A hierarchical Bayesian mixture model for distinguishing active gene expression from transcriptional noise

Ammon Thompson[1,*], Michael R. May[1], Brian R. Moore[1], Artyom Kopp[1]

Department of Evolution and Ecology, University of California, Davis, Storer Hall, One Shields Avenue, Davis, CA 95616, USA

* ammthompson@ucdavis.edu

Abstract.—Identifying the genes that are expressed in each organ and cell is essential to understanding the regulatory mechanisms that translate genetic information into organ function. However, determining which genes are active in a given tissue is challenging because gene regulation, like all biological processes, is noisy. Inactive genes, through rare stochastic events, can still produce detectable numbers of transcripts in RNA-sequencing samples. Conversely, active genes can go undetected in some samples due to technical factors. There is a clear need for robust and objective methods for establishing the boundary between active gene expression and background noise. We describe a hierarchical Bayesian mixture model method that infers the distributions of expression levels exhibited by active and inactive genes in a set of replicate transcriptomes, and estimates the probability that each gene in the genome is in the active or inactive state in the corresponding tissue. Expression probabilities estimated in this fashion reflect an inherent property of the tissue, can be integrated over any number of independent experiments, and can be compared among any number of species and tissues. By analyzing empirical and simulated datasets, we show that model-based inferences of gene expression states make efficient use of experimental replication and are consistent with other sources of information such as chromatin marks and data derived from developmental-genetic studies. We discuss several avenues of research in developmental and evolutionary genomics that will benefit from the analysis of gene expression states. As an example, we identify several genes that underwent a qualitative gain of expression in specific brain regions in human evolution. We implemented our method as a freely available R package called zigzag.

INTRODUCTION

Despite having identical genomes, the cells and tissues within a multicellular organism vary in the ensemble of genes they express; this variation is derived from differences in the presence/absence and concentrations of RNA and proteins within each cell. Much of what we understand about the genomic mechanisms driving diverse cellular phenotypes is based on assays of the molecular intermediate between genome and proteome: the transcriptome. RNA sequencing has made transcriptome analysis tractable on a massive scale, providing opportunities to determine the regulatory mechanisms responsible for cell differentiation and phenotypic evolution. During both development and evolution, tissue-specific transcriptomes can change at two levels: qualitatively, when the expression state of a gene in a particular cell is changed in a discrete, on/off fashion, or quantitatively, through up- or down-regulation of genes that are already expressed in that tissue. The relative contributions of qualitative and quantitative transcriptome variation to phenotypic differences between tissues are not clear. In fact, this intriguing question cannot even be addressed without robust and objective methods for identifying qualitative differences in transcriptome composition. Unfortunately, such methods are still largely lacking.

In addition to distinguishing different cell types, qualitative gain and loss of transcripts from an organ can potentially lead to major phenotypic transitions, perhaps more so than quantitative variation in transcript abundance. Due to additional layers of regulatory control, the relationship between transcript abundance and protein abundance is often complex. In particular, changes in transcript abundance can be buffered by post-transcriptional regulation of protein abundance on both physiological and evolutionary time-scales [1–12]. For example, the PSMD1 gene shows significant differences in transcript abundance, but not protein abundance, among humans, chimpanzees, and rhesus macaques [13]. More generally, the correlation between transcript and protein levels is modest, measured at less than 0.5 in multiple studies of different species [14–16]. In both development and evolution, a qualitative difference in gene-expression state may be more likely to induce a qualitative change in cellular phenotype, since there is little room for post-transcriptional compensation when a transcript is absent. Thus, genes that are expressed in tissue- or cell-restricted patterns are likely candidates for the unique characteristics of those tissues and cells.

A qualitative examination of gene presence/absence in the transcriptomes of different organs or species may usefully augment traditional comparative RNA-seq analyses. However, it remains surprisingly difficult to infer which genes are, and are not, active, especially at low expression levels. Simply put, a read count of zero for a given gene does not necessarily indicate that it is inactive, while a non-zero read count does not necessarily indicate that it is actively expressed. Random transcriptional noise can often yield reads that map to inactive genes; conversely, insufficient sampling (low sequencing depth) can cause active genes to be absent among mapped reads [17–27]. This produces an intrinsic ambiguity between background noise and active but low-abundance genes.

To avoid issues associated with background noise, researchers typically set a hard expression threshold and remove all genes from further analysis that fall below the threshold. A conservative threshold is often used based on estimates linking normalized read counts to the number of transcripts per cell, where fewer than one transcript per cell is often considered as not expressed (*e.g.*, < 1 RPKM/FPKM; [28–30]. Others have used read density in non-genic regions as a proxy for background expression in order to derive an empirical threshold [31]. It is not clear how to accommodate variation across multiple biological replicates of different depth and with non-identical distribution of transcript abundance when setting expression thresholds. The relative expression of genes compared to background noise may also vary across tissues and species, so there is no *a priori* reason to expect a single threshold to be equally appropriate for all samples in a given comparison.

There are signatures in transcriptomic data that provide a way to distinguish active transcrip-

tion from background noise. The relative expression levels of genes in transcriptomes typically fall into two overlapping distributions, of high-abundance and low-abundance genes [18, 32–35]. Genes in the lower expressed population of the bimodal transcriptome distribution tend to have inactive chromatin marks (e.g., H3K27me3), while genes in the higher expressed population have active chromatin marks, (e.g., H3K4me3) [18, 21, 35, 36]. In addition, data from single-cell RNA-seq also indicate that genes exhibit bimodal expression patterns among cells with the lower mode showing few or zero reads, leading researchers to suggest that these patterns indicate the difference between cells actively expressing the gene and cells producing trace amounts of expression due to background stochastic events [27, 37].

These patterns can be exploited to formulate mixture models to more objectively infer the expression state of genes in a given library [27, 34, 36, 38]. Rather than using a hard threshold, the UPC method [38] models expression as a bimodal distribution comprising a mixture of a single inactive and a single active distribution, and uses maximum likelihood and empirical Bayesian statistics to estimate the probability that each gene is actively expressed. Although this approach better accounts for uncertainty in gene activity state [38], this method and those like it [34, 35] can only estimate expression states for individual libraries. Because of natural variation and technical variance, many genes will appear within the noise distribution in some replicate libraries while showing a signal of active expression in others, leaving it unclear how to integrate expression state estimates across multiple libraries. Furthermore, existing approaches rely on point estimates for mixture parameters, and thus do not fully accommodate uncertainty in their estimates. Finally, these approaches assume two-component mixtures, which may be too simple for most datasets where outlier genes with very high expression are often present [39, 40].

In this study, we describe a hierarchical mixture model that accommodates multiple biological and technical variables that may impact patterns of expression within and among replicate RNA-seq libraries. We develop a Bayesian approach to infer the parameters of our hierarchical mixture model from patterns of relative expression across a set of replicate RNA-seq libraries, providing estimates of the posterior probability that each gene in the genome is actively expressed in a given tissue or cell type. We present analyses of simulated datasets that validate the implementation and characterize the statistical behavior of our hierarchical Bayesian mixture model. We also apply our method to several empirical datasets, and demonstrate that the expression-state probabilities inferred using our method are consistent with expectations based on other data types, such as epigenetic marks and developmental-genetic studies.

Method Description

The method relies on relative-expression estimates typically produced at the end of an RNA-seq analysis pipeline using software such as Kallisto or StringTie [41, 42]. These pipelines yield expression estimates that are scaled proportions; each gene-expression level as a proportion of the total number of transcripts. These are ether TPM, which is the proportion $\times 10^6$, or FPKM, which is the proportion scaled by a library-specific scaling factor $\times 10^6$. The data consist, therefore, of G genes in R libraries. We refer to the entire (log-transformed) transcriptomic dataset as $\mathbf{X}$.

Within a transcriptome library, the expression levels of both protein-coding and non-protein-coding genes typically exhibit a characteristic bimodal distribution (Fig. 1 A), where the left mode reflects nonfunctional *background* expression (*e.g.*, promoter leakage), and the right mode reflects *active* transcription, likely associated with a functional role in the cell or tissue.

In any given library, we may fail to detect a given gene owing to biological and/or technical factors; for example, the probability that a given gene is detected in a given library is proportional to the expression level and length of the gene and the sequencing depth of the library (Fig. 1 B). Across multiple libraries, the variance of gene expression shows a characteristic pattern, where the variation in the expression level of a given gene between libraries is inversely proportional to its average expression level among libraries (Fig. 1 C). This leads to individual genes showing

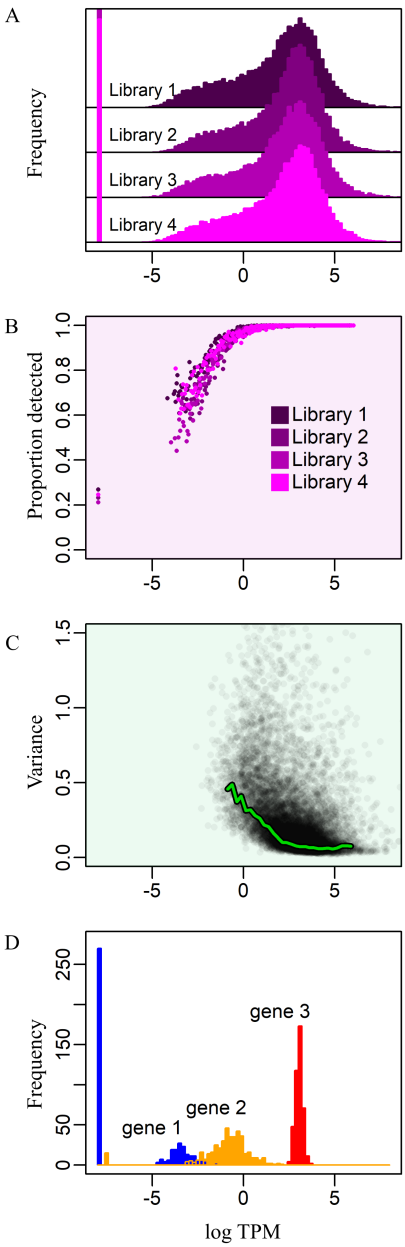

Figure 1: Key properties of RNA-seq data: transcript abundance, probability of detection, and mean-variance relationship. A) The distribution of transcript abundance (log scale) in four RNA-seq libraries from the GTEX dataset of 427 human lung transcriptomes. The long left shoulder reflects a mixture of low-level transcription, transcriptional noise, and technical artifacts. The vertical bar on the left shows the frequency of zeros (undetected) genes. B) The proportion of all genes detected among the 427 GTEX libraries that were also detected in each of the four libraries shown in panel A. The proportion increases with increasing expression until genes with average TPM greater than 1 are virtually guaranteed to be detected in all libraries. The proportion was averaged over each bin of 0.05 log TPM width. Bins that contained fewer than 20 genes were not plotted. C) The gene-wise variances for genes that were detected in each of the 427 libraries, plotted as a function of their mean expression levels. The green trend line shows the empirical mode variance (most frequent variance) at each bin of 0.05 log TPM among genes with greater than zero expression. As expression increases, the variation in expression between libraries decreases. D) Distribution of transcript abundance for three representative genes across the 427 libraries. Gene 1 with the lowest detected expression shows the highest number of libraries in which it is not detected. Gene 3, with the highest expression, shows the lowest variance in expression across libraries.

approximately Normal distributions on the log scale at high expression levels, while genes with low expression have higher variance and zero TPM in some replicates (Fig. 1 D).

We present a hierarchical Bayesian model that describes the processes that generate transcriptomic data—by accommodating the relevant features described above—with the goal of inferring the expression state of each gene; that is, whether that gene is actively expressed in the cell or tissue. We then describe a numerical procedure, Markov chain Monte Carlo, to estimate the parameters of the hierarchical Bayesian model from transcriptomic data. In this way we are able to estimate the marginal posterior probability that each gene is actively expressed in the tissue-specific transcriptome. Finally, we describe an approach for assessing the fit of the data to the hierarchical Bayesian model. Each stage of model development, model testing, and data analysis is described in greater detail in the supplemental text (section S1). We implemented our method in an R package we call `zigzag`.

Model.—Our goal is to infer the expression state of each gene from the observed data, $\mathbf{X}$, (Fig. 1 D) under a hierarchical Bayesian model with two levels. The upper level of the model describes

the distribution of the true (unobserved) expression level of each gene, $\mathbf{Y}$ (Fig. 2 A); the lower level describes the variation in the observed expression levels, $\mathbf{X}$, as a consequence of biological and technical factors (Fig. 2 B-D). The focal parameter of our model, z^a, indicates the expression state of all genes; when $z^a_g = 1$, the expression state of gene g is active, when $z^a_g = 0$, the expression state is inactive. We infer the joint posterior probability distribution of the hierarchical model parameters—including z^a—given the observed expression-level data, $\mathbf{X}$, by applying Bayes' theorem:

$$\overbrace{P(z^a, \theta, \mathbf{Y} \mid \mathbf{X})}^{\text{posterior distribution}} = \frac{\overbrace{P(\mathbf{X} \mid \mathbf{Y}, \theta_1)P(\theta_1)}^{\text{lower level}}\overbrace{P(\mathbf{Y} \mid z^a, \theta_2)P(z^a, \theta_2)}^{\text{upper level}}}{\underbrace{P(\mathbf{X})}_{\text{marginal likelihood}}}, \tag{1}$$

where the first term in the numerator is the joint probability of the lower level of the hierarchical model given the local model parameters, θ_1, the second term is the joint probability of the upper level given the local model parameters, θ_2, and the denominator is the average probability of the expression data under the model (*i.e.*, the marginal likelihood). A graph of the hierarchical model is shown in the supporting text (Figure S1).

The lower level of our hierarchical Bayesian model describes the probability of the observed expression levels for all genes across all libraries, $\mathbf{X}$, assuming that the true expression levels of all genes, $\mathbf{Y}$, and the local model parameters, θ_1, are known: $P(\mathbf{X} \mid \mathbf{Y}, \theta_1)$. We model the observed expression level for a given gene across libraries (Fig. 1 D) as resulting from two sampling processes. The first process describes whether a given transcript is detected in a given library, which models the impact of the true expression level of a gene, its length, and the sequencing depth on detection probability (Fig. 2 B). The second process describes the variation in the observed expression levels for a given gene (which models the inverse relationship between the true expression level and the variation in the observed expression level across libraries) (Fig. 2 C). The lower level is described in more detail in the supplemental text (Section S1.2.1).

The upper level of our hierarchical Bayesian model describes the distribution of the true, unobserved expression levels, $\mathbf{Y}$. We assume that the true expression levels of genes can be divided into two populations; those genes that are actively expressed (Fig. 2 A, red), and those that are not actively expressed (Fig. 2 A blue). The assignment of each gene to the active or inactive expression-state populations is described by parameter z^a_g, where $z^a_g = 1$ indicates that the gene is assigned to the active population, and $z^a_g = 0$ indicates that the gene is assigned to the inactive (background noise) population (Fig. 2 A). We further assume that the population of inactive genes can be subdivided into two subpopulations: one with zero expression, and another with non-zero expression (*e.g.*, as might result from promoter leakage). Similarly, within the active population, genes may cluster into one or more subpopulations with distinctly different expression levels (*e.g.*, housekeeping genes may have higher true expression levels relative to other genes). The upper level is described in the supplemental text (Section S1.2.2).

A challenge for inference using mixture models is the non-identifiability of mixture component means and variances because the likelihood is invariant to index switching between components. This can be overcome with an identifiability constraint such as ordered offsets in the prior distribution for component means where each mean is assumed greater than the previous and the prior distribution is defined over the difference between means. This approach, however, can lead to very slow MCMC burn-in times for some data and starting parameter values. We devised an alternative approach where thresholds can be specified where each mean is assumed to be greater than a component-specific fixed offset; for example, the first mean is assumed greater than 1 and the second greater than 4 (Figure S2). We find that this yields similar estimates to the first approach but reaches the stationary distribution much sooner.

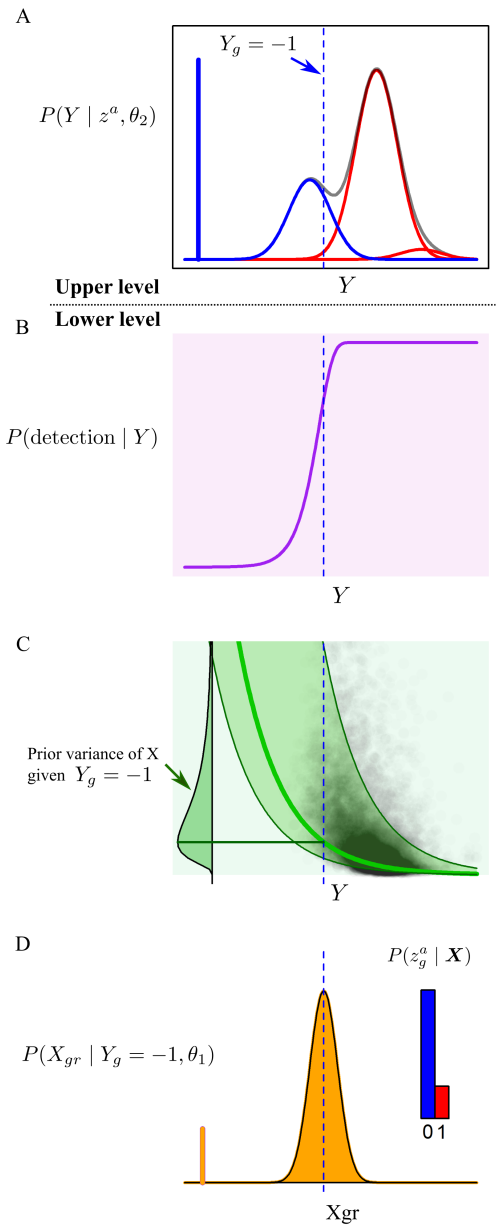

Figure 2: The two-level Hierarchical Bayesian model for determining the active/inactive expression states of genes. A) The bimodal distribution of true expression levels Y . Blue is the spike-and-slab distribution of background inactive expression (i.e. promoter leakage). Red are the distributions of expression levels among genes that are functionally active. z^a indicates whether Y is inactive (blue) or active (red). Y_g, for example, is an inactive gene with (log) expression = -1 (blue dashed lines) which determines subsequent sampling processes. B) The probability of detecting a gene in a library depends on Y. C) The most likely (mode) variance of gene expression is inversely related to Y (thick green trend-line). The prior distribution of variance for gene g is determined by Y_g and by the genome-wide variation around the trend line (flanking green area). For example, the prior distribution for variance given $Y_g = -1$ is shown by the rotated distribution (left). The mode of this distribution is located where the dashed blue line ($Y_g = -1$) intersects the trend line of mode variance. D) The likelihood of a set of observed expression levels for gene g in library r, X_{gr}, given Yg=-1 (orange). The likelihood of not observing any expression (see panel B) is seen as the bar on the left. The right inset shows the marginal posterior probability that g is actively expressed given the data, $\boldsymbol{X}$. For gene g with $Y_g = -1$, $P(z_g^a = 1|\boldsymbol{X}) = 0.2$ (red bar). θ_1 and θ_2 denote the lower and upper level parameters of the hierarchical model, respectively.

Parameter Estimation.—The posterior distribution, $P(\boldsymbol{z}^a, \theta, \mathbf{Y} \mid \mathbf{X})$, cannot be calculated analytically because the marginal likelihood, $P(\mathbf{X})$, is impossible to evaluate in practice. Accordingly, we use a numerical algorithm—Markov chain Monte Carlo (MCMC; [43–46]—to approximate the posterior distribution, (see supplemental text, section S1.3). The MCMC algorithm samples parameter values in proportion to their posterior probability. From these MCMC samples, we compute the posterior probability that a given gene is active as the fraction of MCMC samples where $z_a^g = 1$ (Fig. 2 D, inset). We validated our MCMC implementation by running it under the prior (no data), and by measuring coverage probabilities using simulated data (Figures S3-4), see supplemental text (section S1.4).

Model Checking.—Model-based inference is based on the premise that our model provides an adequate description of the process that gave rise to our observations. The Bayesian approach for assessing model adequacy is called *posterior-predictive assessment* [47]. This approach is based on the following premise: if the inference model provides an adequate description of the process that gave rise to our observed data, then we should be able to use that model to simulate datasets that resemble our original data. The match between the distributions of the observed data and the data simulated from the inference model can be quantified using a summary statistic. By repeatedly simulating datasets from the joint posterior distribution of the model parameters, we construct a predictive distribution of summary statistics that characterize the resemblance between the simulated and observed data. A posterior predictive P-value is then computed by measuring the tail-area beyond zero of the predictive distribution.

We measure model adequacy using three summary statistics: (1) the lower level Wasserstein statistic, which measures the discrepancy between the expression levels for detected genes, $\mathbf{X}$, and the expected expression levels for detected genes given $\mathbf{Y}$ and the lower level model parameters θ_1, $E[\mathbf{X} \mid \mathbf{Y}, \theta_1]$; (2) the upper level Wasserstein statistic, which measures the discrepancy between the true expression levels for detected genes, $\mathbf{Y}$, and the expected true expression levels for detected genes given the upper level mixture model, $E[\mathbf{Y} \mid \theta_2]$ (Figure S5), and; (3) the Rumsfeld statistic, which measures the average discrepancy between the fraction of undetected transcripts in each library and the expected fraction of undetected transcripts in each library given the model. We consider a model inadequate if the 95% predictive intervals of these summary statistics exclude zero, see supplemental text (section S1.5).

RESULTS

Replicates improve the ability to distinguish active expression from noise

First, we assessed the power to discriminate between low levels of active expression and transcriptional noise under the Bayesian mixture model. The ability to discern the activity state of genes will inevitably be sensitive to the degree of overlap between the active and inactive distribution components. We simulated gene-expression data under our model and assessed the impact of this overlap on the ability to predict expression state. As expected, the average probability of correct classification declines as the overlap between active and inactive distributions increases. However, for any given overlap, increasing the number of replicate libraries results in a moderate improvement in classification ability (Figure S6 A). These results indicate that increasing the number of replicates improves the power to distinguish between low levels of active expression and transcriptional noise, but the level of overlap between these expression states sets an upper limit on this power. Details for all analyses are provided in the supplemental text (section S2).

The method is robust to model misspecification

Gene-expression distributions tend to have some highly expressed genes that make the right tail heavier than predicted by a normal distribution. In such datasets, the active component distribution will consist of an unknown number of subcomponents. To determine how this aspect of model uncertainty—and the possible failure to include an adequate number of active subcomponents—impacts our ability to discriminate between active and inactive genes, we used simulated data and assessed our ability to accurately classifying gene-expression states when the number of active subcomponents was either under- or over-specified relative to the true (generating) model. We found that under most scenarios when the active subcomponents largely overlap, estimated expression-state probabilities based on mis-specified models are very close to those inferred under the true model. Estimates using overspecified models (*i.e.*, with one more than the true number of active subcomponents) are virtually identical to those under the true model (Supplemental Fig. S6 B, column 1). Similarly, estimates inferred using underspecified models (*i.e.*, with one fewer than the true number of active subcomponents) are also virtually identical to those under the true model when the subcomponents significantly overlap (Figure S6 B, columns 2 and 3), but perform worse when the separation between active subcomponents is high (column 4). These results indicate that when multiple sub-component distributions significantly overlap and have similar variances, a single distribution is a robust approximation. Moreover, there is little to no cost to over-specifying the model.

Estimates of gene-expression states are consistent with chromatin marks

To evaluate model performance with real data, and to gain insights into qualitative gene expression states in a complex tissue, we analyzed human lung trasncriptome data from the GTEx RNA-seq dataset [30, 48]. We compared mixture models with one, two, three, and four active subcomponents (Figure S7), and fit each model to 25 human lung libraries randomly selected from the GTEx RNA-seq dataset to derive posterior distributions of gene-expression levels for all protein-coding genes. Posterior predictive simulation indicated that that a model with three distinct active subcomponents is the best approximates this dataset (Figure S7). Sensitivity analysis shows the posterior distribution of z_g^a is relatively insensitive to the number of specified components (2 or greater) in the active part of the mixture model (Figure S8). We found that inactive genes tend to have average expression below 1 TPM, while active genes had expression greater than 3 TPM (Fig. 3 A). Although 98% of protein-coding genes have detected expression in at least one of the 25 libraries, this analysis shows that only approximately 67% of protein-coding genes are actively expressed with $P(z_g^a = 1|\mathbf{X}) > 0.5$. We estimate from mean posterior estimates of model parameters, and averaging over gene lengths, that approximately 0.9% of the protein-coding transcriptome emanates from background noise. In other words, 99.1% of protein-coding transcripts in these RNA-seq samples come from actively expressed genes.

Previous research suggests that epigenetic marks associated with inactive genes correlate with expression levels in the transcriptional noise range, while chromatin marks associated with active genes correlate with higher expression levels [18, 35, 49]. To further characterize this pattern, we used an extensive epigenomic dataset for human lung tissue samples from the Roadmap Epigenomics Consortium [48, 50, 51]. In this dataset, information about histone modifications, DNA methylation, and DNA accessibility was used to annotate the regulatory state at each genomic location and identify 15 chromatin states relating to promoter, enhancer, and transcriptional activity state for each non-overlapping 200 bp window. Genes that contain each of these marks in at least one exon show expression levels that correspond with one of the two subcomponents of the mixture distribution (Figure S9). We conservatively defined active genes as those that have at least one active promoter or promoter-flanking mark overlapping with its exon sequences (states 1 and 2), at least one active transcription mark (states $3-5$), and no marks associated with repressed expression [48];. Inactive genes were defined as having no active promoter or transcription marks (states

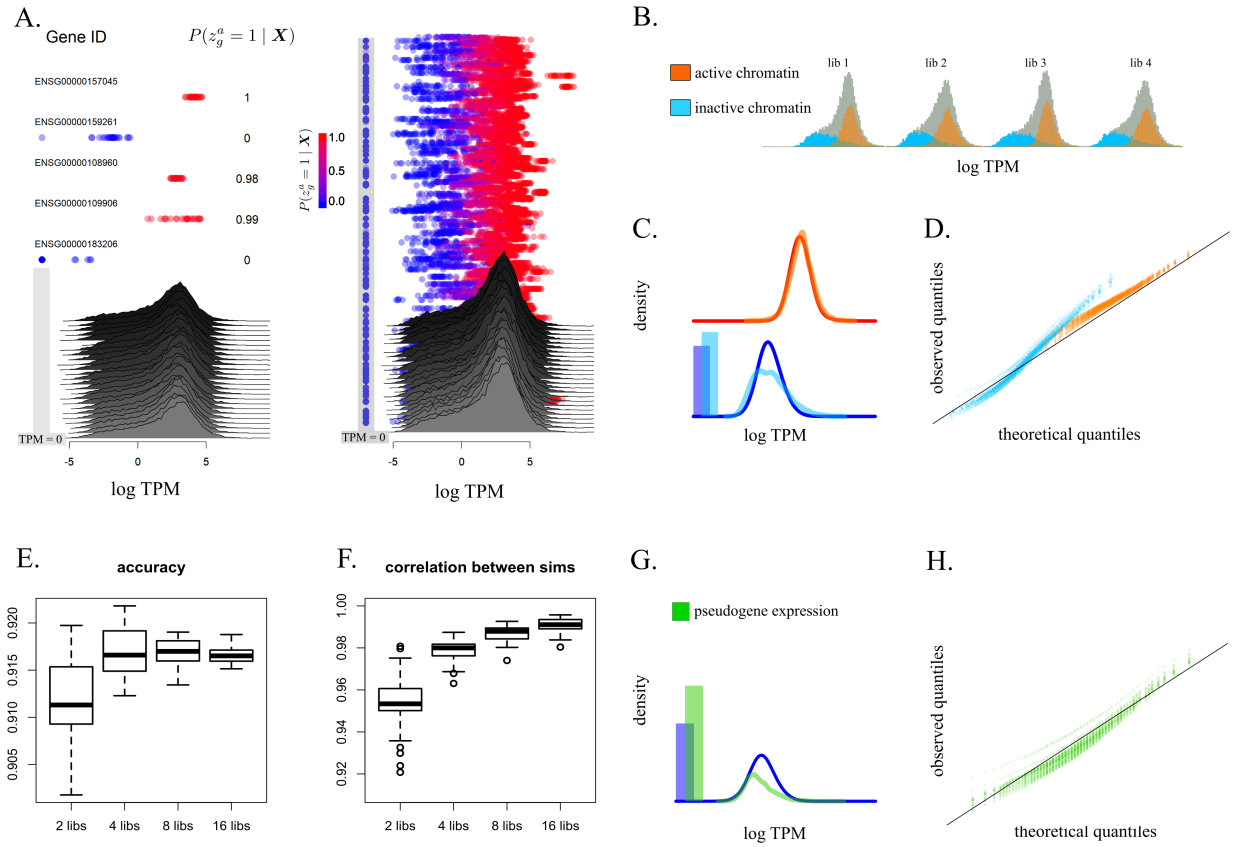

Figure 3: Model fit predicts gene expression states consistent with independent data. The distribution of log mean TPM in 25 human lung RNA-seq libraries (grey). Each row of colored dots gives the expression levels of one gene across 25 libraries; heat of the color indicates the probability of active expression, P(A—X). A, right) Results for 10,000 genes show a genome-wide transition from background noise to active expression between 1 and 3 TPM. B) The distribution of observed expression for genes with active promoter and transcription marks (orange histograms), genes that lack any active promoter or transcription marks (aqua histograms) and all protein-coding genes (gray) in four libraries. C) The orange and aqua distributions from panel B overlaid over the distributions of actively expressed (orange and red, top) or inactive (aqua and blue, bottom) genes predicted by the posterior fit of the model; note the high degree of overlap. D) Quantile-quantile (qq) plots offer another way of visualizing the distribution overlap shown in C. If the distributions are identical, each quantile for one distribution will equal the same quantile in the other distribution, and thus each quantile will fall on the y=x line. The qq-plots for all 25 libraries ranging from the bottom first percentile up through the 99th percentile, by increments of 1 percent, lie close to the y = x line, suggesting high similarity between the distributions inferred from the model and the distributions implied by chromatin marks. E) Classification accuracy (consistency between estimated probability of expression state and the classification of state based on chromatin marks for transcription) increases with the number of replicate libraries. Box and whiskers show medians and standard errors for each of the four experiments (n = 10 for each number of libraries). F) For the same dataset, correlations between the estimated posterior probabilities of expression states increase with the number of libraries (2 16) used to infer these probabilities. G) Empirical distribution of pseudogene expression for a single library (green) overlaps the inferred expression distribution of inactive genes (blue). H) qq-plots of pseudogene expression across all 25 libraries similarly show a high degree of overlap with the distribution of inactive genes.

$1-5$) and having at least one heterochromatin, bivalent enhancer/TSS, Polycomb-repressed, or quiescent marks (states $9-15$) overlapping with at least one exon (see supplemental text section S2.3.1). Of 19,154 protein-coding genes in human lungs, 62% could be classified in this way into either active (7,261 genes) or inactive (4,707 genes). Expression levels of active and inactive genes defined in this way follow well-separated unimodal distributions, which correspond to the upper and lower peaks, respectively, of the bimodal distribution that describes overall transcript abundance in this tissue (Fig. 3 B). The distribution of genes with active chromatin marks overlaps well with the active component of our inference model, while the expression distribution of genes with inactive chromatin marks overlaps well with the inactive component, as shown by quantile-quantile plots (Fig. 3 C–D).

If our model is accurate, genes with high inferred probability of active expression should be classified as active based on chromatin data, while transcripts inferred to be transcriptional noise will come from genes associated with repressed chromatin. We also expect that the accuracy of our method will increase as the number of replicate libraries increases. To test this, we randomly sampled sets of 2, 4, 8 and 16 libraries from > 400 human lung libraries in the GTEx dataset. For each number of replicates ($2-16$), we sampled 10 independent sets of libraries (*e.g.*, 10 sets of 2 libraries, 10 sets of 4 libraries, etc.). We estimated the probability of active expression for each protein-coding gene in each dataset as described above and computed the average probability of accurately predicting the chromatin state of genes averaged over all classified genes (11,974 genes). We found that the average accuracy of our estimates was high (above 90%) for all numbers of replicates. There is a large improvement in accuracy when expression probabilities are inferred from more than two replicate libraries, with marginal and diminishing improvements as the number of replicates increases above four (Fig. 3 E). When we compare independent sets of replicate libraries, we find that between-set correlations increase from 0.94 to almost 1 as the number of replicates per set increases from 2 to 16 (Fig. 3 F). These results indicate that our method produces estimates of active gene expression that are highly consistent with chromatin states, and that relatively few libraries are sufficient to compute an accurate and consistent probability of active expression for each gene.

We also tested how well our model predicted the expression states of pseudogenes. Because these genes are assumed to be non-functional, and presumably expression of these genes is not maintained by selection, we expect their expression distribution to resemble the distribution of functional protein-coding genes that are not actively expressed in human lungs. Indeed, we observe a strong overlap between the empirical distribution of pseudogenes and the inactive distribution component inferred by our model (Fig. 3 G, H), indicating that the model provides an adequate description of transcriptional noise.

Accurate detection of rare genes with known biological functions in complex tissues

Many tissues are highly heterogeneous, containing very rare as well as abundant cell types. When the tissue is analyzed as a whole, genes expressed at low to moderate levels in rare cell types may have expression levels that are barely above the tissue-wide transcriptional noise. To test the ability of our method to detect genes expressed in rare cell types in a complex tissue, we focused on the *Drosophila* testis. The bulk of the testis consists of differentiating gametes at different stages of development. By contrast, germline stem cells and several types of somatic cells that together make up the stem-cell niche at the tip of the testis are each present at $20-30$ cells per testis. Genes that are known, based on developmental-genetic data, to be expressed exclusively in the niche provide a suitable test case for our method.

We created RNA-seq libraries from pooled samples of adult *D. melanogaster* testes from two different strains with two biological replicates per strain, performed Illumina sequencing, and estimated expression in TPM units (see Methods). We jointly analyzed estimates from all four replicates with our hierarchical mixture model to compute genewise posterior probabilities of active expression. In addition to the typical two modes separated around 1 TPM, the testes show a

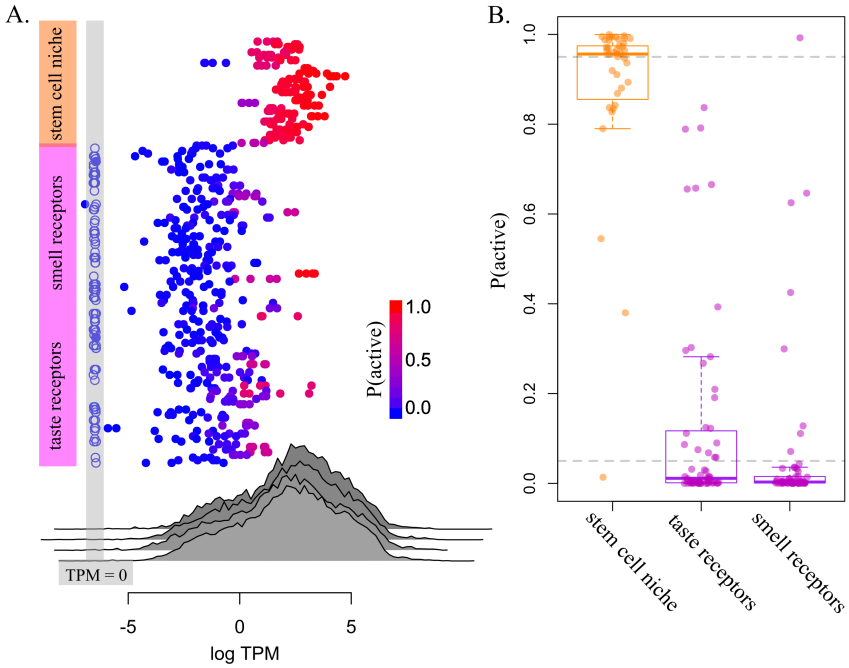

Figure 4: Model fit to the Drosophila testes transcriptome is consistent with developmental-genetic data. A) The distribution of gustatory and olfactory receptor gene expression levels (bottom 120 genes, purple) and genes with characterized expression patterns in the testis stem cell niche (top 39 genes, orange) in 4 replicate libraries, relative to all gene expression levels (gray distributions). Heat of the color indicates the probability of active expression in each library. B) Boxplots show the posterior probabilities of active expression for each class of genes given the expression data in the four libraries, with the dashed lines indicating 0.05 and 0.95 posterior probabilities.

significant subpopulation of transcripts with abundances around 150 TPM, which could represent sperm-specific proteins (Figure 4A). We found that a mixture model with three active subcomponents was sufficiently complex to account for the data (Fig. S10). Sensitivity analysis also supports three components in the active distribution (Figure S11). Like the human lungs, the threshold between inactive and active expression in the testis appears to be around $1 - 3$ TPM, and 71% of protein-coding genes have > 0.5 probability of active expression.

To further test the sensitivity and accuracy of our method, we assembled a list of 39 genes for which developmental-genetic studies provide strong evidence that their expression is restricted to a small group of cells in the stem-cell niche, and that they are functionally important in those cells (see supplemental table S4). We also compiled a set of 121 genes that encode odorant and gustatory receptors or phototransduction components and are therefore unlikely to be expressed in the testis. These genes, however, do not provide a true negative control, since there is no independent data showing that they are not expressed in the testis. The expression probabilities calculated using our method are highly consistent with prior expectations: odorant and gustatory receptors align mainly with the inferred inactive component of the transcriptome, while stem-cell niche genes align with the active component (Fig. 4 A). Among the positive-control genes, the median posterior probability of being active is 0.96, with 37 out of 39 genes having > 0.5 probability of active expression. The median posterior probability of expression among the negative-control genes is 0.005, with 111 out of 120 genes having probability below 0.5 (Fig. 4 B). We conclude that our method is able to identify genes expressed in rare cell types within complex tissues.

Probability estimates classify genes correctly at levels close to theoretical maximum

Compared to the commonly used methods based on hard TPM thresholds (*e.g.*, 0.3 RPKM, [31]; 1 RPKM, [28]; or 3 TPM, [36], our probabilistic method is more objective, accounts for uncertainty

in the data, and is better able to take advantage of replication. We wanted to know whether our method yields improved discrimination between inactive and active genes. To determine the maximum discrimination that can be achieved by threshold-based methods, we eschewed arbitrary numerical thresholds, and instead used the chromatin data to find a library-specific threshold for each library that maximally separated the genes with active and inactive chromatin marks. Genes with log TPM higher than threshold were called active, and those below the threshold inactive. Using the human lung dataset, we calculated the proportion of genes classified as active or inactive based on their chromatin states that were also congruently classified based on each possible TPM threshold. For each library, we progressively changed that threshold from -5 to 5 log TPM in increments of 0.5, and computed classification scores to determine which threshold maximized the congruent classification score, *i.e.*, achieved a score closest to 1 (maximum of curve in Fig. 5 A). In parallel, we calculated expression probabilities across all libraries in the same dataset using our model-based method, classified genes with > 0.5 posterior probability of expression as active and those < 0.5 as inactive, and tested for congruence with the chromatin data using classification scores (horizontal red line in Fig. 5 A). We found that probabilistic classification of genes into active or inactive using our method achieves congruence with the chromatin data that is virtually identical to the maximum possible score that can be achieved with a perfectly set threshold (Fig. 5 A).

Similarly, in the *Drosophila* testis dataset, we found that our method distinguishes between the positive- and negative-control genes with accuracy that is close to the maximum level that is theoretically possible using hard TPM thresholds (Fig. 5 B). Thus, the probabilities of active expression computed using our method based on the patterns of transcript abundance in replicate libraries are maximally congruent both with chromatin states and with information from a large body of experimental work. In this respect, this method performs as well as any analysis based on a hard TPM threshold, and does so without any outside information of gene chromatin state or histochemcial data. We note that in the absence of sample-specific epigenomic or histochemical data, a hard TPM threshold needs to be chosen arbitrarily and cannot be optimized, whereas our probabilistic method has no such disadvantage.

Qualitative changes in gene expression during brain evolution in humans

Our results indicate that the mixture model method is sensitive, able to detect genes expressed in small subpopulations of cells, and yields classification accuracy that is close to theoretical maximum. We next applied it to identify qualitative changes in gene expression in the course of evolution. Comparative gene expression studies indicate that quantitative expression changes are widespread [13, 52]. Recently, a study of several brain regions in human, chimpanzee, and rhesus macaque estimated that 26% of transcripts have significantly diverged in expression in the human lineage [53]. To test whether this divergence includes qualitative changes (inactivated expression, or *de novo* activated expression), we re-analyzed the data from six brain regions: amygdala (AMY), ventral frontal cortex (VFC), dorsal frontal cortex (DFC), superior temporal cortex (STC), striatum (STR), and the area 1 visual cortex (V1C). After posterior-predictive analysis of two alternative models and sensitivity analysis (See Supplemental Text, Figures S12-15) we computed the posterior probability of active expression in each brain region in each of the three species for $17,000 - 18,000$ protein-coding genes using our method. We then compared the probabilities of active expression for approximately 12,000 genes that are conserved as 1:1:1 orthologs. We find that a small proportion of genes show evidence of qualitative expression changes. Ignoring relatedness, we make a rough calculation of the probability of a human-specific gain or loss of expression by multiplying each genes expression probabilities across species. For example, we multiply the probability that the gene is active in human by the probability that its ortholog is not active in both chimpanzee and macaque.

At the 0.95 probability threshold, $9 - 20$ genes gained and $16 - 23$ genes lost expression in each of the six brain regions in the human lineage (Fig. 6 A). We find that most changes in expres-

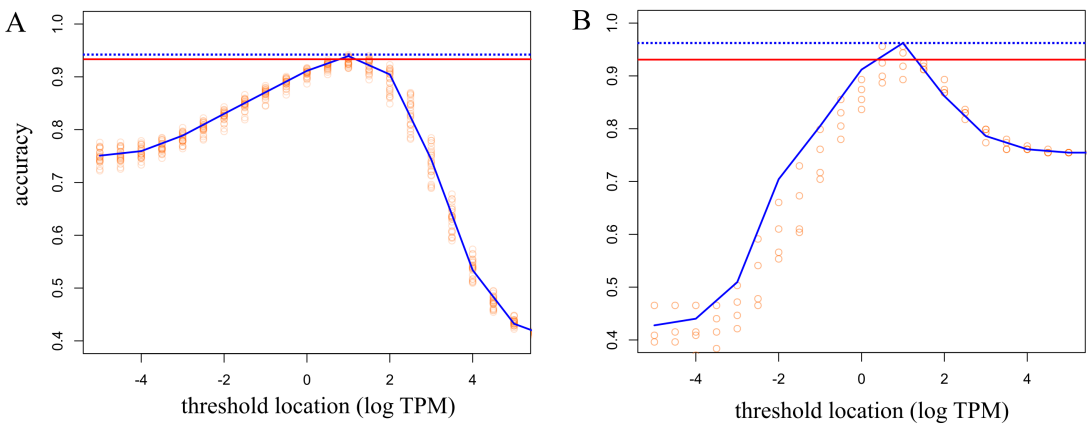

Figure 5: Model-based assignments of gene expression states approach the maximum possible accuracy. Classification accuracy was determined by testing whether the active/inactive assignment based on the hierarchical Bayesian model (red) or arbitrary log TPM thresholds (blue and orange) matched the predictions from chromatin states (A) and developmental-genetic studies (B). A) Classification accuracy in 25 human lung transcriptomes. Horizontal red line indicates the probability of correct classification of genes into active vs inactive based on posterior probabilities estimated from our statistical model. This probability is compared to the active/inactive classification based on arbitrary thresholds ranging from -5 to 5 log TPM on the X axis. Each orange dot shows the accuracy of threshold-based classification for one library at the given threshold. The blue curve shows the accuracy score from the median expression across all libraries at that threshold; the maximum of this curve is indicated by the horizontal blue line. The model predictions approach the maximum possible accuracy, which is near 2 TPM for this dataset. B) In the Drosophila testis transcriptome, model-based classification reaches the maximum possible accuracy compared to active/ inactive assignments based on arbitrary TPM thresholds. In this tissue, the best threshold-based classification would be achieved around 5 TPM.

sion states are not restricted to a single brain region. However, human expression gains are more likely to be strictly region-specific than human expression losses (Fig. 6 A). The greatest number of qualitative expression changes have occurred in the striatum. In summary, while many genes show quantitative expression divergence, a much smaller fraction show evidence for qualitative changes in expression state. Future investigation could determine if these qualitative changes on average have more phenotypic impact than the average impact of the thousands of genes previously detected to have differential expression.

Genes that gained expression in the human brain make intriguing candidates for the genetic changes behind human evolution. For example, the *Slc17a6* gene, which encodes a glutamate transporter, is predicted with high probability under our model to be expressed in the human striatum and not expressed in Macaque or Chimpanzee striatum(Fig. 6 B). Expression data indicate this gene is also not expressed in the mouse striatum [54] suggesting expression in the human striatum resulted from a novel activation in the human lineage. This gene, along with its two paralogs, is commonly used as a marker for excitatory neurons, as the protein it encodes, VGLUT2, is responsible for loading glutamate, the principal excitatory neurotransmitter, into synaptic vesicles [54–56]. Variation in these transporters can impact the concentrations of neurotransmitters contained in synaptic vesicles, and thus the quantal size of neuroexocytosis, which in turn influences the degree of response in the post-synaptic neuron [54]. The expression of *Slc17a6* in the human striatum appears to be novel, with interesting implications. The striatum is classically known to be involved in motor planning, where it receives afferents from other midbrain structures and the prefrontal cortex, and projects axons to the thalamus as well as the prefrontal cortex and other cortical areas. Recent data indicate that the striatum may also be involved in high-order cognitive processes [57]. *Slc17a6* expression is conserved in all the other brain regions studied. The possible evolutionary gain of this glutamate transporter specifically in the human striatum may indicate a substantial change in the function of this brain region either through the gain of a novel cell type or change in the activity of an ancestral cell type. Our method opens the way for testing whether genes that undergo qualitative expression changes have a particularly strong impact on

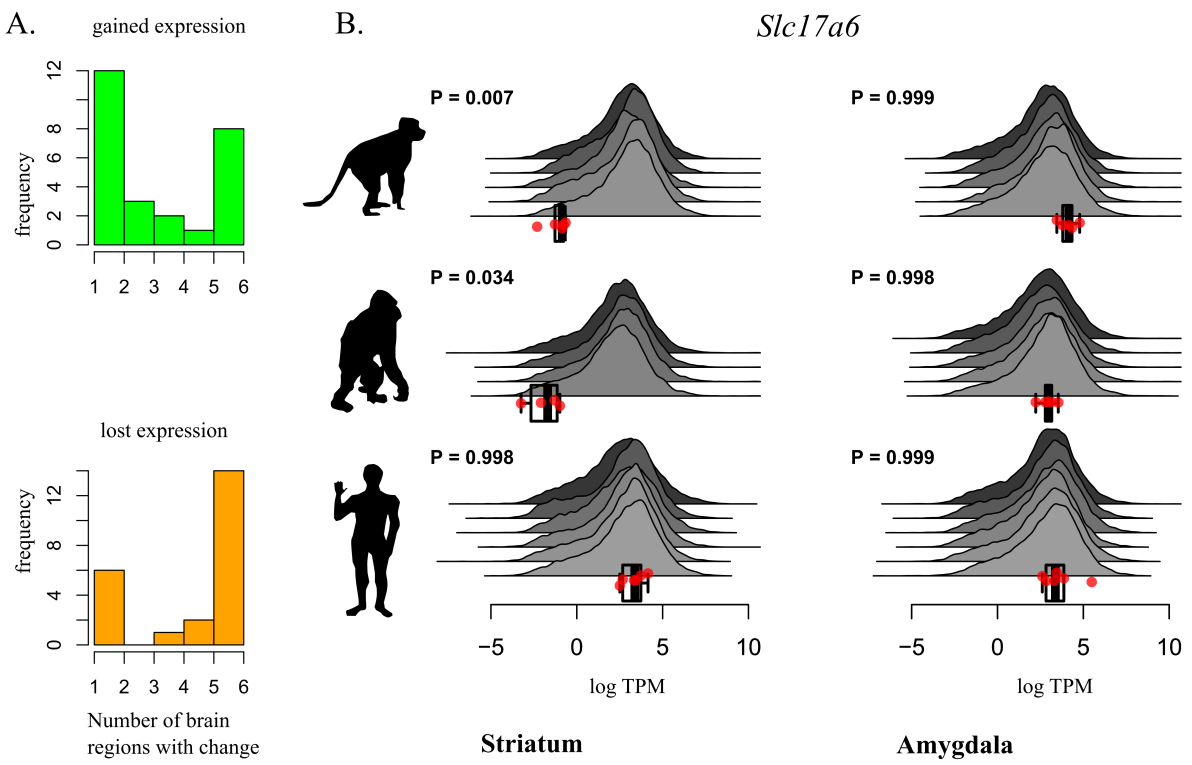

Figure 6: Estimated gains and losses of gene expression during human brain evolution. A) Human-specific gains and losses of gene expression across six brain regions. In green, each bar shows the number of genes with > 0.95 probability of being expressed in humans, and < 0.05 probability of being expressed in chimpanzees and macaques, in 1 to 6 brain regions. Orange bars show the number of genes with > 0.95 probability of being expressed in chimpanzees and macaques, and < 0.05 probability of being expressed in humans. A gain or loss in one brain region indicates tissue-specific evolution, a gain/loss in all 6 regions indicates a likely brain-wide change. B) Slc17a6 is expressed with high probability in the human striatium ($P > 0.99$), but with low probability in both chimpanzee and macaque striatum ($P < 0.04$). This gene is actively expressed in the amygdala in all three species. The distributions of log TPM for all protein-coding genes in 4 6 replicate libraries per species are shown in grey. The red dots show the measured log TPM for *Slc17a6* in each replicate library.

phenotypic evolution.

DISCUSSION

The goal of this method is to answer a seemingly simple but surprisingly difficult question: which genes are expressed, and which genes are not expressed, in a tissue or cell type? There is abundant evidence suggesting gene expression is both a quantitative trait and a binary qualitative trait. Models of transcription kinetics distinguish between two discrete promoter states: active and inactive. Importantly, these models often consider transcription as occurring in both of these states, but with large differences in rates of transcription between them [20, 21, 26, 27]. This implies that genes which are regulated to be inactive in a tissue or cell type will tend to produce expression levels much lower than genes regulated to be active in that tissue type, resulting in a bimodal distribution for the tissue transcriptome as a whole. This prediction is borne out empirically: two major classes of expression levels are broadly observed across species at the level of transcriptomes, whether in transcriptomes sampled from complex tissues or in transcriptomes derived from homogenous bulk samples of a single cell type [18, 35].

Substantial evidence indicates that the widespread low level of transcription observed throughout genomes is not functional [17–21, 24, 26, 31]. Epigenomic investigations, GO analysis and studies using "ground-truth" gene sets with known expression states indicate that the low and

high modes of the observed expression distribution are associated with inactive and active expression, respectively [18, 35, 36]. Expression leakage of inactive genes may be influenced by the chromatin dynamics of their genomic neighborhood [19, 58–63]. These biological processes, along with technical errors such as read mis-mapping, underscore the importance of distinguishing transcriptional noise from functionally relevant expression when making inferences about patterns of gene expression within and between RNA-seq datasets.

Our hierarchical Bayesian mixture model integrates relative abundance measurements across any number of replicate libraries, and leverages information from all genes in the dataset, to infer attributes of the inactive and active expression distributions and estimate the posterior probability that each gene is actively expressed. This is a significant improvement over current methods that use *ad hoc* library-specific thresholds or fit mixture models to individual libraries [18, 34, 38]. The variation in expression of genes across replicate libraries means it is plausible that, in some libraries, active genes go undetected or fall within the inactive part of the transcriptome distribution, while inactive genes are nevertheless detected in some replicates. Because single-library methods treat the expression state of a gene as a library-specific attribute, the expression states of genes as a tissue or species property cannot be directly inferred. By contrast, our hierarchical model treats the expression level and expression state of each gene as population-level parameters that are inferred from an independent sample of replicate RNA-seq libraries. We can therefore jointly infer the shape of the transcriptome distribution and the expression state of each gene within the transcriptome distribution given the patterns of detection and variation among all genes in all replicate libraries.

Sharing information between genes.—Given the true expression of any given gene, accurately characterizing the variance of that gene is essential to inference about its expression state. This is a challenge that is compounded by the fact that RNA-seq datasets often contain few replicate libraries. Method developers have addressed this challenge by leveraging the clear and consistent relationship between the mean and variance of gene expression to shrink gene-wise variances toward the trend inferred from all genes in the dataset [64–66]. Our hierarchical model of gene-specific variance jointly estimates mean-variance trends and the amount of variation around that trend with all gene variances. For a given gene-expression level, if all genes show similar levels of variation across libraries, then the trend strongly predicts gene-specific variance given gene-specific expression level (Figure 2C, green shaded region). In other words, if there is a strong relationship between the expression level and the amount of variation, the prior distribution for gene-specific variance will be concentrated around the trend line and thus the expression level of a gene will be highly informative about its variation between libraries. As the number of libraries in the dataset increases, the posterior distribution of the variance for each gene becomes relatively free from the influence of the trend line. Our parameterization therefore jointly estimates the predictive value of the trend along with the individual gene-wise variances to provide robust classification estimates for even small datasets.

Differential expression methods employ library-specific scaling factors in their tests [64–66]. In our model, this translates to the parameterization of the mean-variance trend-line where the component of the variance trend that is independent of the expression level of genes is influenced by random effects such as scaling factors. If the scale of library expression levels is highly variable, then the trend line will be shifted upward. This should have similar impacts on estimates of expression state for genes but with a single parameter rather than estimating scaling factors for each library in the dataset.

Inference from patterns of detection across libraries.—Another common challenge in RNA-seq datasets involves making inferences about genes that are only detected in a subset of libraries. The probability of observing zero expression for a given gene in RNA-seq data depends on many factors, including its relative expression level, transcript length, variability in sequencing depth, read qual-

ity among libraries, and differences between the genetic content of annotations and the biological sample (*e.g.*, gene-copy number variation between populations, or annotation errors; [37, 67]). Our approach is to model zero-expression calls as being caused by two processes: either the true expression is zero and no amount of sequencing effort will detect the gene, or the true expression is greater than zero but the gene goes undetected due to binomial sampling error where the detection probability for a gene is a function of its expression level, its length, and a library-specific scaling factor. Because we explicitly model the generation of zero-expression calls in RNA-seq data, our approach is more robust to variability in sequencing depth across libraries and to discrepancies between the sample and the reference annotation. While many studies filter out zero-expressed or low-expressed genes before comparing gene expression between samples, our approach uses all the data to make biologically meaningful inferences, especially at the low end of the expression distribution. We have demonstrated through posterior-predictive checking that our performs well in predicting patterns of zero-expression calls observed in data.

Parameterizing the mixture model of the transcriptome.—The upper level of the model describes the distribution of relative expression of the set of genes that comprise the transcriptome, which we model as an overlapping mixture distribution of inactive and active genes. While the active expression component of the mixture distribution can in principle consist of a single peak, datasets often include clusters of very highly expressed genes in the right tail of the distribution. Several studies have shown that a log-Normal distribution provides a good fit to the bulk of the empirical distribution of transcript abundance around the mode, but that the tails are heavier than expected [39, 40]. We improved the accuracy of classifying the expression state of genes by accommodating multiple active sub-components in our mixture model. We have shown that under-specifying the number of mixture components can impact the classification accuracy of genes at the margins, especially for genes with intermediate expression probabilities, *i.e.*, those that are inherently more difficult to classify. Thus, improving model fit produces more accurate classification. Posterior-predictive checks indicate this greatly improves the fit of the mixture distribution to the patterns of gene expression implied by the data.

Applications of qualitative gene expression analysis.—One of the key advantages of our method is that it allows biological insights to be drawn from tissue-specific transcriptomes without the need to compare different samples. The probability that a gene is expressed in a given tissue is an inherent property of that tissue, independent of other tissues. Once these probabilities are estimated, hitherto problematic comparisons become possible: for example, we can determine which genes are tissue-specific and which are expressed in multiple tissues. Importantly, expression probabilities are "portable" between studies and datasets. We have used our method to classify expression states of genes in diverse datasets generated by different labs using different protocols, in different species and tissues, and found that about $60 - 70\%$ of protein-coding genes are typically expressed in a given tissue. Differential analysis of qualitative expression states using our method will be a valuable way to augment tissue-specific transcriptome studies. In addition, for studies in which quantitative expression differences are the primary interest, this method can serve as an objective pre-filtering step to remove inactive genes from the comparison, for example by setting a conservative posterior probability of active expression as a cutoff. This allows the researcher to use a consistent prefiltering methodology before differential expression analysis.

Similarly, a comparison of expression probabilities between species can be used to identify the genes that have conserved expression states, and those that do not. These comparisons can be performed directly across multiple, independently generated datasets. We used this method to examine human brain evolution and found that several genes may have undergone a qualitative gain of expression (that is, went from inactive to active) in specific brain regions in the human lineage. There is a clear potential for more complicated studies that track gene expression states in multiple tissues and across many species to reconstruct the evolutionary expansion and contrac-

tion of gene expression domains and, by implication, the pleiotropic functions of these genes. By modeling the activity state of genes as a discrete trait, expression evolution can be studied using well developed phylogenetic methods.

The robustness of phenotype to quantitative changes in relative gene expression suggests that qualitative differences in expression state may be more likely to impact phenotype. Ribosome profiling and high-throughput protein sequencing have shown a modest relationship between RNA abundance and protein abundance, often with less than 0.5 correlation [14–16]. Evolutionary analyses indicate that relative transcript levels evolve more rapidly than protein levels. The emerging picture is that protein concentration is robust to changes in RNA concentration to varying degrees; therefore, quantitative differences in expression for some genes may not indicate a significant difference in protein abundance [1–9, 13]. Genes with qualitatively different expression states may be, on average, more likely to affect phenotypic differences because of the inherent limits of post-transcriptional compensation for such binary differences. Although it remains to be tested whether qualitative gene expression changes play especially prominent roles in the evolution of organ function, we emphasize that this question could not even be asked without an objective method for identifying gene expression states. Our method provides a practical way to address this question and is effective at generating candidate genes for follow-up functional studies

Conclusion.—For many purposes, qualitative comparisons of gene expression states between species and tissues will be more straightforward and interpretable than quantitative metrics and will provide a useful complement to quantitative measures of expression divergence. Posterior-predictive simulation suggests that our model fits diverse datasets and provide a means of measuring model fit for future improvements of this method. We anticipate that further developments will improve the realism and accuracy of this approach. As this approach to transcriptomics matures, we can explore the relative importance of qualitative and quantitative gene expression differences in cell differentiation and phenotypic evolution. Phylotranscriptomics is emerging as a powerful tool for investigating the genomic mechanisms of phenotypic change [52, 53, 70, 71]. We are optimistic that our method will provide a powerful addition to the comparative transcriptomics arsenal, opening the way for characterizing qualitative differences in gene expression between taxa and streamlining the identification of candidate genes that may be responsible for the evolution of organ structure and function.

METHODS

See supplemental Text 1 for a detailed description of the method and analyses.

ACKNOWLEDGEMENTS

The Authors thank Li Zhao and David Begun for technical advice and comments on the manuscript. We also thank Nerisa Riedl and Olga Barmina for technical assistance. This work was supported by NIH grant 5F32GM125107-02 to AT and NIH grant R35GM122592 to AK.

References

[1] Tamar Geiger, Juergen Cox, and Matthias Mann. Proteomic Changes Resulting from Gene Copy Number Variations in Cancer Cells. *PLOS Genetics*, 6(9):e1001090, September 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001090. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001090.

[2] Jon M. Laurent, Christine Vogel, Taejoon Kwon, Stephanie A. Craig, Daniel R. Boutz, Holly K. Huse, Kazunari Nozue, Harkamal Walia, Marvin Whiteley, Pamela C. Ronald, and Edward M. Marcotte. Protein abundances are more conserved than mRNA abundances across diverse taxa. *PROTEOMICS*, 10(23):4209–4212, December 2010. ISSN 1615-9861. doi: 10.1002/pmic.201000327. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201000327.

[3] Silvia Stingele, Gabriele Stoehr, Karolina Peplowska, Jrgen Cox, Matthias Mann, and Zuzana Storchova. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Molecular Systems Biology*, 8(1):608, January 2012. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2012.40. URL http://msb.embopress.org/content/8/1/608.

[4] Carlo G. Artieri and Hunter B. Fraser. Evolution at two levels of gene expression in yeast. *Genome Research*, 24(3):411–421, March 2014. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.165522.113. URL http://genome.cshlp.org/content/24/3/411.

[5] C. Joel McManus, Gemma E. May, Pieter Spealman, and Alan Shteyman. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Research*, 24(3):422–430, March 2014. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.164996.113. URL http://genome.cshlp.org/content/24/3/422.

[6] Linfeng Wu, Sophie I. Candille, Yoonha Choi, Dan Xie, Lihua Jiang, Jennifer Li-Pook-Than, Hua Tang, and Michael Snyder. Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79–82, July 2013. ISSN 1476-4687. doi: 10.1038/nature12223. URL https://www.nature.com/articles/nature12223.

[7] Noah Dephoure, Sunyoung Hwang, Ciara O'Sullivan, Stacie E Dodgson, Steven P Gygi, Angelika Amon, and Eduardo M Torres. Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife*, 3:e03023, July 2014. ISSN 2050-084X. doi: 10.7554/eLife.03023. URL https://elifesciences.org/articles/03023.

[8] Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation from RNA to protein. *Science*, page 1260793, December 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1260793. URL http://science.sciencemag.org/content/early/2014/12/17/science.1260793.

[9] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550, April 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.03.014. URL http://www.sciencedirect.com/science/article/pii/S0092867416302707.

[10] Joel M. Chick, Steven C. Munger, Petr Simecek, Edward L. Huttlin, Kwangbom Choi, Daniel M. Gatti, Narayanan Raghupathy, Karen L. Svenson, Gary A. Churchill, and Steven P. Gygi. Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, 534(7608):500–505, June 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature18270. URL http://www.nature.com/articles/nature18270.

[11] Koji Ishikawa, Koji Makanae, Shintaro Iwasaki, Nicholas T. Ingolia, and Hisao Moriya. Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes. *PLOS Genetics*, 13(1):e1006554, January 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006554. URL https://dx.plos.org/10.1371/journal.pgen.1006554.

[12] Sidney H. Wang, Chiaowen Joyce Hsiao, Zia Khan, and Jonathan K. Pritchard. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biology*, 19(1):83, 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1451-z.

[13] Zia Khan, Michael J. Ford, Darren A. Cusanovich, Amy Mitrano, Jonathan K. Pritchard, and Yoav Gilad. Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures. *Science*, 342(6162):1100–1104, November 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1242379. URL http://science.sciencemag.org/content/342/6162/1100.

[14] Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232, April 2012. ISSN 1471-0064. doi: 10.1038/nrg3185. URL https://www.nature.com/articles/nrg3185.

[15] Christine Vogel, Raquel de Sousa Abreu, Daijin Ko, Shu-Yun Le, Bruce A. Shapiro, Suzanne C. Burns, Devraj Sandhu, Daniel R. Boutz, Edward M. Marcotte, and Luiz O. Penalva. Sequence signatures and mRNA concentration can explain twothirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6(1):400, January 2010. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2010.59. URL http://msb.embopress.org/content/6/1/400.

[16] Bjrn Schwanhusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011. ISSN 1476-4687. doi: 10.1038/nature10098.

[17] Kevin Struhl. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology*, 14(2):103–105, February 2007. ISSN 1545-9993. doi: 10.1038/nsmb0207-103. URL http://www.nature.com/nsmb/journal/v14/n2/abs/nsmb0207-103.html.

[18] Daniel Hebenstreit, Miaoqing Fang, Muxin Gu, Varodom Charoensawan, Alexander van Oudenaarden, and Sarah A. Teichmann. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology*, 7(1):497, January 2011. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2011.28. URL http://msb.embopress.org/content/7/1/497.

[19] Harm van Bakel, Corey Nislow, Benjamin J. Blencowe, and Timothy R. Hughes. Most Dark Matter Transcripts Are Associated With Known Genes. *PLOS Biology*, 8(5):e1000371, May 2010. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000371. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000371.

[20] A. Singh, C. A. Vargas, and R. Karmakar. Stochastic analysis and inference of a two-state genetic promoter model. In *2013 American Control Conference*, pages 4563–4568, June 2013. doi: 10.1109/ACC.2013.6580542.

[21] Lifang Huang, Zhanjiang Yuan, Peijiang Liu, and Tianshou Zhou. Effects of promoter leakage on dynamics of gene expression. *BMC systems biology*, 9:16, March 2015. ISSN 1752-0509. doi: 10.1186/s12918-015-0157-z.

[22] Alexander F. Palazzo and T. Ryan Gregory. The Case for Junk DNA. *PLOS Genetics*, 10(5):e1004351, May 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004351. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004351.

[23] Rafik Neme and Diethard Tautz. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, 5:e09977, February 2016. ISSN 2050-084X. doi: 10.7554/eLife.09977. URL https://elifesciences.org/articles/09977.

[24] John P. Lloyd, Zing Tsung-Yeh Tsai, Rosalie P. Sowers, Nicholas L. Panchy, and Shin-Han Shiu. A Model-Based Approach for Identifying Functional Intergenic Transcribed Regions and Noncoding RNAs. *Molecular Biology and Evolution*, 35(6):1422–1436, June 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy035. URL https://academic.oup.com/mbe/article/35/6/1422/4938759.

[25] Mihaela Pertea, Alaina Shumate, Geo Pertea, Ales Varabyou, Florian P. Breitwieser, Yu-Chi Chang, Anil K. Madugundu, Akhilesh Pandey, and Steven L. Salzberg. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, 19(1):208, December 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1590-2. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1590-2.

[26] Simone Tiberi, Mark Walsh, Massimo Cavallaro, Daniel Hebenstreit, and Brbel Finkenstdt. Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics*, 34(17):i647–i655, September 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty568. URL https://academic.oup.com/bioinformatics/article/34/17/i647/5093237.

[27] Zhijin Wu, Yi Zhang, Michael L. Stitzel, and Hao Wu. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics*, 34(19):3340–3348, October 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty329. URL https://academic.oup.com/bioinformatics/article/34/19/3340/4984507.

[28] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1226.

[29] Samuel Marguerat, Alexander Schmidt, Sandra Codlin, Wei Chen, Ruedi Aebersold, and Jrg Bhler. Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell*, 151(3):671–683, October 2012. ISSN 0092-8674. doi: 10.1016/j.cell.2012.09.019. URL http://www.cell.com/article/S0092867412011269/abstract.

[30] Marta Mel, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, Jakob M. Goldmann, Dmitri D. Pervouchine, Timothy J. Sullivan, Rory Johnson, Ayellet V. Segr, Sarah Djebali, Anastasia Niarchou, The GTEx Consortium, Fred A. Wright, Tuuli Lappalainen, Miquel Calvo, Gad Getz, Emmanouil T. Dermitzakis, Kristin G. Ardlie, and Roderic Guig. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, May 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa0355. URL https://science.sciencemag.org/content/348/6235/660.

[31] Daniel Ramskld, Eric T. Wang, Christopher B. Burge, and Rickard Sandberg. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLOS Computational Biology*, 5(12):e1000598, December 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000598. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000598.

[32] Nicholas D. Hastie and John O. Bishop. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell*, 9(4, Part 2):761–774, December 1976. ISSN 0092-8674. doi: 10.1016/0092-8674(76)90139-2. URL http://www.sciencedirect.com/science/article/pii/0092867476901392.

[33] Daniel Hebenstreit and Sarah A. Teichmann. Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Physical Biology*, 8(3):035013, May 2011. ISSN 1478-3975. doi: 10.1088/1478-3975/8/3/035013. URL https://doi.org/10.1088%2F1478-3975%2F8%2F3%2F035013.

[34] Gnter P. Wagner, Koryu Kin, and Vincent J. Lynch. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences = Theorie in Den Biowissenschaften*, 132(3):159–164, September 2013. ISSN 1611-7530. doi: 10.1007/s12064-013-0178-3.

[35] Traver Hart, H Komori, Sarah LaMere, Katie Podshivalova, and Daniel R Salomon. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics*, 14(1):778, 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-778. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-778.

[36] Koryu Kin, MaurisC. Nnamani, VincentJ. Lynch, Elias Michaelides, and GnterP. Wagner. Cell-type Phylogenetics and the Origin of Endometrial Stromal Cells. *Cell Reports*, 10(8):1398–1409, March 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.01.062. URL http://www.sciencedirect.com/science/article/pii/S2211124715001047.

[37] Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, July 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2967. URL https://www.nature.com/articles/nmeth.2967.

[38] Stephen R. Piccolo, Michelle R. Withers, Owen E. Francis, Andrea H. Bild, and W. Evan Johnson. Multiplatform single-sample estimates of transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17778–17783, October 2013. ISSN 1091-6490. doi: 10.1073/pnas.1305823110.

[39] David C. Hoyle, Magnus Rattray, Ray Jupp, and Andrew Brass. Making sense of microarray data distributions. *Bioinformatics (Oxford, England)*, 18(4):576–584, April 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.4.576.

[40] Chuan Lu and Ross D. King. An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems. *Bioinformatics (Oxford, England)*, 25(16):2020–2027, August 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp360.

[41] Nicolas L. Bray, Harold Pimentel, Pll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, May 2016. ISSN 1546-1696. doi: 10.1038/nbt.3519. URL https://www.nature.com/articles/nbt.3519.

[42] Mihaela Pertea, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, March 2015. ISSN 1546-1696. doi: 10.1038/nbt.3122. URL https://www.nature.com/articles/nbt.3122.

[43] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1699114. URL http://aip.scitation.org/doi/10.1063/1.1699114.

[44] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL https://academic.oup.com/biomet/article/57/1/97/284580.

[45] STUART Geman and DONALD Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. In Martin A. Fischler and Oscar Firschein, editors, *Readings in Computer Vision*, pages 564–584. Morgan Kaufmann, San Francisco (CA), January 1987. ISBN 978-0-08-051581-6. doi: 10.1016/B978-0-08-051581-6.50057-X. URL http://www.sciencedirect.com/science/article/pii/B978008051581650057X.

[46] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, December 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.711. URL https://academic.oup.com/biomet/article/82/4/711/252058.

[47] Andrew Gelman, Xiao-Li Meng, and Hal Stern. POSTERIOR PREDICTIVE ASSESSMENT OF MODEL FITNESS VIA REALIZED DISCREPANCIES. *Statistica Sinica*, 6(4):733–760, 1996. ISSN 1017-0405. URL https://www.jstor.org/stable/24306036.

[48] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1262110. URL https://science.sciencemag.org/content/348/6235/648.

[49] Leehee Weinberger, Yoav Voichek, Itay Tirosh, Gil Hornung, Ido Amit, and Naama Barkai. Expression Noise and Acetylation Profiles Distinguish HDAC Functions. *Molecular Cell*, 47(2):193–202, July 2012. ISSN 1097-2765. doi: 10.1016/j.molcel.2012.05.008. URL http://www.sciencedirect.com/science/article/pii/S1097276512003917.

[50] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011. ISSN 1476-4687. doi: 10.1038/nature09906. URL https://www.nature.com/articles/nature09906.

[51] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015. ISSN 1476-4687. doi: 10.1038/nature14248. URL https://www.nature.com/articles/nature14248.

[52] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gbor Csrdi, Patrick Harrigan, Manuela Weier, Anglica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grtzner, Sven Bergmann, Rasmus Nielsen, Svante Pbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, October 2011. ISSN 1476-4687. doi: 10.1038/nature10532. URL https://www.nature.com/articles/nature10532.

[53] Andr M. M. Sousa, Ying Zhu, Mary Ann Raghanti, Robert R. Kitchen, Marco Onorati, Andrew T. N. Tebbenkamp, Bernardo Stutz, Kyle A. Meyer, Mingfeng Li, Yuka Imamura Kawasawa, Fuchen Liu, Raquel Garcia Perez, Marta Mele, Tiago Carvalho, Mario Skarica, Forrest O. Gulden, Mihovil Pletikos, Akemi Shibata, Alexa R. Stephenson, Melissa K. Edler, John J. Ely, John D. Elsworth, Tamas L. Horvath, Patrick R. Hof, Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger, Mark Reimers, Richard P. Lifton, Shrikant M. Mane, James P. Noonan, Matthew W. State, Ed S. Lein, James A. Knowles, Tomas Marques-Bonet, Chet C. Sherwood, Mark B. Gerstein, and Nenad Sestan. Molecular and cellular reorganization of neural circuits in the human lineage. *Science*, 358(6366):1027–1032, November 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan3456. URL http://science.sciencemag.org/content/358/6366/1027.

[54] sa Walln-Mackenzie, Hanna Wootz, and Hillevi Englund. Genetic inactivation of the vesicular glutamate transporter 2 (VGLUT2) in the mouse: What have we learnt about functional glutamatergic neurotransmission? *Upsala Journal of Medical Sciences*, 115(1):11–20, February 2010. ISSN 0300-9734. doi: 10.3109/03009730903572073. URL https://doi.org/10.3109/03009730903572073.

[55] Richard J. Reimer and Robert H. Edwards. Organic anion transport is the primary function of the SLC17/type I phosphate transporter family. *Pflgers Archiv European Journal of Physiology*, 447(5):629–635, February 2004. ISSN 0031-6768, 1432-2013. doi: 10.1007/s00424-003-1087-y. URL http://link.springer.com/10.1007/s00424-003-1087-y.

[56] Michael L. Wallace, Arpiar Saunders, Kee Wui Huang, Adrienne C. Philson, Melissa Goldman, Evan Z. Macosko, Steven A. McCarroll, and Bernardo L. Sabatini. Genetically Distinct Parallel Pathways in the Entopeduncular Nucleus for Limbic and Sensorimotor Output of the Basal Ganglia. *Neuron*, 94(1):138–152.e5, April 2017. ISSN 0896-6273. doi: 10.1016/j.neuron.2017.03.017. URL http://www.sciencedirect.com/science/article/pii/S0896627317302003.

[57] Jean-Sebastien Provost, Alexandru Hanganu, and Oury Monchi. Neuroimaging studies of the striatum in cognition Part I: healthy individuals. *Frontiers in Systems Neuroscience*, 9, October 2015. ISSN 1662-5137. doi: 10.3389/fnsys.2015.00140. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4596942/.

[58] Albert Jordan, Patricia Defechereux, and Eric Verdin. The site of HIV1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *The EMBO Journal*, 20(7):1726–1738, April 2001. ISSN 0261-4189, 1460-2075. doi: 10.1093/emboj/20.7.1726. URL http://emboj.embopress.org/content/20/7/1726.

[59] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology*, 4(10):e309, September 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040309. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040309.

[60] Siddharth S. Dey, Jonathan E. Foley, Prajit Limsirichai, David V. Schaffer, and Adam P. Arkin. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Molecular Systems Biology*, 11 (5):806, May 2015. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20145704. URL http://msb.embopress.org/content/11/5/806.

[61] Ron Skupsky, John C. Burnett, Jonathan E. Foley, David V. Schaffer, and Adam P. Arkin. HIV Promoter Integration Site Primarily Modulates Transcriptional Burst Size Rather Than Frequency. *PLOS Computational Biology*, 6(9): e1000952, September 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000952. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000952.

[62] Roy D. Dar, Brandon S. Razooky, Abhyudai Singh, Thomas V. Trimeloni, James M. McCollum, Chris D. Cox, Michael L. Simpson, and Leor S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1213530109. URL https://www.pnas.org/content/109/43/17454.

[63] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer Control of Transcriptional Bursting. *Cell*, 166(2):358–368, July 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.05.025. URL http://www.sciencedirect.com/science/article/pii/S0092867416305736.

[64] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. page 12, 2010.

[65] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp616. URL http://bioinformatics.oxfordjournals.org/content/26/1/139.

[66] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.

[67] Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R50, 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-5-r50. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2898062/.

[68] Jakob Lovn, DavidA. Orlando, AllaA. Sigova, CharlesY. Lin, PeterB. Rahl, ChristopherB. Burge, DavidL. Levens, TongIhn Lee, and RichardA. Young. Revisiting Global Gene Expression Analysis. *Cell*, 151(3):476–482, October 2012. ISSN 0092-8674. doi: 10.1016/j.cell.2012.10.012. URL http://www.sciencedirect.com/science/article/pii/S0092867412012263.

[69] Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, August 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty175. URL https://academic.oup.com/bioinformatics/article/34/16/2870/4956011.

[70] Christian Rdelsperger, Waltraud Rseler, Neel Prabh, Kohta Yoshida, Christian Weiler, Matthias Herrmann, and Ralf J. Sommer. Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. *Current Biology*, 28(19):3123–3127.e5, October 2018. ISSN 0960-9822. doi: 10.1016/j.cub.2018.07.041. URL http://www.sciencedirect.com/science/article/pii/S0960982218309400.

[71] Amy L. Bauernfeind, Erik J. Soderblom, Meredith E. Turner, M. Arthur Moseley, John J. Ely, Patrick R. Hof, Chet C. Sherwood, Gregory A. Wray, and Courtney C. Babbitt. Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. *Genome Biology and Evolution*, 7(8):2276–2288, August 2015. doi: 10.1093/gbe/evv132. URL https://academic.oup.com/gbe/article/7/8/2276/556944.