance in gene expression explained by first 10 gene expression principal components. (B) Variance explained of each gene expression principal component using sample covariates. Adjusted R^2 was used to handle categorical sample covariates. Detailed explanation of each sample covariate can be found in Table S1.

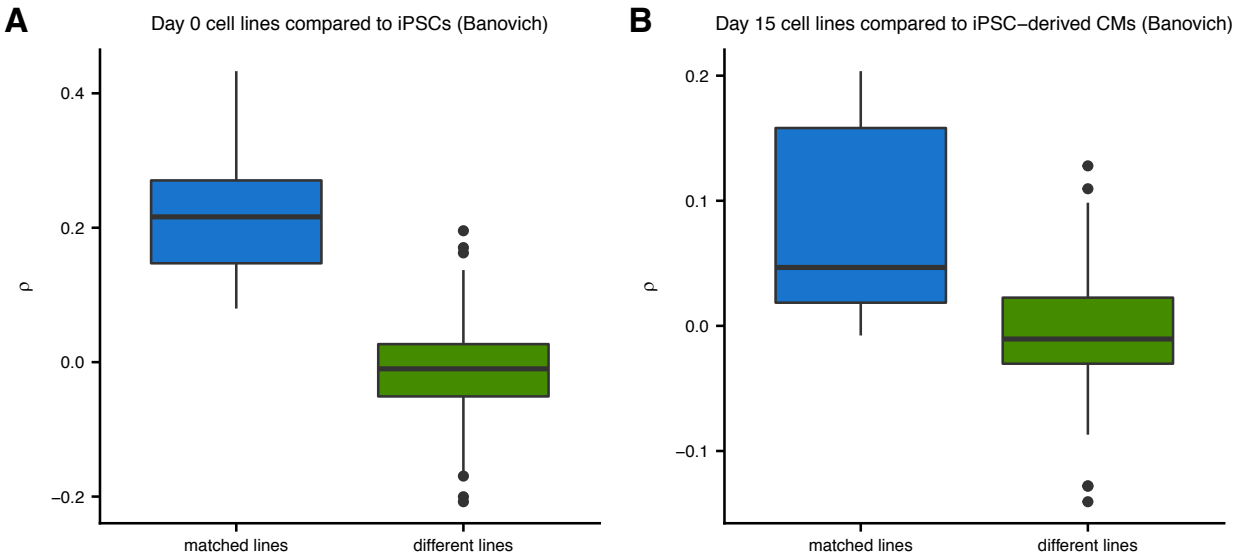

Figure S4: Biological replication of day 0 and day 15 cells. We compared day 0 and day 15 cell lines with matched iPSC lines and iPSC-derived cardiomyocyte lines, respectively, from Banovich et al. (*9*). This analysis was restricted to cell lines present in both data sets. Spearman correlation across genes observed in both data sets between (A) day 0 cell lines and iPSC lines and between (B) day 15 cell lines and iPSC-derived cardiomyocyte cell lines. Distribution of spearman correlations shown for matched cell lines (blue) and different cell lines (green). The correlation of gene expression is greater for matched cell lines compared to different cell lines ($p < .05$ for both comparisons, Wilcoxon rank-sum test).

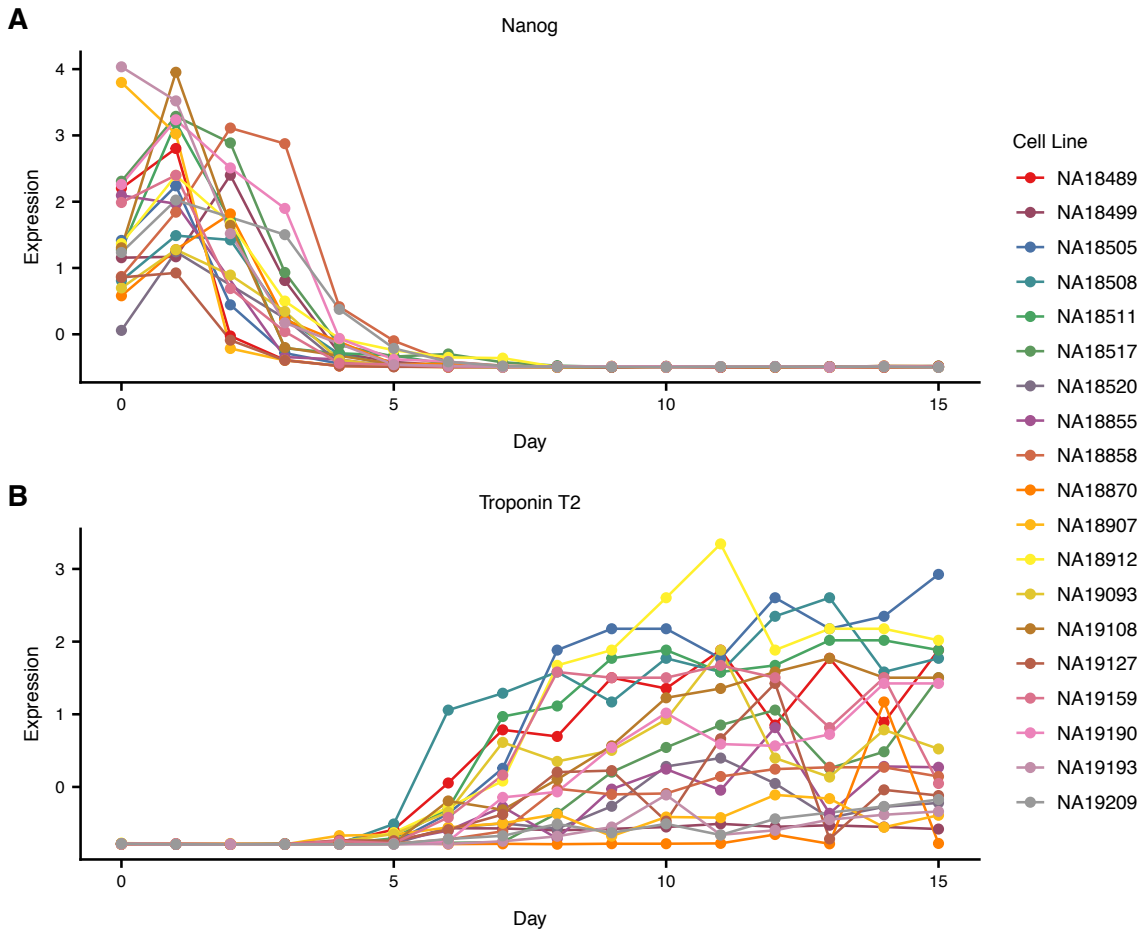

Figure S5: Expression time course of known cell type specific marker genes. Standardized gene expression levels of *Nanog* (A, stem cell marker gene) and *Troponin T2* (B, cardiomyocyte marker gene) across 16 time points (x-axis) and 19 cell lines (colors).

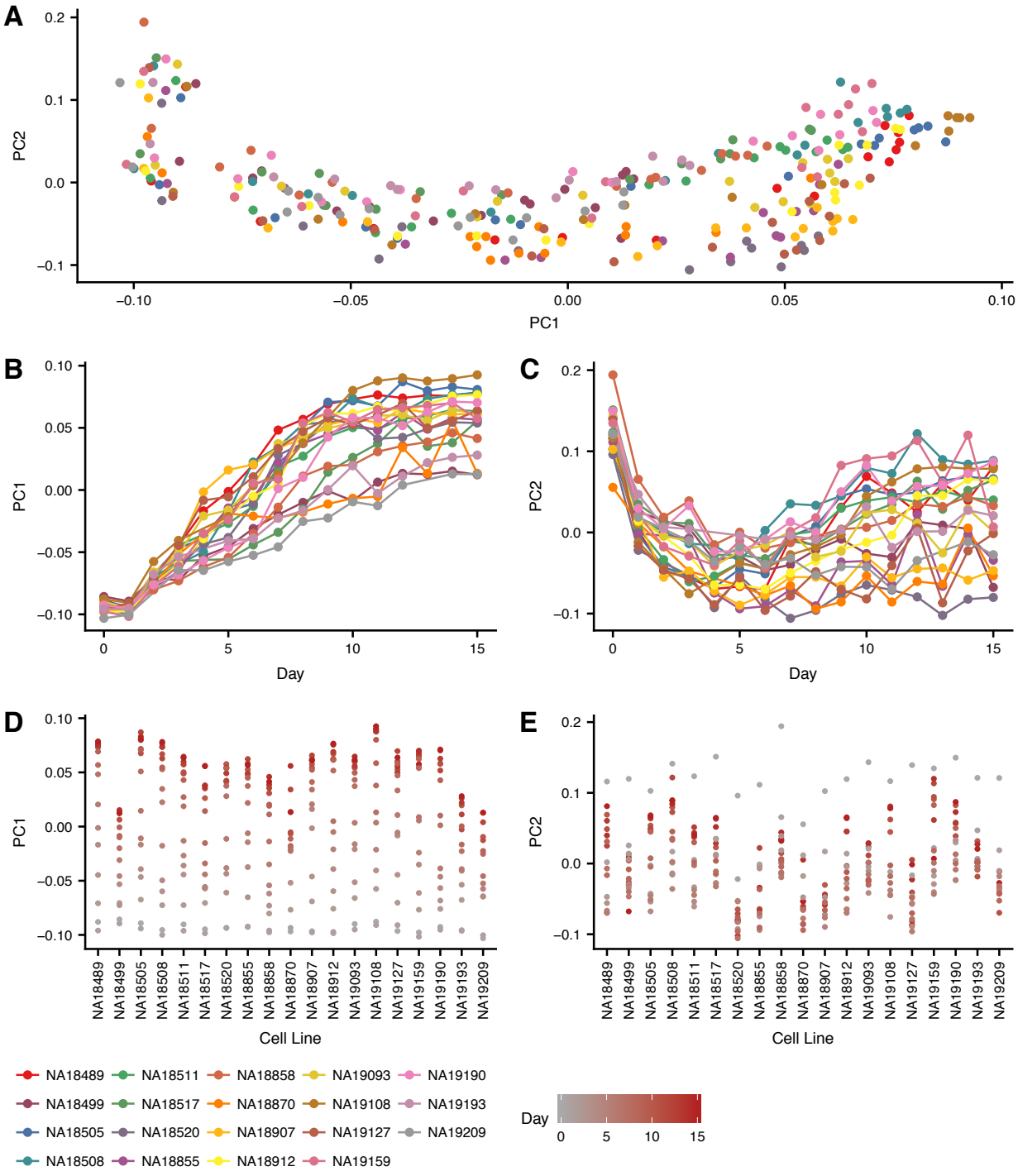

Figure S6: Principal component analysis separated by cell line identity. (A) First two gene expression principal component loadings for all 297 RNA-seq samples, where each sample is colored by its cell line identity. (B, C) Principal component 1 and 2 loadings across 16 time points (x-axis) and 19 cell lines (colors). (D, E) Principal component 1 and 2 loadings across 19 cell lines (x-axis) and 16 time points (colors).

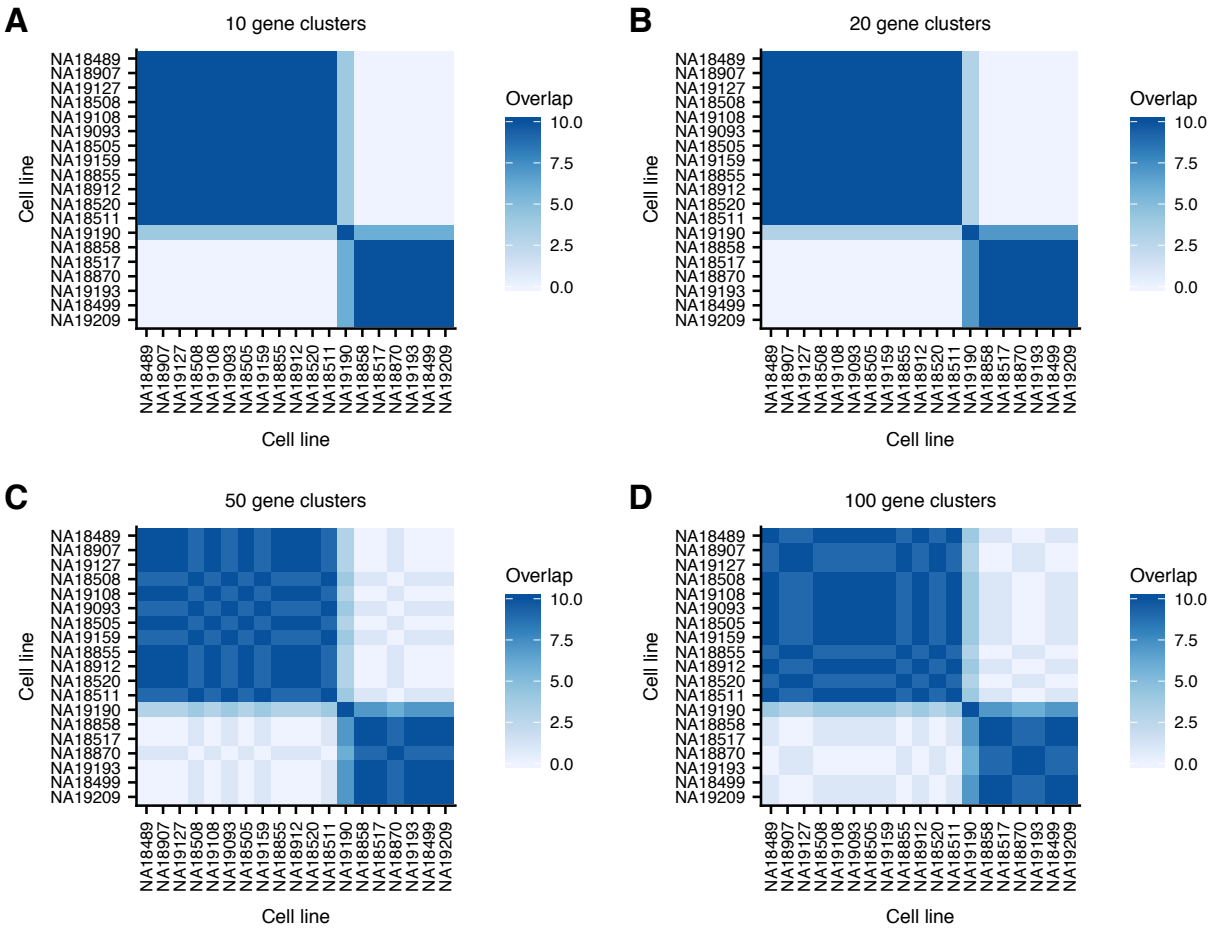

Figure S7: split-GPM cell line cluster assignment robust to hyper-parameter choice. Number of times (out of 10 split-GPM runs with independent, random initializations) that each cell line pair was assigned to the same cell line cluster when 10 (A), 20 (B), 50 (C), and 100 (D) gene clusters were used. Cell lines are ordered by their cell line collapsed PC1 loadings.

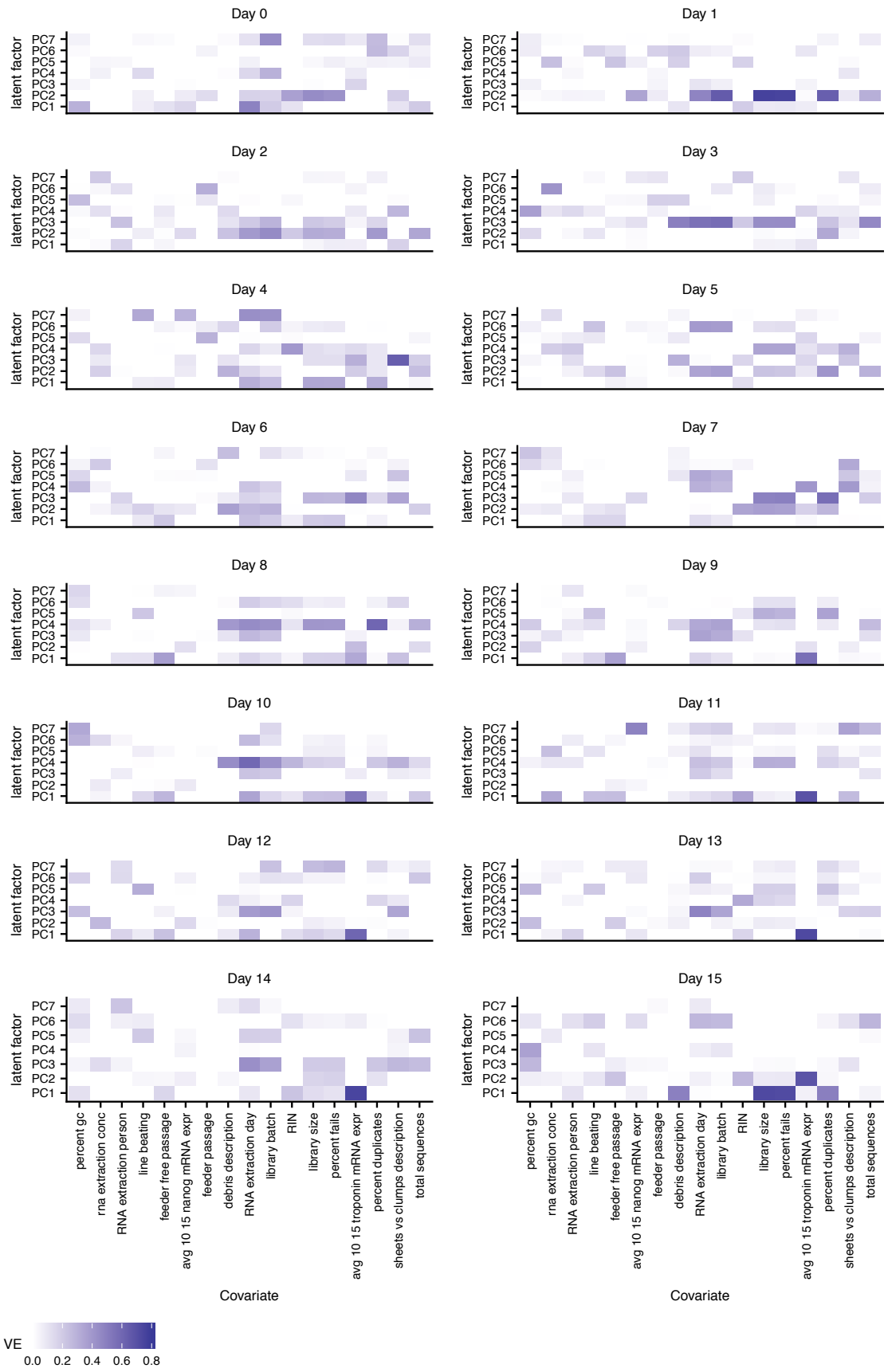
Day 0
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
Day 8
Day 9
Day 10
Day 11
Day 12
Day 13
Day 14
Day 15
latent factor
PC7
PC6
PC5
PC4
PC3
PC2
PC1
percent gc
rna extraction conc
RNA extraction person
line beating
feeder free passage
avg 10 15 nanog mRNA expr
feeder passage
debris description
RNA extraction day
library batch
RIN
library size
percent fails
avg 10 15 troponin mRNA expr
percent duplicates
sheets vs clumps description
total sequences
Covariate
VE
0.0
0.2
0.4
0.6
0.8

Figure S8: Explaining time step principal components with sample covariates. In each time point independently, variance explained of each raw read count expression principal components (from samples belonging to the corresponding time point) using sample covariates. Adjusted R^2 was used to handle categorical sample covariates. Sample categorical covariates with more than 8 categories were excluded from this analysis due to the small sample size when considering time points, independently. Detailed explanation of each sample covariate can be found in Table S1.

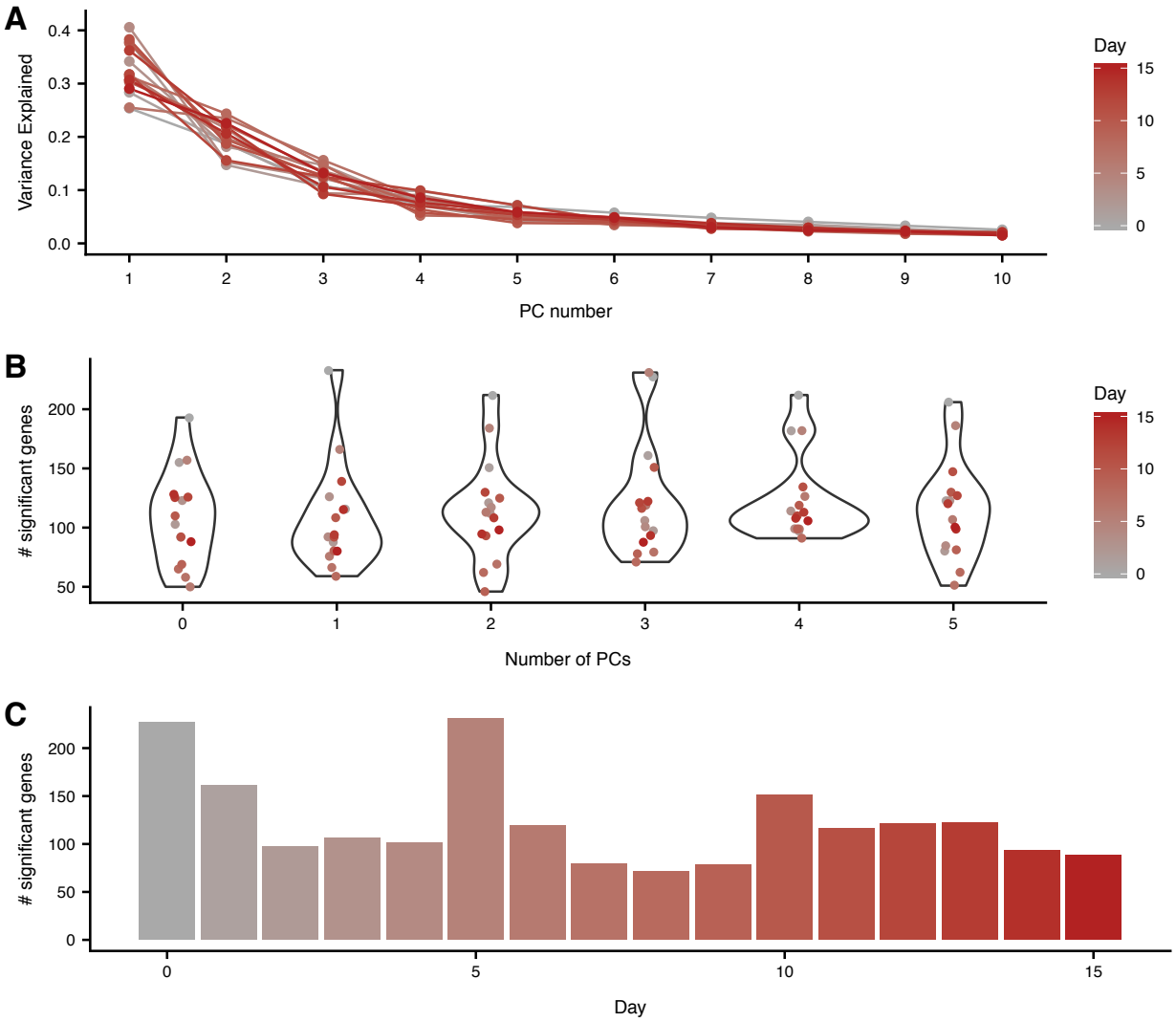

Figure S9: Number of genes with non-dynamic eQTLs. (A) Variance explained of gene expression from samples belonging to a particular time point (color) by the first 10 gene expression PCs (x-axis) computed on samples belonging to that time point. (B) The number of genes with a significant eQTL (eFDR <= .05) in each time point (color) as a function of number of expression PCs controlled for (x-axis). (C) The number of genes with a significant eQTL (eFDR <= .05) in each time point when controlling for three expression PCs.

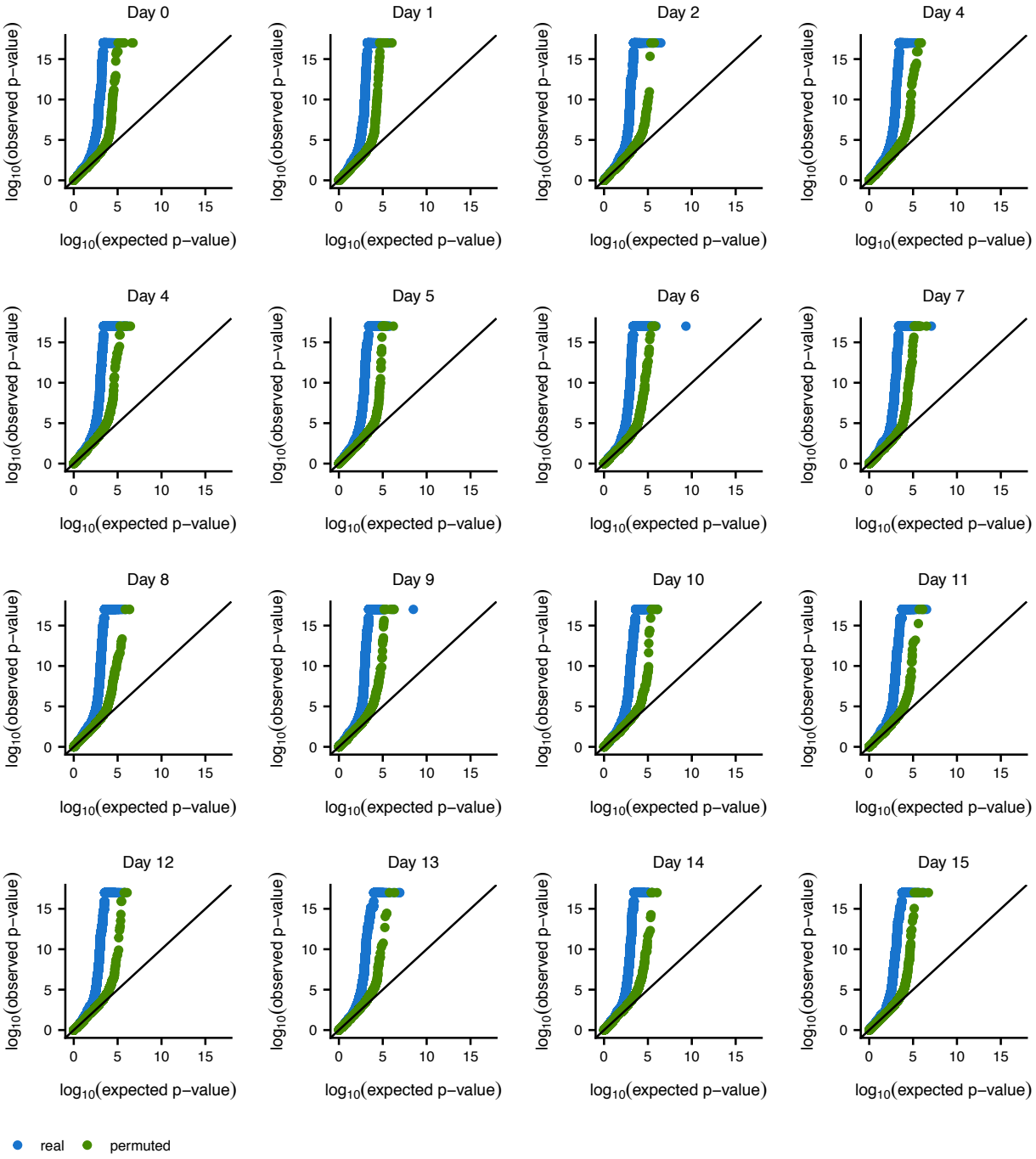

Figure S10: Q-Q plots for non-dynamic eQTLs. Q-Q plot for non-dynamic eQTLs in all 16 time steps. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data (using WASP's permutation strategy) relative to uniformly distributed p-values.

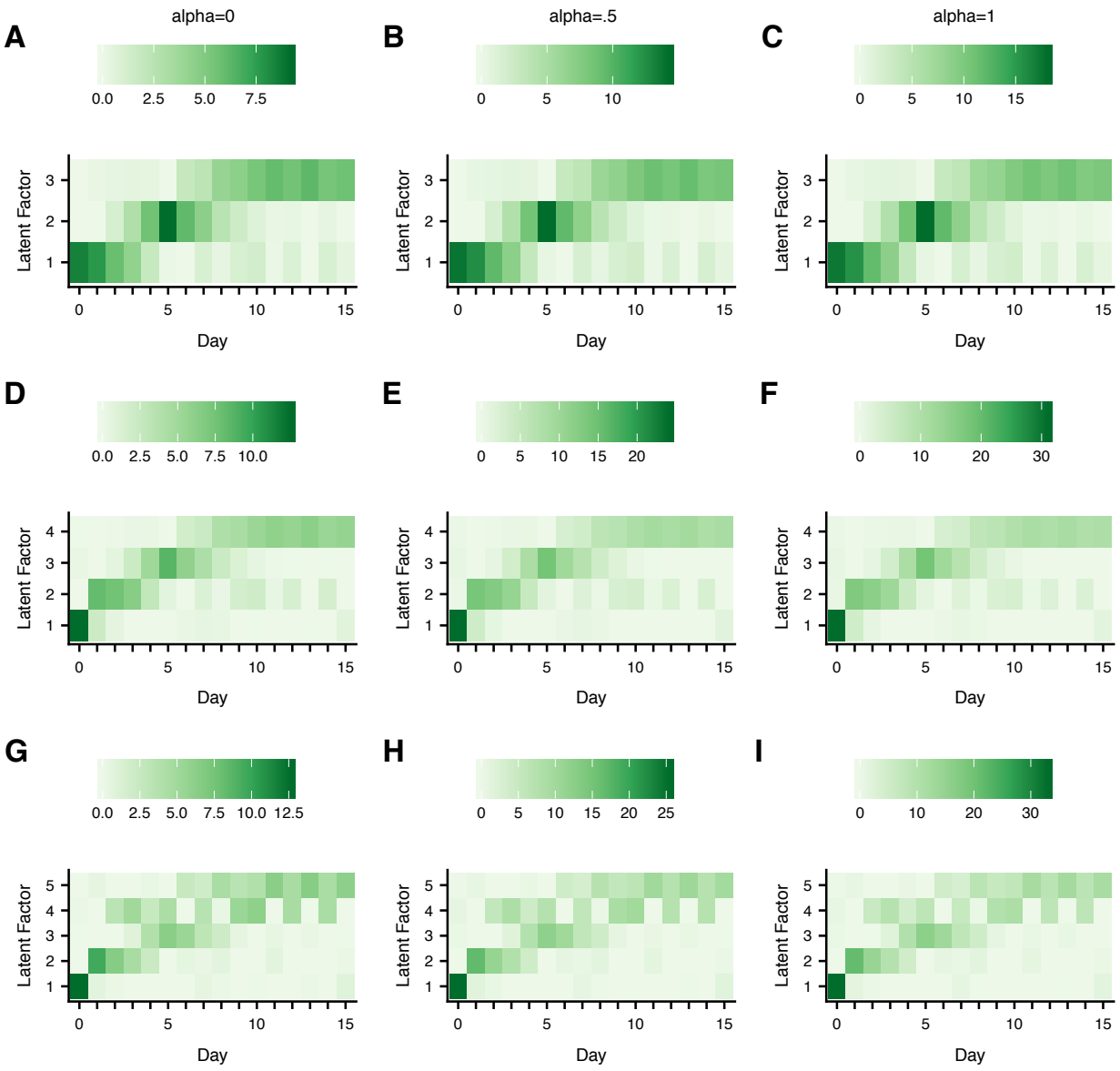

Figure S11: Matrix factorization of eQTL summary statistics. Latent factors identified via sparse non-negative matrix factorization of non-dynamic eQTL -$\log_{10}$ p-values shown for a range of sparse prior choices (alpha; columns) when using 3, 4, and 5 factors (rows).

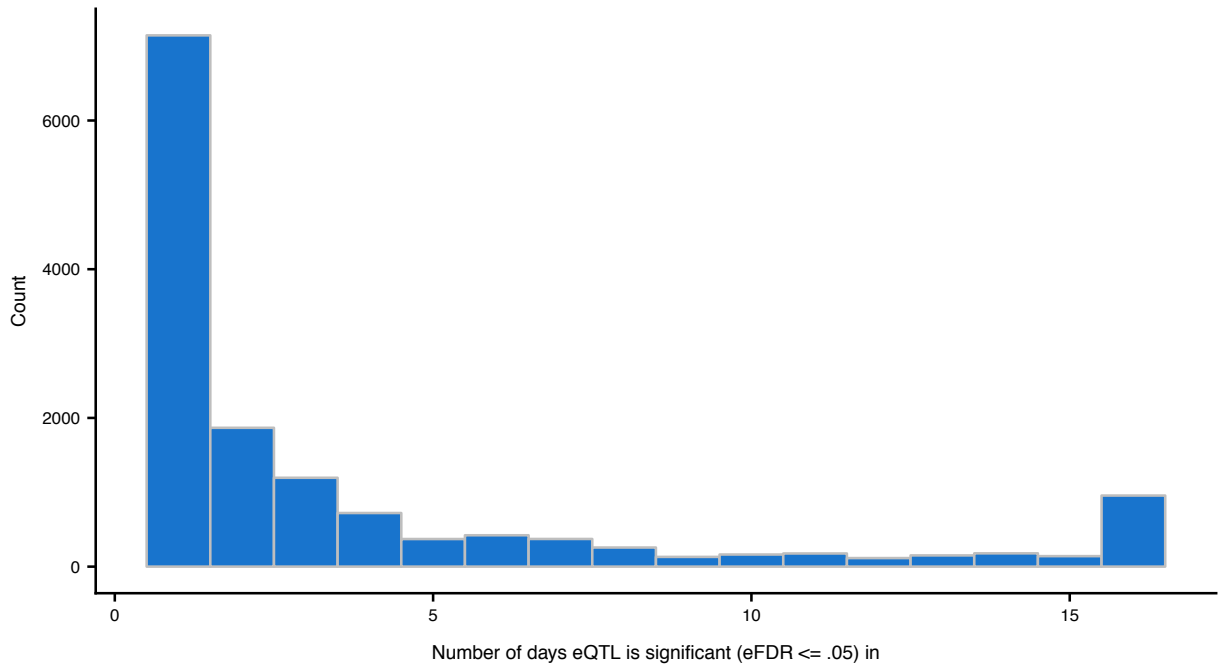

Figure S12: eQTL sharing across time points. The number of days in which each non-dynamic eQTL is significant (eFDR <= .05) for all variant-gene pairs that are significant in at least one day.

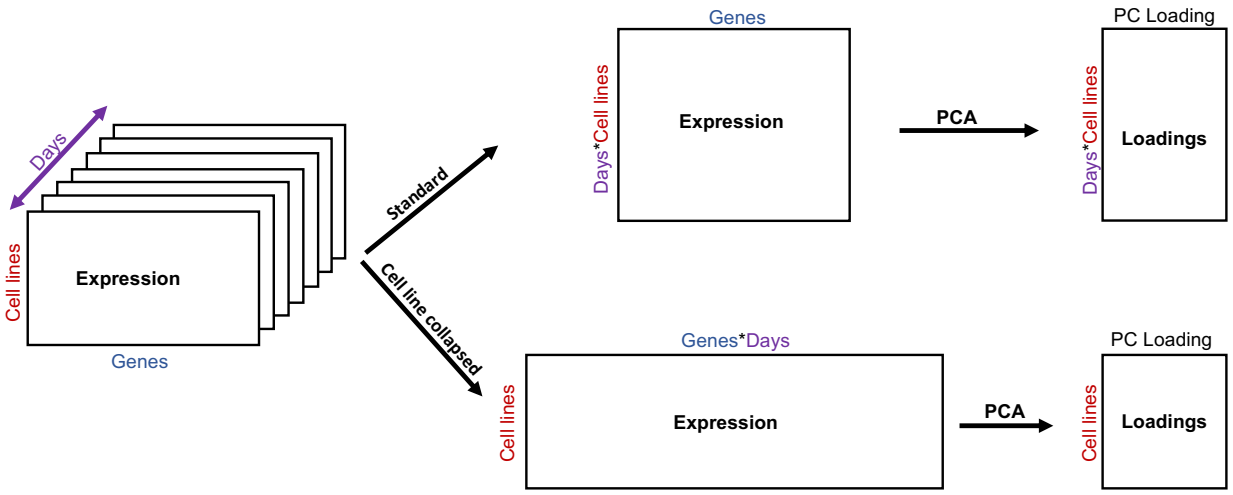

Figure S13: Overview of cell line collapsed PCA. Gene expression can be represented as a three-dimensional matrix spanning days, cell lines, and genes. For standard PCA (top row), we rearrange this gene expression matrix such that rows now correspond to cell lines at specific days (e.g., RNA-seq samples) and columns correspond to genes. Here, PCA will learn a low dimensional representation for cell lines at specific days. For cell line collapsed PCA (bottom row), we rearrange this gene expression matrix such that rows now correspond to cell lines and columns correspond to genes at specific days. Here, PCA will learn a low dimensional representation for each cell line.

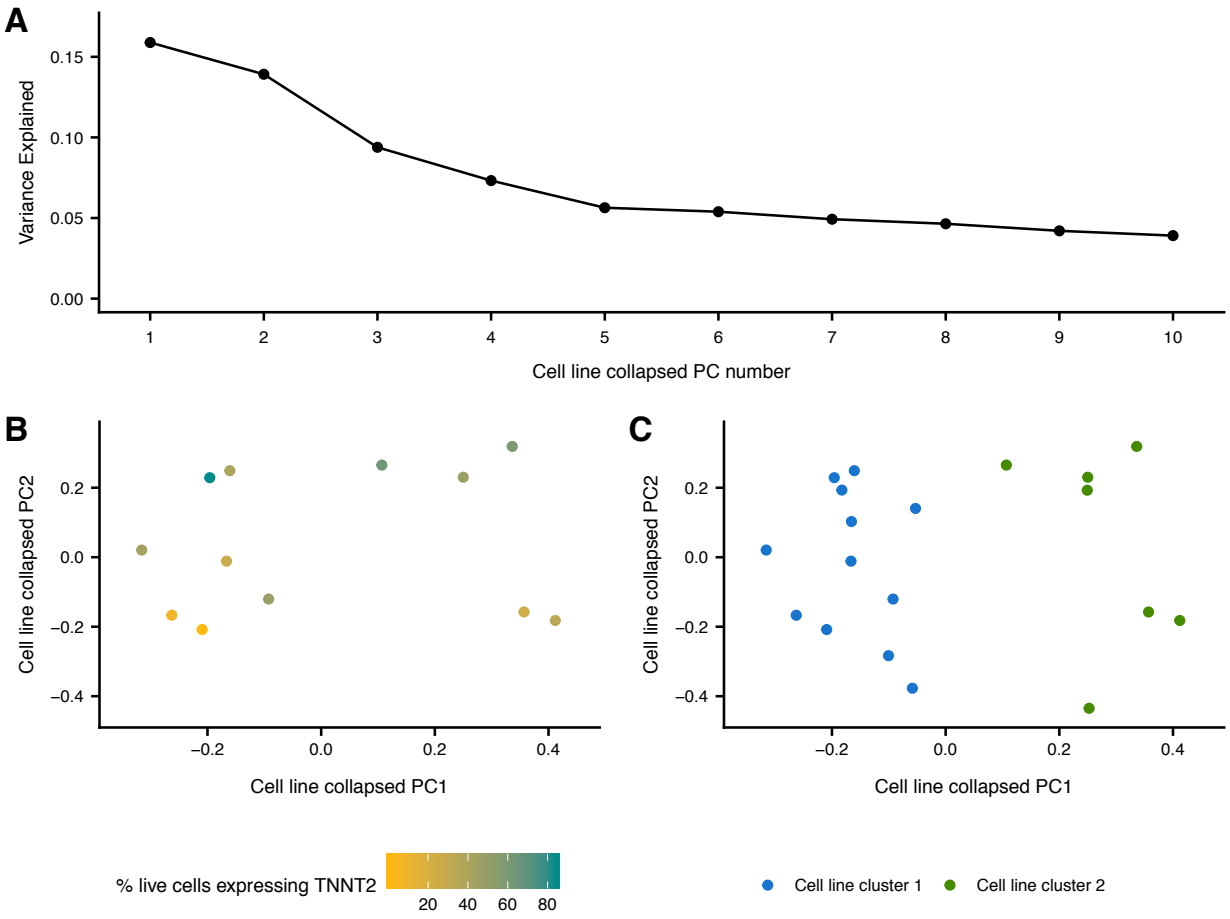

Figure S14: Analysis of cell line collapsed PCs. (A) Variance explained of gene expression by first 10 cell line collapsed principal components. (B, C) First two cell line collapsed principal components where each data point is a cell line colored by its (B) percentage of live cells expressing TNNT2 at time point 15 and (C) split-GPM cell line cluster assignment.

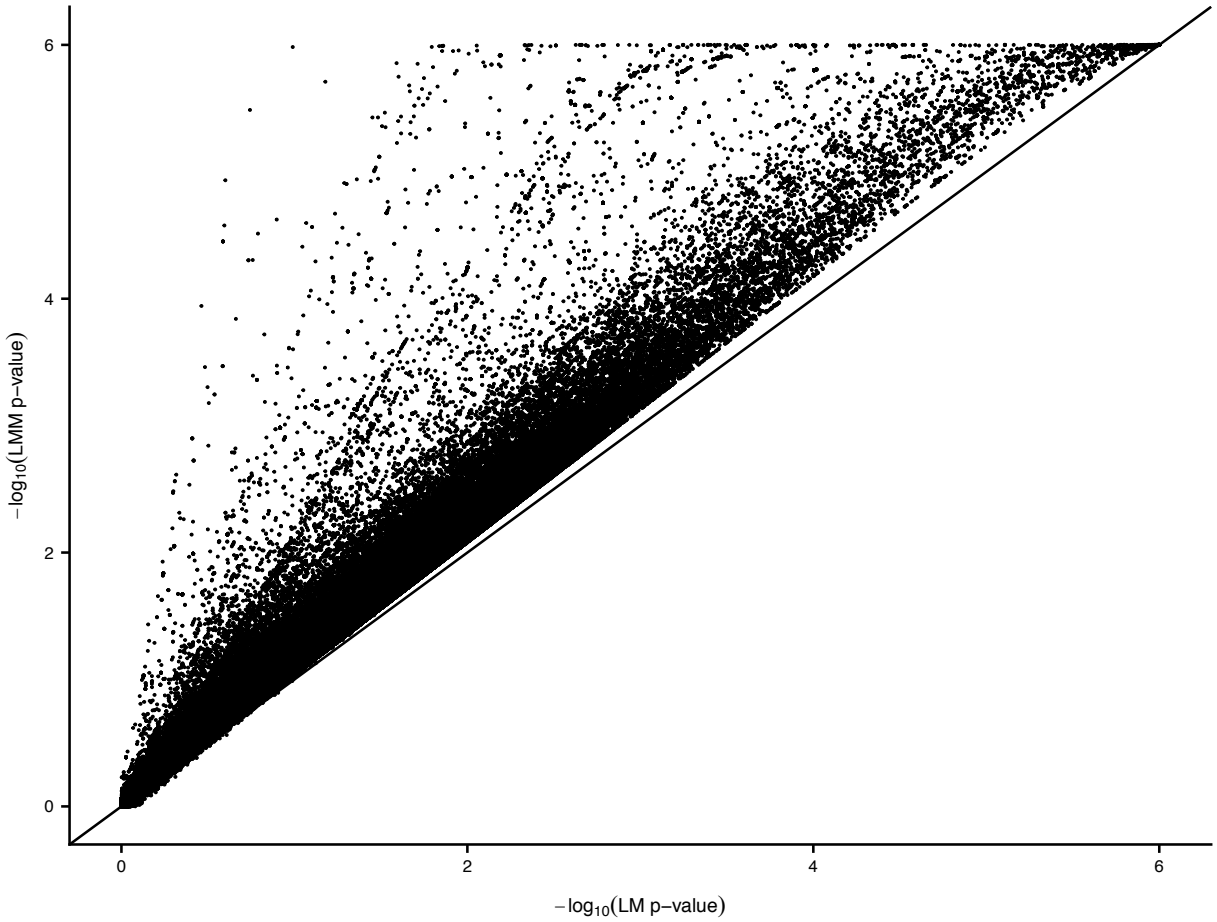

Figure S15: Detecting dynamic eQTLs with gaussian linear mixed model: Comparison of linear dynamic eQTL p-values between gaussian linear model (x-axis) and gaussian linear mixed model with cell line specific random effect (y-axis) across all tested variant-gene pairs (Pearson correlation=.983).

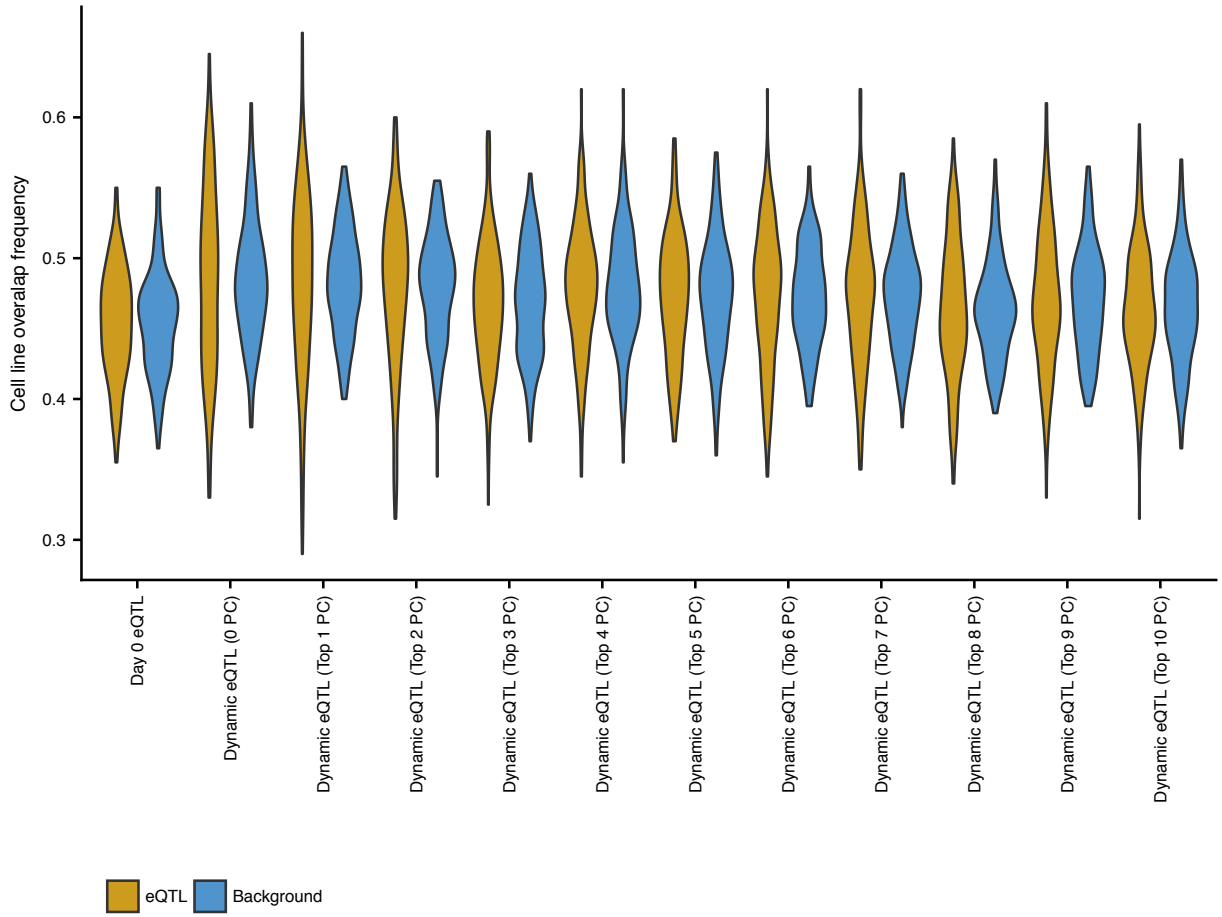

Figure S16: Frequency of cell line overlap in genotype bins. Frequency at which each cell line pair is in the same genotype bin ({0,1,2}) across the strongest associated variants of the 200 most significant eQTL genes (gold) compared to MAF-matched randomly selected background variants (blue). Analysis shown for linear dynamic eQTLs while controlling for a range of the top cell line collapsed PCs. Non-dynamic eQTLs (from day 0) are also shown as a control.

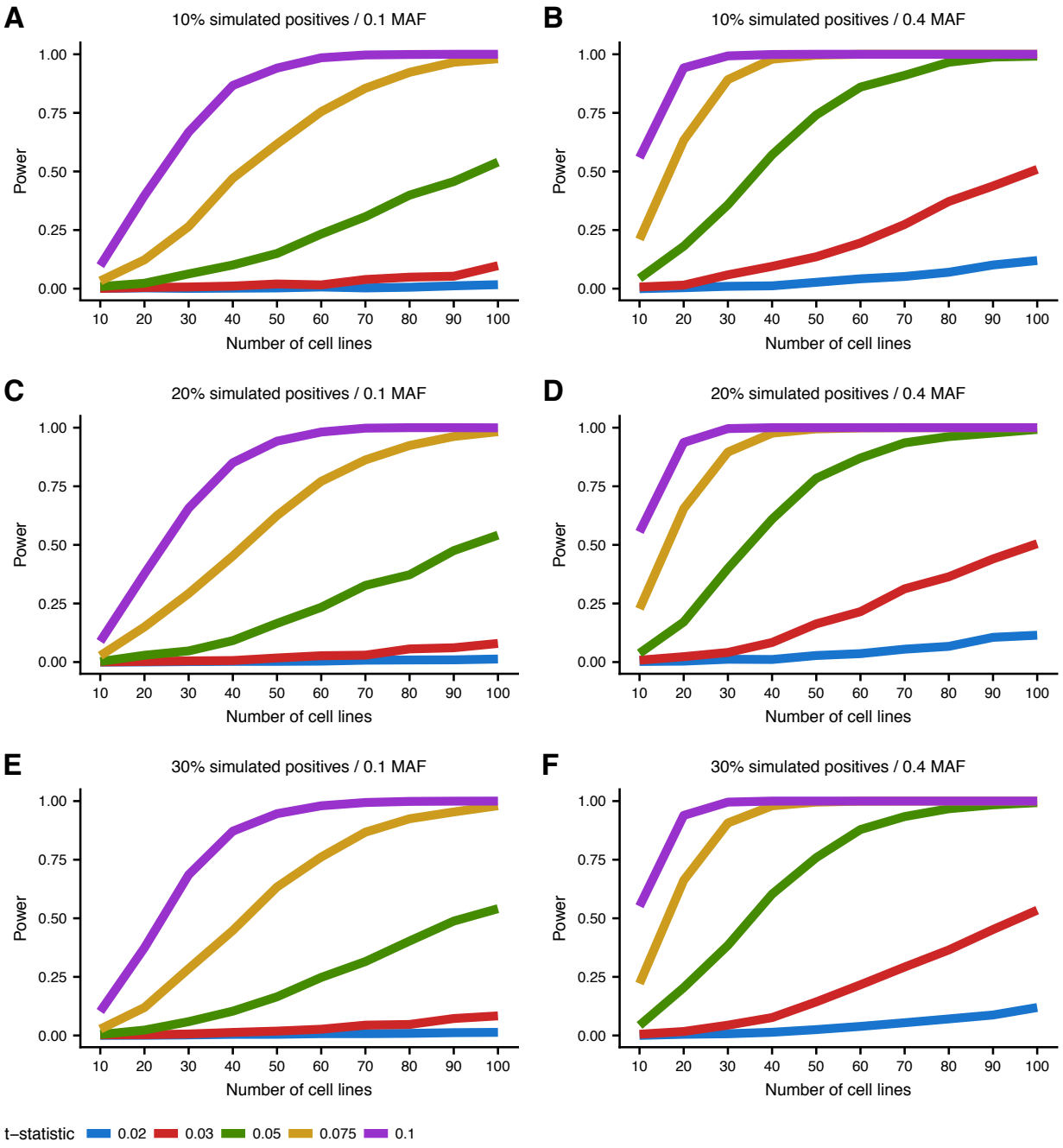

Figure S17: Simulated power analysis for linear dynamic eQTLs. Power to detect simulated linear dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). t-statistic represents the ratio of the effect size of the interaction term and the standard deviation term used to simulate the expression data. We additionally vary (A-F) both the simulated MAF (columns) and the proportion of those tests that were simulated according to the alternative hypothesis (true dynamic eQTLs; rows).

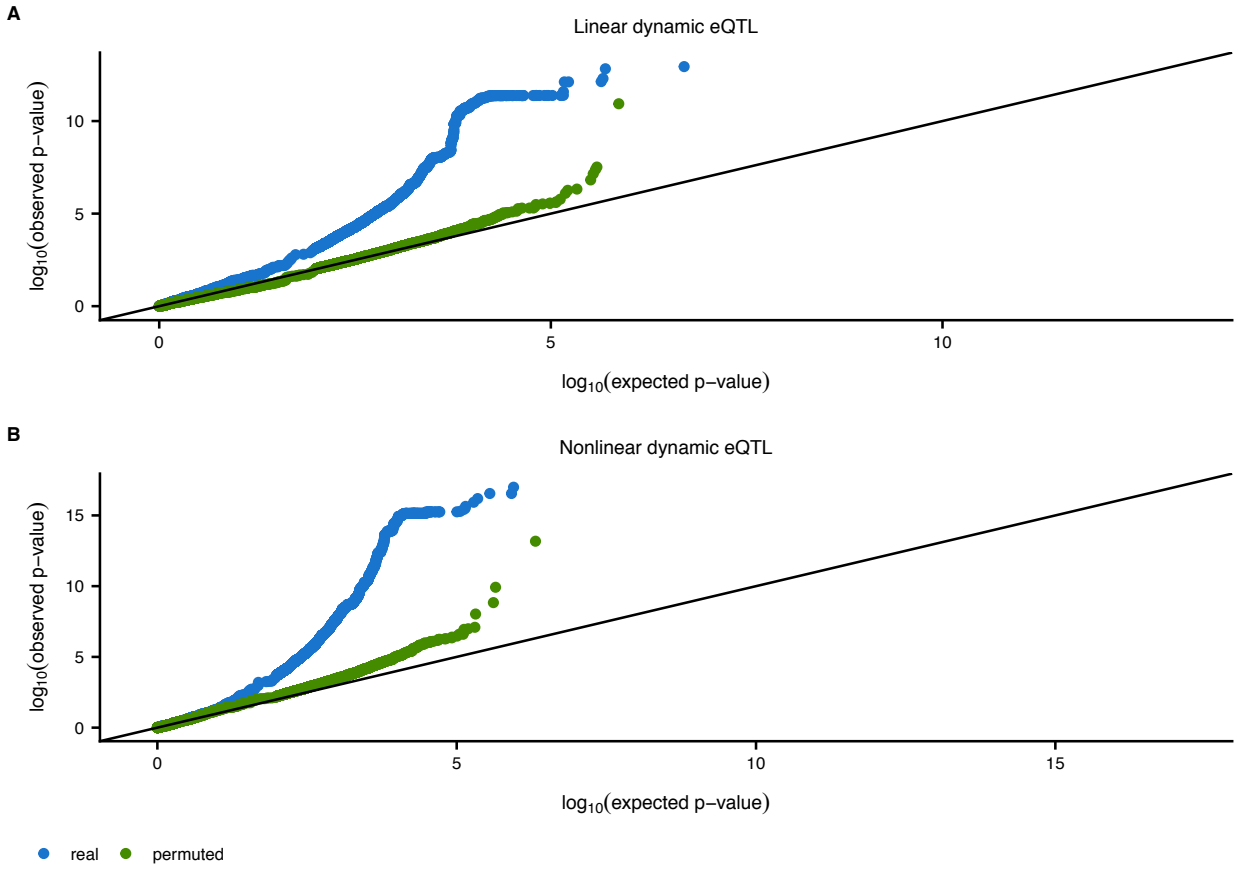

Figure S18: Q-Q plots for linear and non-linear dynamic eQTLs. Q-Q plot for (A) linear and (B) non-linear dynamic eQTLs. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data relative to uniformly distributed p-values.

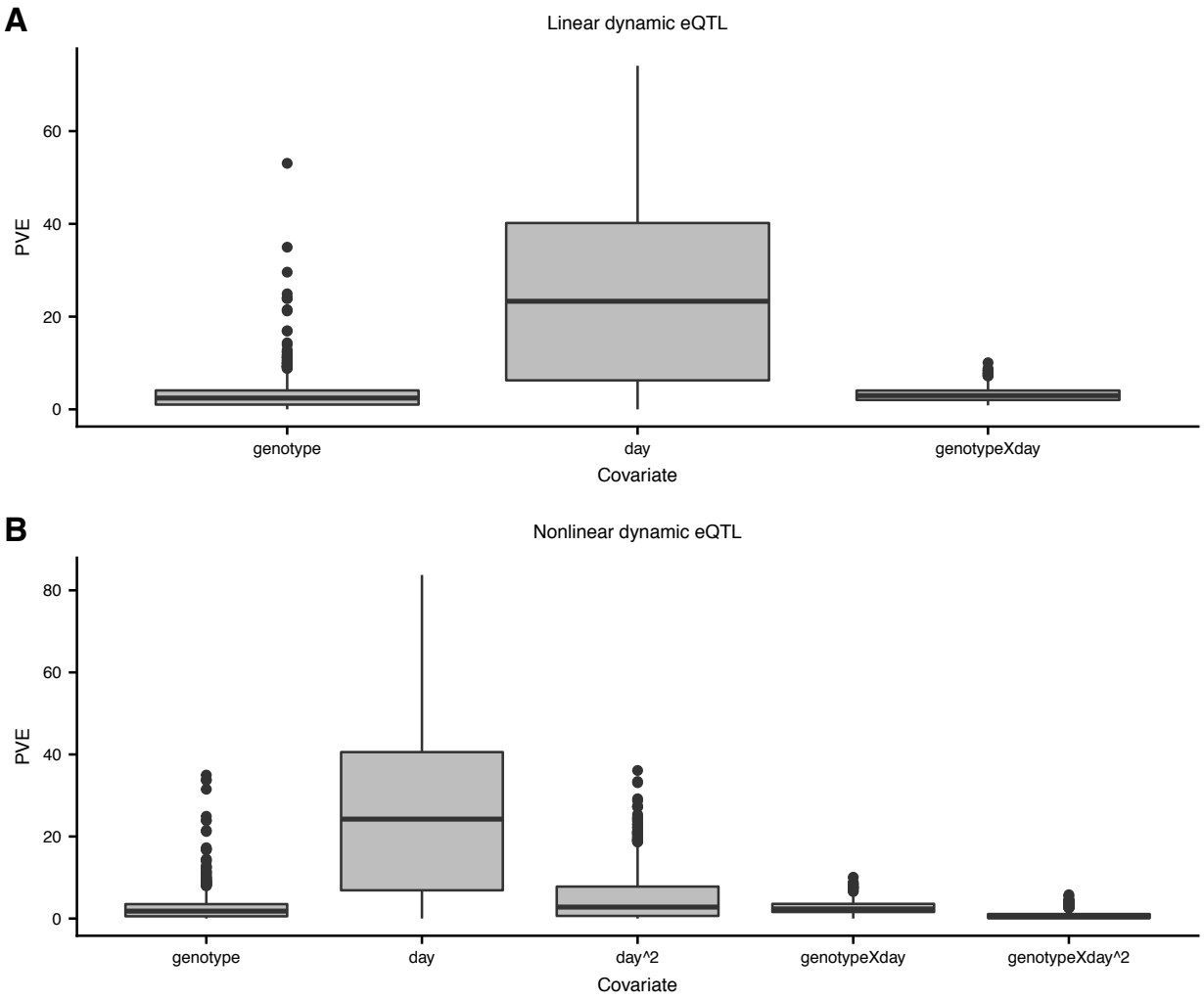

Figure S19: Percent variance explained of dynamic eQTL covariates. Distribution of percent variance explained (PVE; y-axis) of each covariate (x-axis) across significant (eFDR <= .05) (A) linear dynamic eQTLs and (B) nonlinear dynamic eQTLs. For linear dynamic eQTLs, the interaction term (genotypeXday) explains on average 3.16 % of the variance. For nonlinear dynamic eQTLs, the linear interaction term (genotypeXday) and the nonlinear interaction term (genotypeXday^2) explain on average 2.69 and 0.78 % of the variance, respectively. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For linear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, and then genotypeXday. For nonlinear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, day^2, genotypeXday, and then genotypeXday^2.

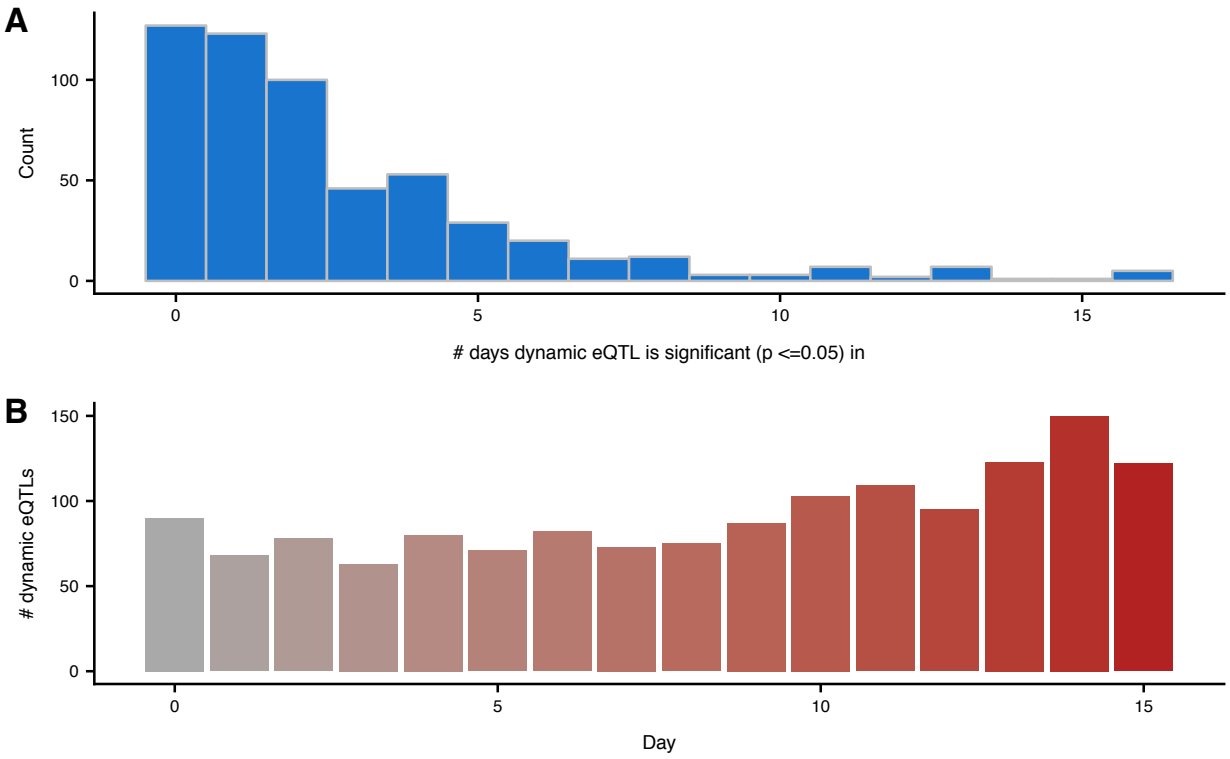

Figure S20: Comparing linear dynamic eQTLs to non-dynamic eQTLs. (A) The number of time points in which the dynamic eQTLs (most significant variant per dynamic eQTL gene) have a nominally significant (p <= .05) non-dynamic eQTL. (B) The number of dynamic eQTLs (most significant variant per dynamic eQTL gene) that are nominally significant (p <= .05) in each time point.

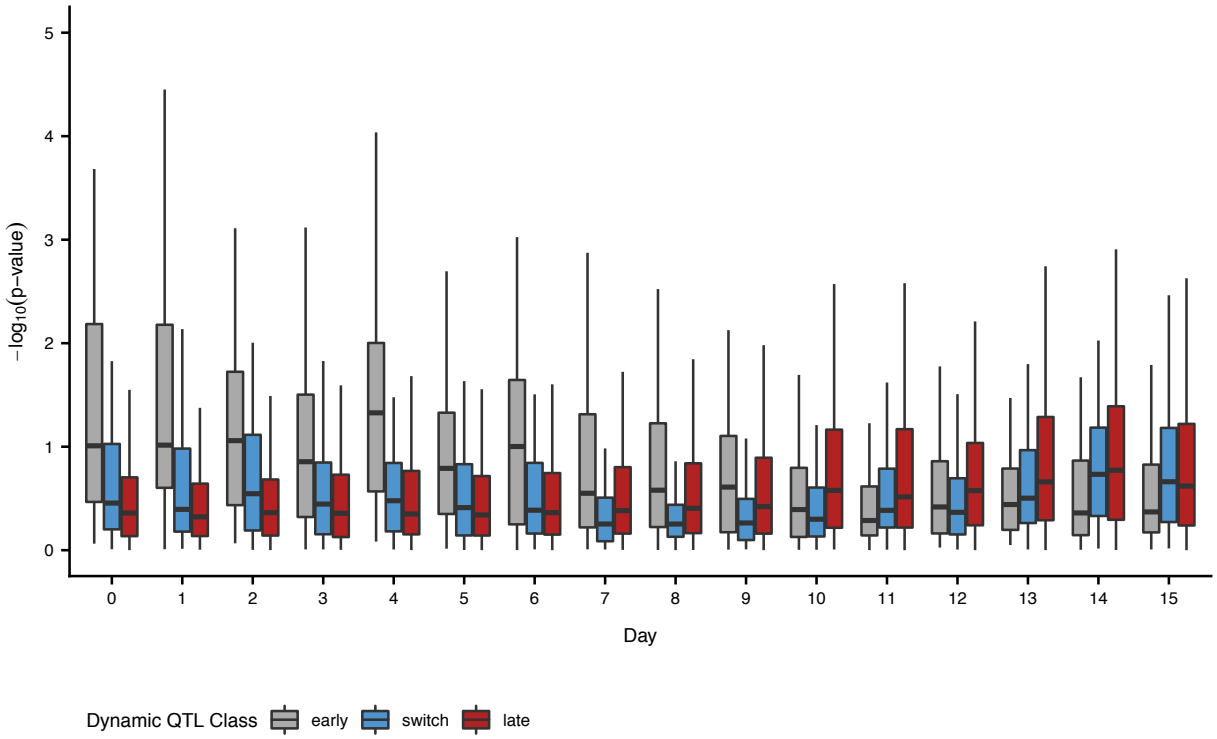

Figure S21: Comparing linear dynamic eQTLs with non-dynamic eQTLs: Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of linear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by linear dynamic eQTL classifications (early, switch, and late).

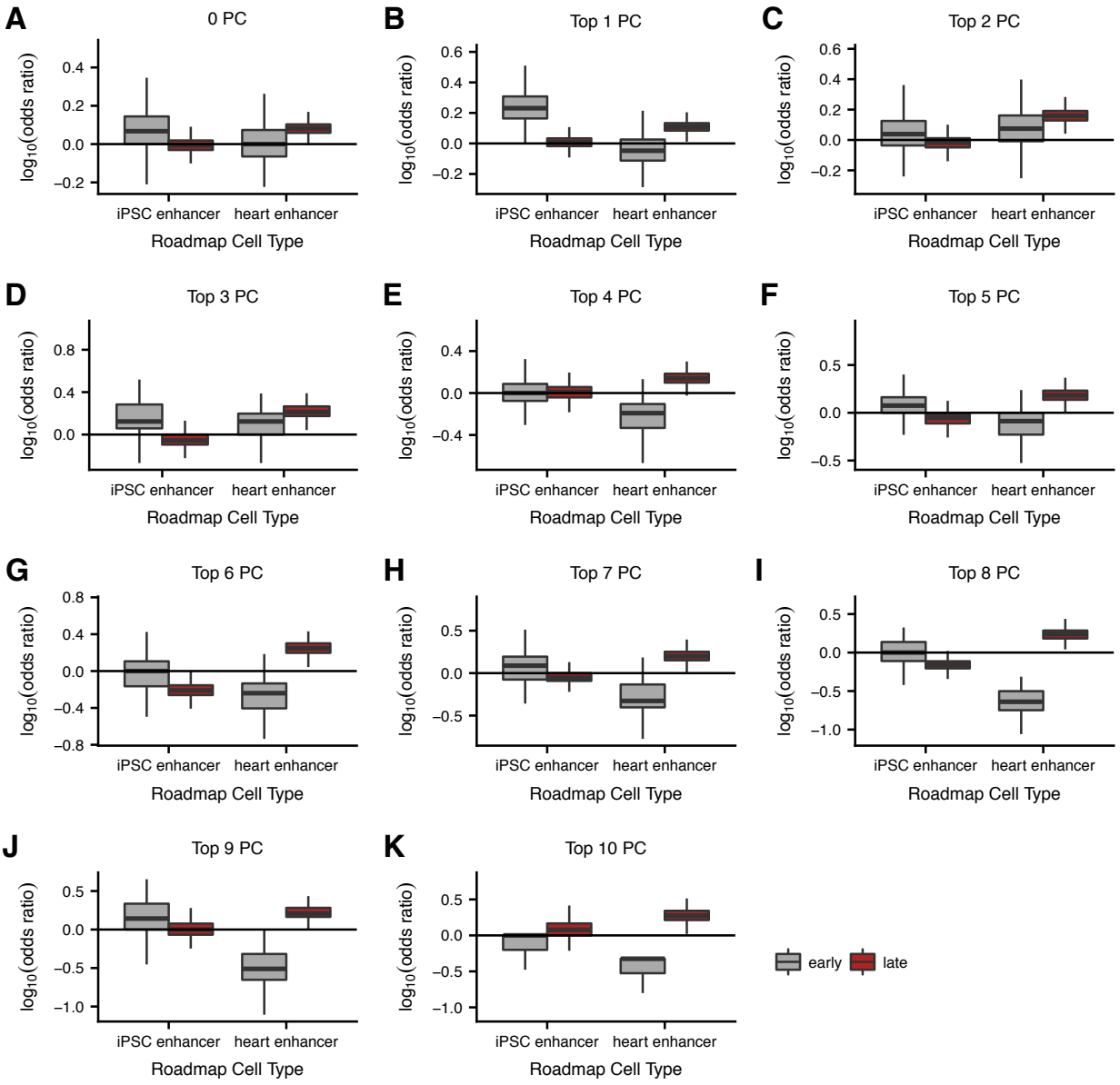

Figure S22: Dynamic eQTL enhancer enrichment. Enrichment of dynamic eQTLs within cell type specific chromHMM enhancer elements relative to 1000 sets of randomly selected background variants matched for distance to transcription start site and minor allele frequency. Dynamic eQTLs were classified as early (eQTL effect size decreasing over time) or late (eQTL effect size increasing over time). Analysis shown for linear dynamic eQTLs while controlling for a range of the top cell line collapsed PCs (A-K).

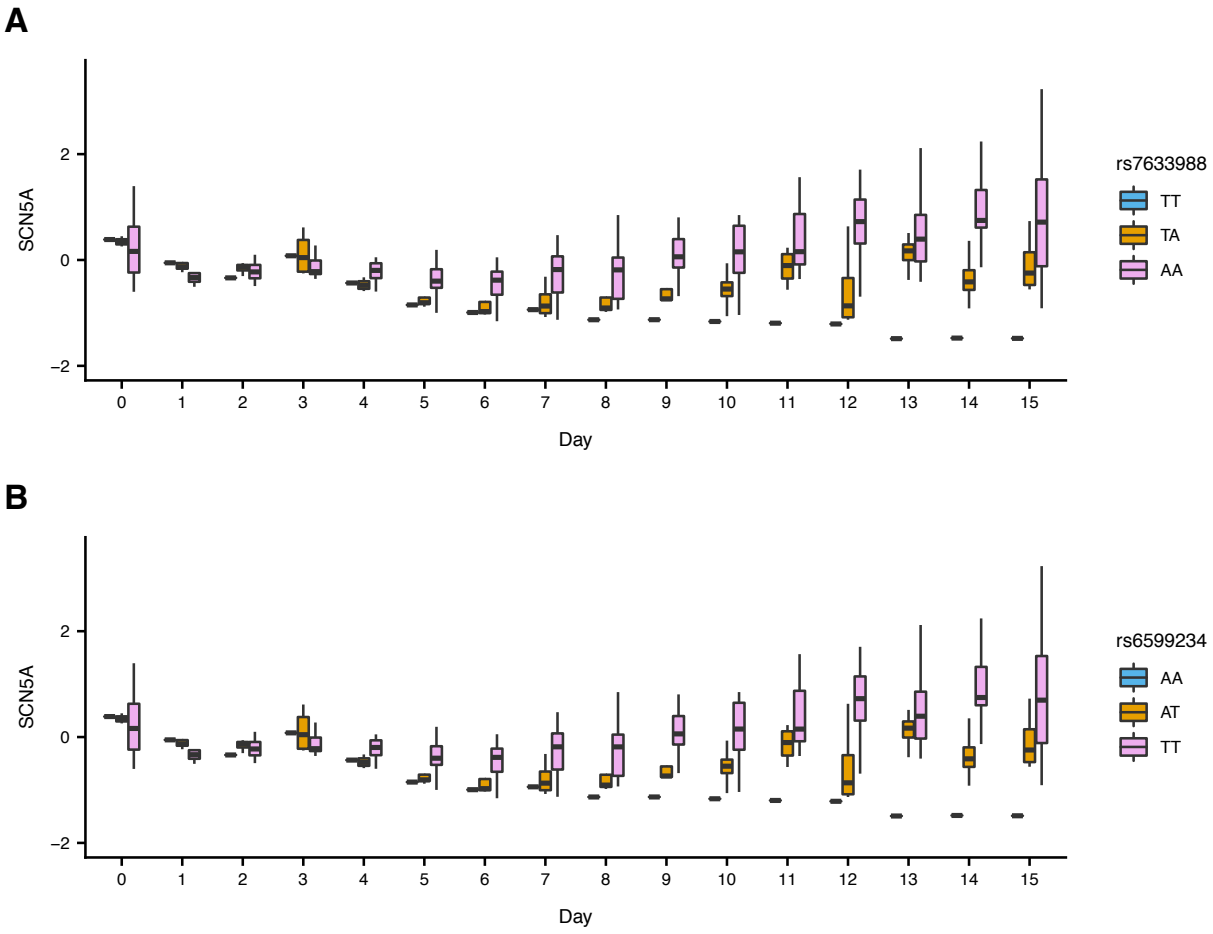

Figure S23: Two significant linear dynamic eQTLs are known GWAS variants. Linear interaction association between time point (x-axis) and genotype (color) of (A) rs7633988 and (B) rs6599234 on residual gene expression (cell line effects regressed on expression) of *SCN5A* (y-axis).

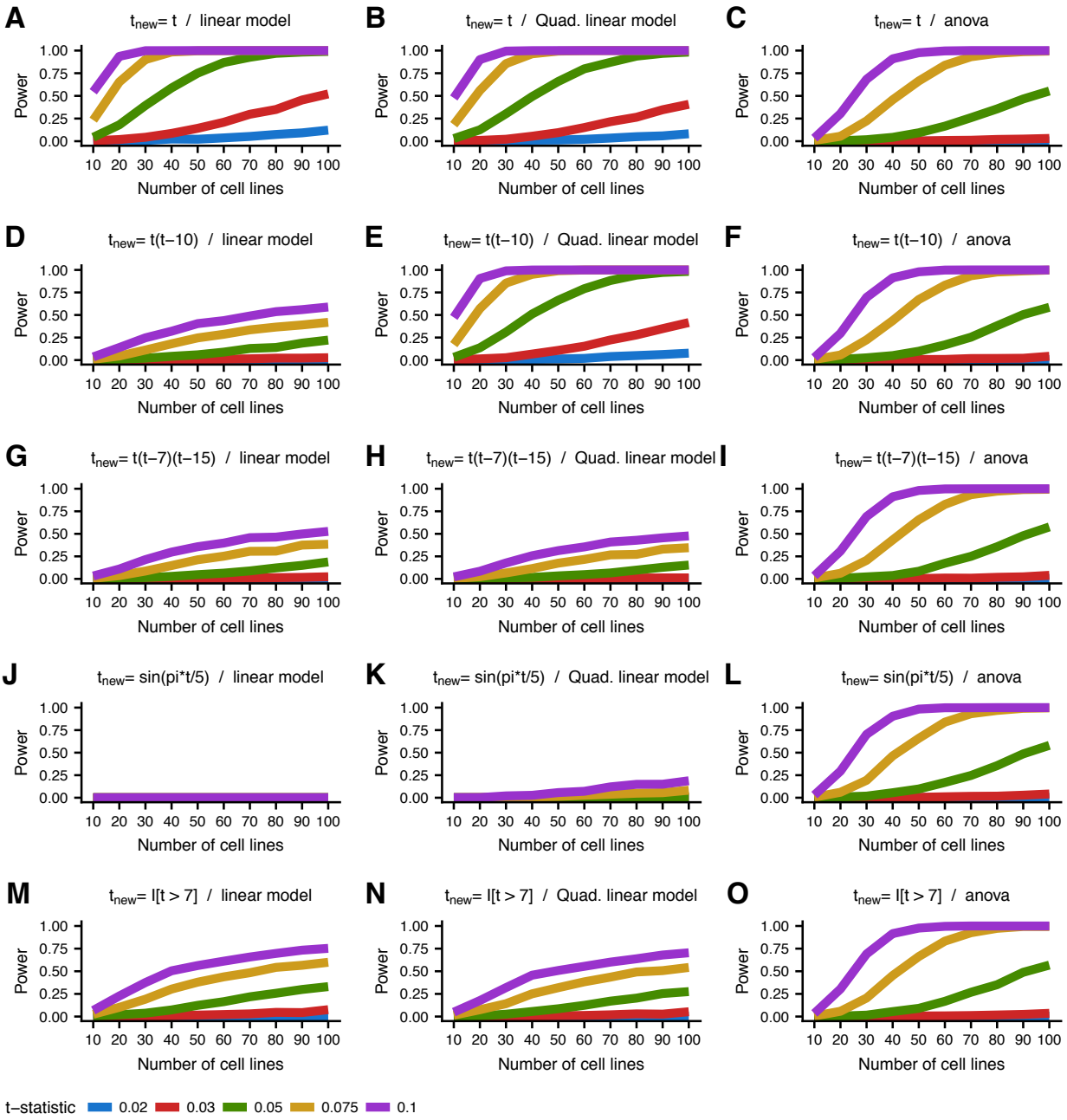

Figure S24: Non-linear simulated power analysis. Power to detect simulated dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). t-statistic represents the ratio of the effect size of the interaction term and the standard deviation term used to simulate the expression data. Simulated expression was generated based on various transformations (t_{new}; rows) of the original values of differentiation time (t). Transformed differentiation time was scaled to have the same standard deviation as the original values of differentiation time. Three different statistical models were used to identify dynamic eQTLs (columns): linear model (linear dynamic eQTL), quadratic linear model (nonlinear dynamic eQTL), and categorical ANOVA analysis. The simulated MAF was .4 and 30% of all simulated tests were drawn from the alternative hypothesis.

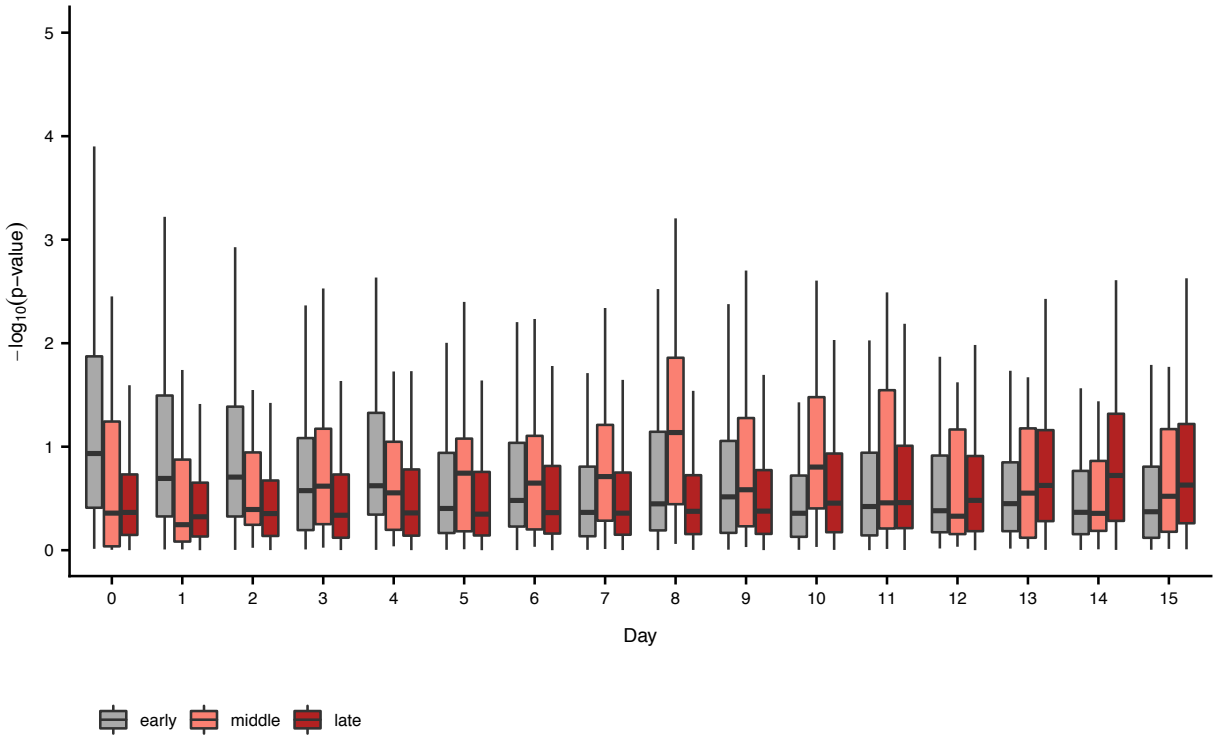

Figure S25: Comparing nonlinear dynamic eQTLs to non-dynamic eQTLs. Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of nonlinear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by nonlinear dynamic eQTL classifications (early, middle, and late).

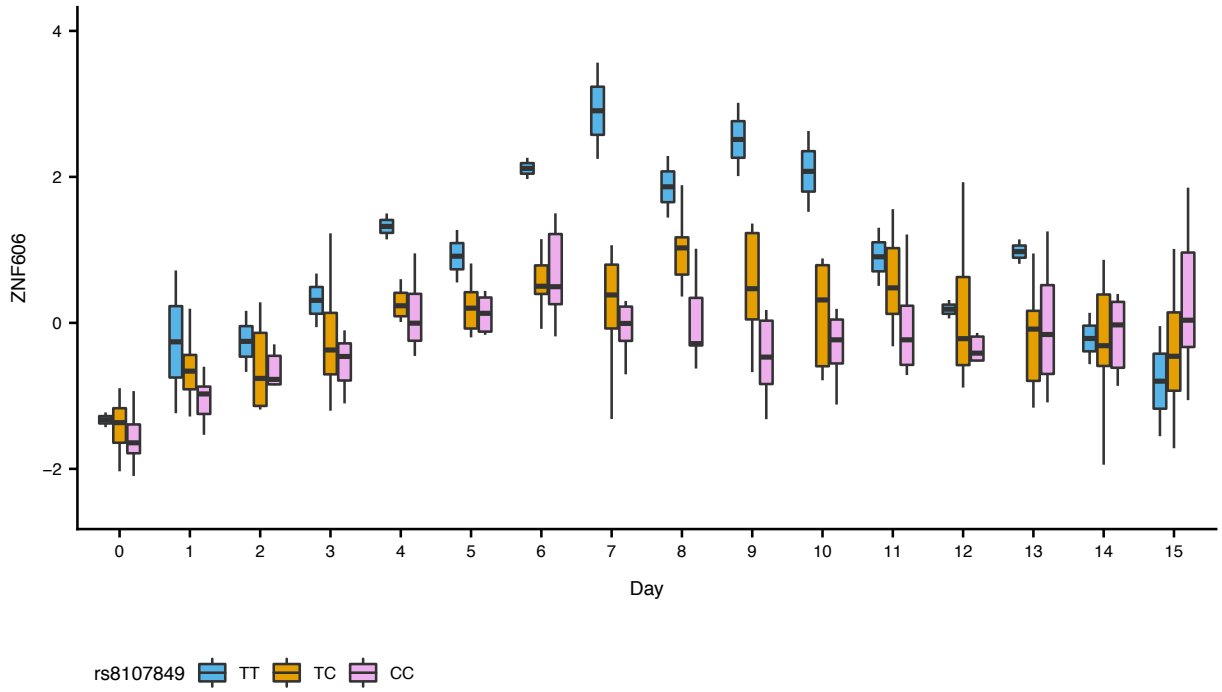

Figure S26: Middle dynamic eQTL example: Nonlinear interaction association between genotype (color) of rs8107849 and time point (x-axis) on residual gene expression (cell line effects regressed on expression) of *ZNF606* (y-axis).

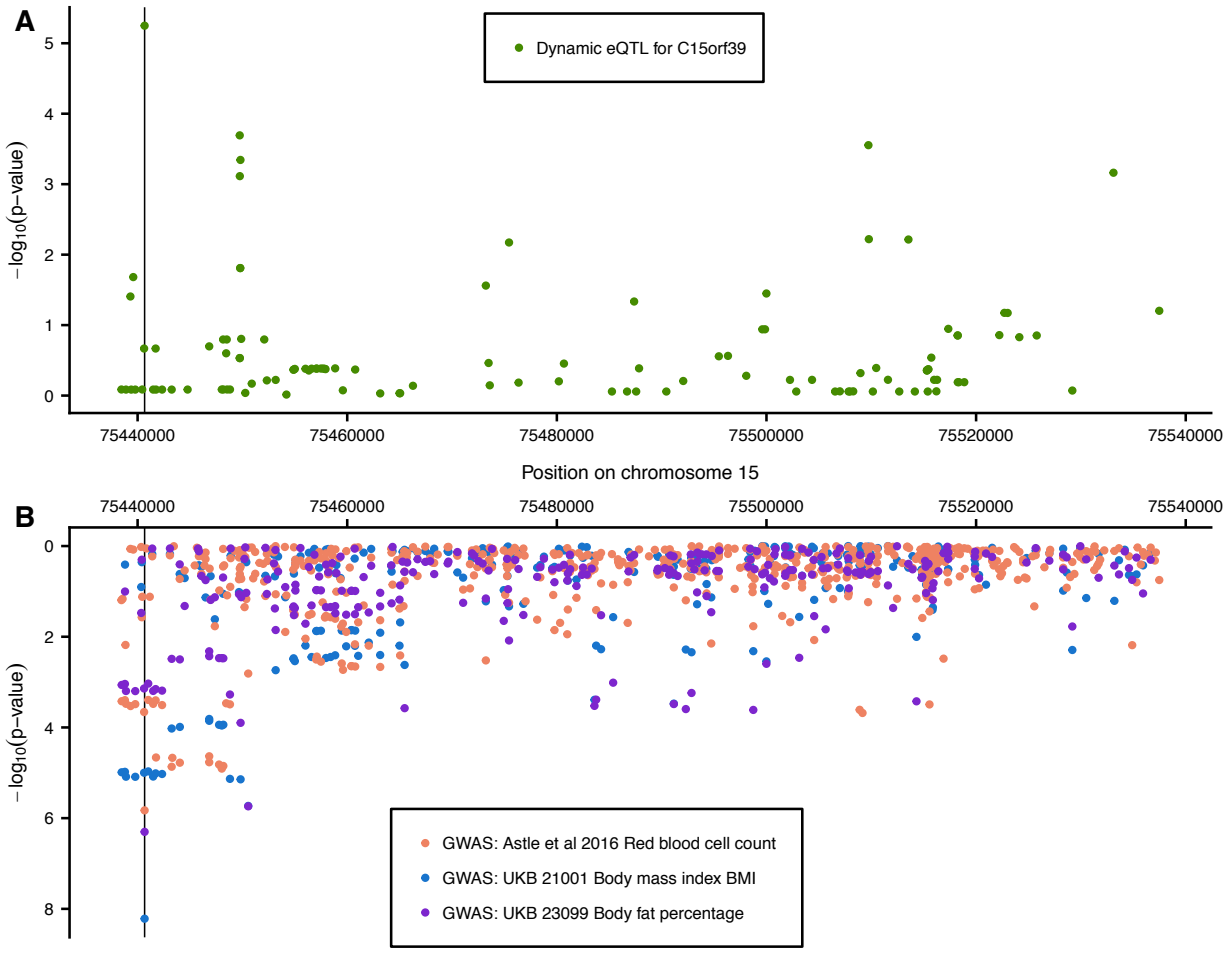

Figure S27: Nonlinear dynamic eQTL overlaps GWAS variant: (A) Manhattan plot showing interaction association p-values for *C15orf39* according to nonlinear dynamic eQTL calling for all variants tested within 50KB of the *C15orf39* transcription start site. (B) Manhattan plot showing GWAS p-values on the same region surrounding *C15orf39* from three different GWAS studies (colors) (*23, 24*). Vertical line depicts genomic location of most significant nonlinear dynamic eQTL (rs28818910) for *C15orf39*. p-values shown for body mass index and body fat percentage are based on round 1 of UK Biobank (UKB) (*23*). Body mass index and body fat percentage p-values for rs28818910 according to the round 2 of UKB (*31*) become slightly less extreme (p=1.322e-07 and p= 2.521e-06, respectively), but are still significant after multiple testing correction for all significant (eFDR <= .05) nonlinear dynamic eQTL variants (Bonferroni p= 0.000902 and Bonferroni p=.0172, respectively).

Supplementary Tables

Table S1. Available as an excel file online. Sheet ‘A-Sample meta-data’ contains meta-data for each RNA-seq sample. Sheet ‘B-meta data description’ contains descriptions of each meta-data variable collected.

Cell Line	Percent of Live Cells Expressing TNNT2
18489	44.3
18499	24.2
18505	NA
18508	83.9
18511	NA
18517	47.8
18520	NA
18855	NA
18858	NA
18870	NA
18907	7.9
18912	47.8
19093	27
19108	NA
19127	1.1
19159	39.8
19190	63.2
19193	59.5
19209	33.4

Table S2: Flow cytometry results for each cell line at day 15 of cardiomyocyte differentiation. The percent of live cells expressing cardiac troponin (TNNT2) for every cell line at day 15 of differentiation. Cells with an NA indicate that flow cytometry was not performed on this cell line.

Hallmark gene set	Gene cluster 2	Gene cluster 4	Gene cluster 5	Gene cluster 6	Gene cluster 9	Gene cluster 11	Gene cluster 13	Gene cluster 16
TNFA signaling via NFKB	1	1	1	.000208	1	1	1	1
Mitotic spindle	1	1	1	1	.0166	1	1.80e-14	1
TGF beta signaling	1	1	1	.348	.000624	1	1	1
DNA repair	1	1	.000242	1	1	1	3.73e-7	1
G2M checkpoint	1	1	1	1	1	1	2.87e-63	.594
Myogenesis	9.29e-14	1	1	1.05e-5	1	1	1	1
Protein secretion	.00384	1	1	1	1	1	1	1
Complement	1	1.98e-5	1	1	1	1	1	1
Unfolded protein response	1	1	6.99e-5	1	1	1	1	1
MTORC1 signaling	1	1	2.07e-10	1	1	.696	1	1
E2F targets	1	1	.0111	1	1	1	5.47e-73	.0458
MYC targets V1	1	1	3.03e-25	1	1	.329	1.28e-16	1.16e-5
MYC targets V2	1	1	7.04e-21	1	1	1	.981	1
Epithelial mesenchymal transition	1	.000310	1	2.05e-5	1	1	1	1
Xenobiotic metabolism	1	.000435	1	1	1	1	1	1
Oxidative phosphorylation	1	1	1	.134	1	8.11e-11	1	1
Heme metabolism	1.24e-6	1	1	1	1	1	1	1
Coagulation	1	1.72e-16	1	1	1	1	1	1
Bile acid metabolism	1	.00392	1	1	1	1	1	1
Spermatogenesis	1	1	1	1	1	1	.00433	1
KRAS signaling up	1	.00536	1	.622	1	1	1	1

Table S3: Hallmark gene set enrichment of split-GPM gene clusters: Bonferroni corrected p-values (Fisher's exact) from gene set enrichment of gene clusters (columns) from split-GPM within Hallmark gene sets (rows). Only gene clusters and gene sets with at least one significant enrichment (Bonferroni p-value <= .05) are shown.

# of cell line collapsed PCs	# genes with significant dynamic eQTL (eFDR <= .05)	# genes with significant dynamic eQTL (eFDR <= .01)
0	2256	931
1	1943	785
2	1247	294
3	648	250
4	608	186
5	550	150
6	533	113
7	556	212
8	456	110
9	288	22
10	213	79

Table S4: Number of linear dynamic eQTLs detected. The number of genes with a significant linear dynamic eQTL (eFDR <= .05 and eFDR <= .01) as a function of the number cell line collapsed PCs used as covariates.

Table S5. Available as a text file online. This table reports the percent variance explained (PVE) by the linear dynamic eQTL model's fixed effects (excluding fixed effects related to cell line collapsed PCs) for all significant (eFDR <= .05) linear dynamic eQTLs. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For linear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, and then genotypeXday.

Hallmark gene set	0 PCs	1 PC	2 PCs	3 PCs	4 PCs	5 PCs
KRAS signalling dn	.0076	.0007	.472	1.0	1.0	1.0
Hypoxia	1	1	.33	.00095	.0048	.02
Myogenesis	.91	.01	1	.055	.011	.002
Interferon Gamma Response	1	.08	.39	.39	.086	.016

Hallmark gene set	6 PCs	7 PC	8 PCs	9 PCs	10 PCs
KRAS signalling dn	1.0	1.0	1.0	1.0	1.0
Hypoxia	.33	.022	1.0	1.0	.33
Myogenesis	.24	.055	.055	1.0	1.0
Interferon Gamma Response	.086	.0026	.086	1.0	.39

Table S6: Hallmark gene set enrichment of linear dynamic eQTLs. Bonferroni corrected p-values (Fisher's exact) from gene set enrichment within Hallmark gene sets (rows) of the 200 genes with the strongest linear dynamic eQTLs as a function of the number of cell line collapsed PCs used as covariates (columns). Only Hallmark gene sets with at least one significant enrichment (Bonferroni p-value <= .05) are shown.

Number of cell line collapsed PCs	Enrichment p-value
0	.08
1	.01
2	.01
3	.00099
4	6.8e-5
5	.00099
6	.01
7	.00099
8	.08
9	.08
10	.08

Table S7: Dilated cardiomyopathy gene set enrichment of linear dynamic eQTLs. p-values (Fisher's exact) from gene set enrichment within dilated cardiomyopathy gene set of the 200 genes with the strongest linear dynamic eQTLs as a function of the number of cell line collapsed PCs used as covariates.

Table S8. Available as a text file online. This table reports the percent variance explained (PVE) by the nonlinear dynamic eQTL model's fixed effects (excluding fixed effects related to cell line collapsed PCs) for all significant (eFDR <= .05) nonlinear dynamic eQTLs. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For nonlinear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, day^2, genotypeXday, and then genotypeXday^2.