Supplement: Flexible isoform-level differential expression analysis with Ballgown

Alyssa C. Frazee[1], Geo Pertea[2,3],Andrew E. Jaffe[1,3,4], Ben Langmead[1,2,3,5], Steven L. Salzberg[1,2,3,5], & Jeffrey T. Leek[1,3,*]

March 2014

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
2. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine
3. Center for Computational Biology, Johns Hopkins University
4. Lieber Institute for Brain Development, Johns Hopkins Medical Campus
5. Department of Computer Science, Johns Hopkins University

∗ *Correspondance to jtleek@gmail.com*

1 *Tablemaker* output files

Tablemaker outputs the following set of related tab-delimited text files. *Tablemaker* is designed to be run on the output of *Cufflinks* and *Cuffmerge* but *Ballgown* can be used with any assembly output that can be converted into the following sets of tab-delimited files.

- *e_data.ctab*: exon-level expression measurements. One row per exon. Columns are *e_id* (numeric exon id), *chr*, *strand*, *start*, *end* (genomic location of the exon), and the following expression measurements for each sample:
  - *rcount*: reads overlapping the exon
  - *ucount*: uniquely mapped reads overlapping the exon
  - *mrcount*: multi-map-corrected number of reads overlapping the exon
  - *cov*: average per-base read coverage
  - *cov_sd*: standard deviation of per-base read coverage
  - *mcov*: multi-map-corrected average per-base read coverage

  - *mcov_sd*: standard deviation of multi-map-corrected per-base coverage

- *i_data.ctab*: intron- (i.e., junction-) level expression measurements. One row per intron. Columns are *i_id* (numeric intron id), *chr*, *strand*, *start*, *end* (genomic location of the intron), and the following expression measurements for each sample:
  - *rcount*: number of reads supporting the intron
  - *ucount*: number of uniquely mapped reads supporting the intron
  - *mrcount*: multi-map-corrected number of reads supporting the intron

- *t_data.ctab*: transcript-level expression measurements. One row per transcript. Columns are:
  - *t_id*: numeric transcript id
  - *chr*, *strand*, *start*, *end*: genomic location of the transcript
  - *t_name*: Cufflinks-generated transcript id
  - *num_exons*: number of exons comprising the transcript
  - *length*: transcript length, including both exons and introns
  - *gene_id*: gene the transcript belongs to
  - *gene_name*: HUGO gene name for the transcript, if known
  - *cov*: per-base coverage for the transcript (available for each sample)
  - *FPKM*: Cufflinks-estimated FPKM for the transcript (available for each sample)

- *e2t.ctab*: table with two columns, *e_id* and *t_id*, denoting which exons belong to which transcripts. These ids match the ids in the *e_data* and *t_data* tables.

- *i2t.ctab*: table with two columns, *i_id* and *t_id*, denoting which introns belong to which transcripts. These ids match the ids in the *i_data* and *t_data* tables.

2 Data and Notation

There are two distinct components to the data that *Ballgown* is equipped to analyze: the actual structure of the assembled transcriptome: (1) genomic locations of features and the relationships between exons, introns, transcripts and (2) genes and the expression measurements for the features in the transcriptome. Here we precisely define both the assembly structure and the associated data.

Assembly structure

The transcriptome is assembled based on a set R of aligned RNA-seq reads. We denote the yth read from the zth sample with r_{yz}, where $y = 1, ..., N_z$ and $z = 1, ..., n$, so there are n samples in the study, and sample z has N_z aligned reads.

The transcriptome assembled from the reads consists of four types of features: transcripts, genes, exons, and introns. These features all have start and finishing positions on the genome, which represent using the functions $s()$ and $f()$, e.g., $s(x)$ represents the start position of feature x. The K assembled transcripts are denoted by t_k, where $k = 1, ..., K$. These transcripts can be organized into G genes, denoted by g_l, $l = 1, ..., G$. Each gene can be represented by a set of transcripts falling within its boundaries:

$$g_l = \{t_k : s(t_k) > s(g_l) \text{ and } f(t_k) < f(g_l)\}$$

The assembly also contains M exons, each of which we represent as a closed interval of genomic locations:

$$e_m = [s(e_m), f(e_m)], m = 1, ..., M$$

With this notation, we can then represent transcript k as a subset of the M exons comprising the assembly:

$$t_k = \{e_m : m \in I_k\}, I_k \subset \{1, ..., M\}$$

Here, I_k represents the indices of the exons that make up transcript k. Note that the exon e_m can belong to several different transcripts. We can then easily define $s(t_k)$ and $f(t_k)$ in terms of exon boundaries:

$$s(t_k) = \min\{s(e_m) : m \in I_k\}$$
$$f(t_k) = \max\{f(e_m) : m \in I_k\}$$

Finally, let w_k represent the wth element of I_k. Then we can denote the wth intron in transcript k with an open interval:

$$i_{kw} = (f(e_{w_k}), s(e_{(w+1)_k}))$$

In other words, i_{kw} is simply the genomic interval between the wth and $w + 1$th exons of transcript k.

With these definitions in place, we can now precisely define the reads r_{yz}. An RNA-seq read is simply a subsequence of an RNA transcript. Using set notation, we can define each read using the form:

$$r_{yz} = \left\{ x \in [E, E'] : E < E' \text{ and } x, E, E' \in \bigcup_{m \in I_k} e_m \text{ for some } k \right\}$$

An assembly algorithm applied to the set of reads r_{yz} produces estimates of the exons: $\hat{e}_m, m = 1, \ldots, M$, transcripts: $\hat{t}_k, k = 1, \ldots K$ of the transcripts and genes: $\hat{g}_l, l = 1, \ldots, G$. Most current statistical models treat this assembly as fixed and correct when performing analyses. But as we will demonstrate in the methods section, assembled transcripts are subject to error and may be improved through statistical analysis [11, 18].

Expression Data

Next we can define expression measurements for each type of feature given a particular assembled set of transcripts. Here we define sensible expression measurements that are currently implemented in the *Ballgown* package, but the statistical models are flexible enough to handle other types of measures as well.

For each sample z, each transcript $\hat{t}_k$ has two measurements that are calculated by our upstream *Ballgown* preprocessing software: average per-base read coverage: $cov(t_k, z)$ and FPKM (fragments per kilobase of transcript per million mapped reads): $FPKM(t_k, z)$. Currently, these transcript-level measurements are estimated in *Cufflinks* via maximum likelihood; the procedure is described in detail by [19].

Each gene g_l has one expression measurement for each sample, $FPKM(g_l, z)$. This measurement is reconstructed from the transcripts in g_l as follows: first, the number of fragments per million mapped reads for sample z for each $t_k \in g_l$ is calculated by multiplying $FPKM(t_k, z)$ by the length of transcript t_k in kilobases. The gene's total fragments per million mapped reads is the sum of the transcript-level fragments per million mapped reads for all the transcripts in the gene. Finally, the gene-level FPKM is calculated by dividing the gene's total fragments per million mapped reads by the gene's length.

The *Ballgown* preprocessor also calculates average per-base read coverage for each exon in the assembly, given the assembly structure and the aligned reads R. For sample z, we have:

$$cov(e_m, z) = \frac{\sum_{r_{yz} \in R} \sum_{bp \in [s(e_m), f(e_m)]} \mathbb{1}\{bp \in r_{yz}\}}{f(e_m) - s(e_m) + 1}$$

Each exon also has a raw read count, defined as the number of reads whose alignments overlap that exon:

$$rcount(e_m, z) = \sum_{r_{yz} \in R} \mathbb{1}\{r_{yz} \cap e_m \neq \emptyset\}$$

The main expression measurement for introns is also raw read count, defined as the number of reads whose alignments support the intron in the sense that their alignments are split across that intron's neighboring exons:

$$rcount(i_{kw}, z) = \sum_{r_{yz} \in R} \mathbb{1}\{s(r_{yz}) \in e_m \text{ and } f(r_{yz}) \in e_{m'}\}$$

where $m \leq w_k$ and $m' \geq (w+1)_k$.

Statistical methods for detecting differential expression

After exploring the structure of the assembled transcriptome and performing any necessary transcript post processing, the next step is to identify transcripts or genes that are differentially expressed across groups. Here we outline a framework for statistical analysis of transcript and gene abundances. To make the ideas concrete we use FPKM as the expression measurement and transcripts as the feature of interest, but these can be replaced in the

following model definitions with any of the expression measurements and any of the available genomic features in the assembly (genes, transcripts, exons, or introns).

Differential expression tests are implemented as follows: for each transcript $\hat{t}_k$, the following model is fit:

$$h(FPKM(\hat{t}_k, z)) = \alpha_k + \sum_{p=1}^{P} \beta_{pk} X_{zp} + \varepsilon_{zk} \tag{1}$$

where:

- $FPKM(\hat{t}_k, z)$ is the FPKM expression measurement for transcript k for sample z
- h is a transformation [2] to reduce the impact of mean-variance relationships observed in the counts [1]. For example, the transformation $h(\cdot) = \log_2(\cdot + 1)$ is commonly applied in the analysis of sequence-count data [8].
- α_k represents the baseline expression for transcript k
- X_{zp} represents covariate p for sample z. These covariates differ by experiment type. X_{z1} generally represents a library size adjustment for sample z; `ballgown`'s default for this value is $X_{1z} = \text{median}_k\{FPKM(\hat{t}_k, z)\}$
- β_{pk} quantifies the association of covariate p on te expression of transcript k
- ε represents residual measurement error

A flexible approach to differential expression is to compare nested sub models of model (1) using parametric F-tests [15]. The null hypothesis can be as flexible as any linear contrast of the coefficients β_{pk} but for simplicity we focus on null hypotheses of the form: $H_0 : \beta_{pk} = 0, p \in \mathcal{S}$ versus the alternative that all β_{pk} are nonzero. The general principle is that a model including any potential confounders plus the covariate(s) of interest – a 0/1 indicator for group in the two-group comparison, several indicator variables for the multi-group comparison, or a generalized additive model [6] for a time variable for timecourse experiments – is compared with a model that includes only the potential confounders. For the two models fit for each transcript k, *Ballgown* calculates the statistic

$$F = \frac{\frac{RSS_0 - RSS_1}{P_1 - P_0}}{\frac{RSS_1}{n - P_1}}$$

where RSS_0 represents the residual sum of squares from the model without group or time covariates, RSS_1 represents the residual sum of squares from the model including the covariates of interest, P_0 is the number of covariates in the smaller model, P_1 is the number of covariates in the larger model, and n is the total number of samples. Under the null hypothesis that the larger model does not fit the data significantly better than the smaller model, this statistic follows an F distribution with $(P_1 - P_0, n - P_1)$ degrees of freedom, so p-values can be generated by comparing the two models for each transcript k [10]. We control for multiple testing using standard FDR controlling procedures [16].

3 Simulation studies

We performed two separate simulation studies. For both studies, reads were generated from 2745 annotated transcripts on Chromosome 22 from Ensembl [5], using genome build GRCh37 and Ensembl version 74. Data was generated for 20 biological replicates, divided into two groups of 10, where 274 transcripts were randomly chosen to be differentially expressed (at a 6x increase in expression level) in one of the two groups, randomly chosen.

The first simulation study is presented in the main manuscript and was set up as follows:

- Expression was measured in FPKM. Each transcript's mean FPKM value was set to be the mean nonzero FPKM value from a randomly selected transcript with mean FPKM larger than 100 in the assembled GEUVADIS dataset.
- We defined a log-log relationship between a transcript's mean expression level and the variance of its expression levels:

$$\log \text{variance} = 2 \log \text{mean} + 0.5$$

  in order to encompass biological and technical variability.
- Then, for each transcript, we randomly drew FPKM expression values from a log-normal distribution with the pre-set mean and variance. For the differentially expressed transcripts, the pre-set mean FPKM was 6 times larger in one group than in the other.
- For each transcript, we also set a sample's expression level to 0 with probability p_0, which was estimated from the GEUVADIS data: for each simulated transcript, p_0 was randomly drawn from the empirical distribution of the proportion of samples with zero expression, over transcripts in the GEUVADIS dataset with mean FPKM larger than 100.
- To translate the pre-set FPKM value into a number of reads to be generated from a transcript for a given sample, we used the definition of FPKM and calculated the number of "fragments" (reads) that should be generated from a transcript by multiplying the set FPKM value by the transcript's length over 1000, then multiplying by an approximate library size of 150,000 reads, over 1 million.

This simulation setup made it such that more reads were generated from longer transcripts, as is expected with RNA-seq protocols.

A second simulation was also conducted with a slightly simpler setup:

- Expression was defined directly by the number of reads being generated from each transcript (instead of using FPKM).
- The mean number of reads generated from each transcript was set to be 300, unless the transcript was randomly selected to be overexpressed in one group, in which case, that group's mean read number for that transcript was 1800.

- The actual number of reads to be simulated from a transcript was drawn from a negative binomial distribution with mean $\mu = 300$ or 1800, and size equal to 0.005μ (so, 1.5 for $\mu = 300$ and 9 for $\mu = 1800$). Note that in the negative binomial distribution, the variance is equal to $\mu + \mu^2/\text{size}$.
- Each sample's read counts were scaled and rounded such that approximately 600,000 reads were generated per sample.

For both these scenarios, the specified number of reads was then generated from transcripts using the *polyester* package. These simulated reads were then processed through the *Tophat2 -Cufflinks -Cuffdiff2* pipeline and the *Tophat2 -Cufflinks -Tablemaker -Ballgown* pipeline (Figure 1a, main manuscript).

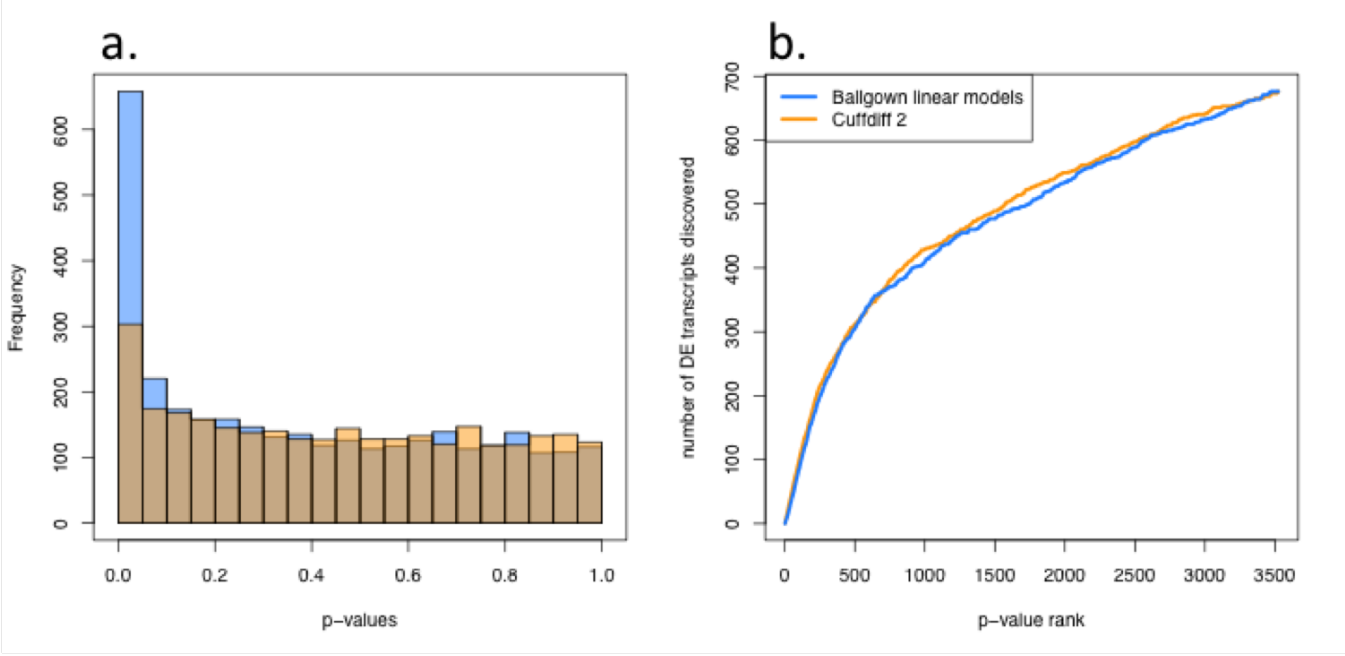

Figure 1: **Simulation results for transcript-length-independent simulation a.** Histograms of p-values from the transcript-length-independent simulation study, comparing the 10 cases to the 10 controls. In this study, *Cuffdiff2* performed adequately; the conservative bias observed in more realistic data (Figure 2(c), main text) was not seen here. **d.** A plot of the ranking of transcripts from most differentially expressed to least (x-axis) versus the number of truly differentially expressed transcripts (y-axis), using the transcript-length-independent simulated dataset. Among the top 100 transcripts ranked by each method for differential expression, 98 are truly differentially expressed for *Cuffdiff2* and 84 are for *Ballgown*.

Model fitting in simulated data

In the simulated data we fit the nested set of linear models:

$$
\begin{aligned}
H_A &: \log_2(FPKM_i + 1) = \beta_0^* + \beta_1^* grp_i + \eta^* q75_i + \epsilon_i^* \\
H_0 &: \log_2(FPKM_i + 1) = \beta_0 + \eta q75_i + \epsilon_i
\end{aligned}
$$

where grp_i is the value of the group indicator for sample i and $q75$ is a library-size normalizing constant equal to the sum of the nonzero FPKM values to the 75th percentile [12].We then tested the hypothesis $H_0 : \beta_1^* = 0$ versus the alternative that the coefficient was non-zero. For the analysis with average coverage we replaced $FPKM_i$ with $acov_i$ in the above equations.

Results from transcript-length-independent simulation

In this second, less-realistic simulation scenario, *Cuffdiff2* and *Ballgown* performed comparably (Supplementary Figure 1).

4 Data Analyses

Preprocessing GEUVADIS

We downloaded the FASTQ files from the GEUVADIS project from `http://www.ebi.ac.uk/ena/data/view/ERP001942` and ran the pipeline *Tophat2 - Cufflinks - Cuffmerge- Tablemaker* to create the set of tables described in the previous section. We then created a *Ballgown* object using the *Ballgown* package and matched the phenotype data available from `http://www.ebi.ac.uk/ena/data/view/ERP001942` along with additional QC data.

InSilico DB analysis

InSilico DB [3] includes processed data from public experiments on the Sequence Read Archive. We downloaded the *Cuffdiff2* output from the cancer versus normal and developmental data sets from InSilico DB on March 5th, 2014. We extracted the p-values for differential expression for the cancer versus normal comparison[7] and the embryonic stem cells versus preimplantation blastomeres data. We also reformatted the FPKM values from this analysis and applied the linear models included in the *Ballgown* package to perform the comparison. The versions and parameters for the software used by *InSilico DB* were *cufflinks, cuffmerge, cuffdiff: v 2.0.2*, *cufflinks -p 6 -q*, *tophat: v 2.0.4 –mate_inner_dist 80 –no-coverage-search* (personal communication Alain Coletta from the InSilico DB).

RIN analysis

We filtered to the 464 unique replicates as described by GEUVADIS [9] and analyzed only transcripts with FPKM > 0.1. We first searched for differential expression with respect to RNA quality (RIN) using the following set of nested linear models to each transcript.

$$H_A \;:\; \log_2(FPKM_i + 1) = \beta_0^* + \sum_{t=1}^{4} \beta_t^* \text{spline}_t(RIN_i) + \sum_{p=1}^{5} \gamma_p^* 1(Pop_i = p) + \eta^* q75_i + \epsilon_i^*$$

$$H_0 \;:\; \log_2(FPKM_i + 1) = \beta_0 + \sum_{p=1}^{5} \gamma_p 1(Pop_i = p) + \eta q75_i + \epsilon_i$$

Here i indicates sample and the subscript for transcript has been suppressed for clarity. H_0 denotes the null model and H_A denotes the alternative. The first set of terms encode a natural cubic spline fit with 4 degrees of freedom between the RIN values and the FPKM levels; the term $\text{spline}_t(RIN_i)$ refers to the tth B-spline basis term for sample i. The second set of terms encode a factor model for the relationship between population and FPKM and the last term is a library size normalization term that consists of the sum of the non-zero FPKMs up to the 75th percentile for that sample [12]. We then tested the hypothesis that $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus the alternative that at least one coefficient was non-zero. All transcripts with a Q-value [17] less than 0.05 were called significant.

Next we attempted to identify transcripts that were significantly better explained by a non-linear polynomial fit, rather than a linear trend. We fit the following nested set of models:

$$H_A \;:\; \log_2(FPKM_i + 1) = \beta_0^* + \sum_{t=1}^{3} \beta_t^* RIN_i^t + \sum_{p=1}^{5} \gamma_p^* 1(Pop_i = p) + \eta^* q75_i + \epsilon_i^* \quad (2)$$

$$H_0 \;:\; \log_2(FPKM_i + 1) = \beta_0 + \beta_1 RIN_i + \sum_{p=1}^{5} \gamma_p 1(Pop_i = p) + \eta q75_i + \epsilon_i \quad (3)$$

and tested the hypothesis that $H_0 : \beta_2 = \beta_3 = 0$ versus the alternative that at least one of the higher order polynomial coefficients was nonzero. Again, all transcripts with a Q-value [17] less than 0.05 were called significant.

The transcripts in the figure were statistically significant at the FDR 5% level for this second analysis. In the plots, the curves represent the fitted values for the average library size within each population. We show one example each of a positive and negative relationship between expression and RIN. While there were several examples of associations in both directions, there were more positive associations, as expected (Supplementary Figure 2).

eQTL analysis

We downloaded genotype information for the GEUVADIS cohort from `ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/genotypes/`. We filtered to only SNPs with a minor allele frequency greater than 5%. We used the processed transcriptome data from *Tablemaker* as described above. We removed samples that were

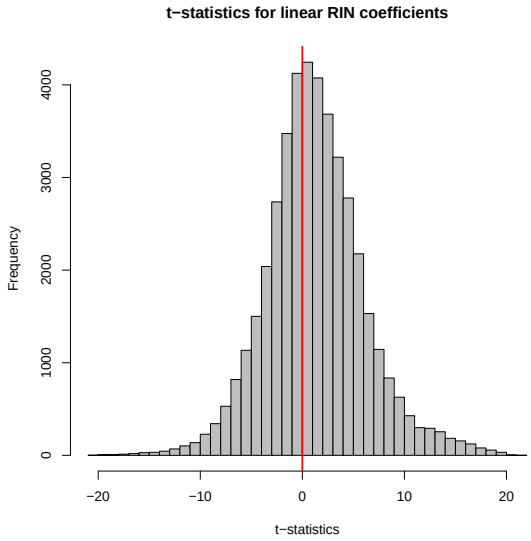

Figure 2: **Distribution of t-statistics for the linear RIN term for GEUVADIS transcripts.** These are moderated t-statistics calculated with *limma* for the β_1 coefficient in model (3), indicating directionality of the RIN-FPKM relationship. We observe associations in both directions, but as expected, there are more positive associations.

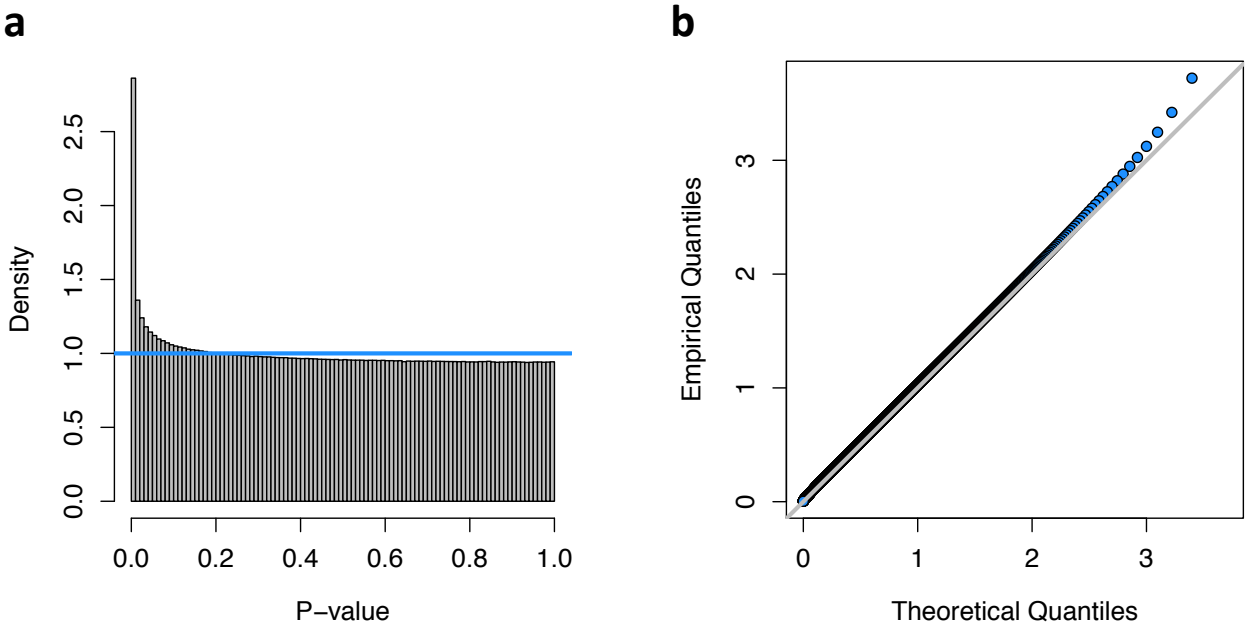

Figure 3: **Distribution of statistical significance scores for all cis-eQTL tests a.** P-value histogram for all p-values from cis-eQTL tests, the estimated fraction of null hypotheses is 94.2%. **b.** QQ-plot of -log10(p-values) versus theoretical quantiles shows no gross deviation from expected behavior.

sequenced multiple times according to the protocol described by GEUVADIS [9]. We calculated the first three principal components of the genotype data using the Plink software [13]. We filtered to transcripts with an average FPKM > 0.1 and took the log2 transform of the FPKM values. We then used the MatrixEQTL package [14] to perform the eQTL analysis testing an additive linear regression model for the SNPs adjusting for three expression prinicipal components and three genotype principal components. We filtered to only transcript-SNP pairs that were no more than 1000Kb apart.

We recorded the histogram of p-values from all transcript-SNP pairs. We calculated an estimate of the fraction of null hypotheses based on the distribution of observed p-values [17] and obtained an estimate of $\hat{\pi_0} = 0.942$. The p-value histogram (Figure 3a) and QQ-plot of -log10(p-values) (Figure 3b) versus their theoretical distribution under the null do not show any gross deviation suggesting unmodeled confounding [4].

For the transcript overlap analysis we downloaded the list of significant cis-eQTL from `ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/genotypes/` for the EUR and YRI populations. We identified all Ensembl genes overlapped to any degree by each assembled transcript. We then calculated the number of gene-SNP pairs in common between the GEUVADIS EUR and YRI analyses and our eQTL analysis.

5 Software

1. *Ballgown* - `https://github.com/alyssafrazee/ballgown/` Installation instructions and tutorial for use are available at `https://github.com/alyssafrazee/ballgown/blob/master/README.md`

2. *Tablemaker* - `https://github.com/alyssafrazee/ballgown/tree/master/tablemaker` Installation instructions available at `https://github.com/alyssafrazee/ballgown/blob/master/README.md`

3. *polyester* - `https://github.com/alyssafrazee/ballgown/tree/master/polyester` Installation instructions for polyester are here: `https://github.com/alyssafrazee/ballgown/blob/master/polyester/README.md`

6 Scripts and Data

Scripts and data will be available at: `https://github.com/alyssafrazee/ballgown_code/`

References

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[2] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[3] Alain Coletta, Colin Molter, Robin Duqué, David Steenhoff, Jonatan Taminau, Virginie De Schaetzen, Stijn Meganck, Cosmin Lazar, David Venet, Vincent Detours, et al. Insilico db genomic datasets hub: an efficient starting point for analyzing genome-wide studies in genepattern, integrative genomics viewer, and r/bioconductor. 2012.

[4] B Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[5] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2014. *Nucleic acids research*, 42(D1):D749–D755, 2014.

[6] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical science*, pages 297–310, 1986.

[7] Sang Cheol Kim, Yeonjoo Jung, Jinah Park, Sooyoung Cho, Chaehwa Seo, Jaesang Kim, Pora Kim, Jehwan Park, Jihae Seo, Jiwoong Kim, et al. A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. *PloS one*, 8(2):e55596, 2013.

[8] Ben Langmead, Kasper D Hansen, Jeffrey T Leek, et al. Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome Biol*, 11(8):R83, 2010.

[9] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013.

[10] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[11] Bo Li and Colin Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[12] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 2013.

[13] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[14] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[15] Gordon K Smyth et al. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):3, 2004.

[16] J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[17] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[18] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2012.

[19] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.