Estimating cis and trans contributions to differences in gene regulation

Ingileif B. Hallgrímsdóttir[1], **Maria Carilli**[1], **and Lior Pachter**[1,2]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125, USA
[2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125, USA

We describe a coordinate system and associated hypothesis testing framework for determining whether *cis* or *trans* regulation is responsible for differences in gene expression between two homozygous strains or species. We apply our framework to data from single replicate studies on yeast strains and human-chimpanzee hybrid cells, as well as to data from a mouse study with replicates, showing marked differences between our gene regulatory assignments and those previously reported. We also show how our multi-sample framework can determine the context dependency of *cis* and *trans* effects as well as explicitly model different hypotheses regarding the underlying mechanism of *trans* regulation.

Correspondence: ***Ingileif Hallgrimsdottir (ingileif@caltech.edu)***

Introduction

In 1961, Jacob and Monod developed a theory of gene regulation in which they distinguished local effects (*cis*) from distal regulation (*trans*) (1). Their work immediately raised the question of the relative contributions of these two regulation modalities (2, 3). One approach for assessing whether *cis* or *trans* regulation is responsible for differences in gene expression between strains or species is to compare differences in expression of genes in parents to allele-specific differences in F_1 hybrids. This approach was explored in (4), who used crosses of C57BL/6J and CAST/Ei mice to study regulatory mechanisms that could explain differences in gene expression between parental strains. Their approach was developed in (5, 6), who used pyrosequencing to study the regulation differences between *D. melanogaster* and *D. simulans*. With the advent of RNA-seq, genome-wide scans were possible, and (7) examined RNA-seq from F_1 crosses of C57BL/6J and CAST/EiJ to tease apart *cis* and *trans* contributions to gene expression differences between the parental strains. Similarly, (8) performed such an RNA-seq analysis using *Drosophila* lines.

Formally, the idea of using crosses to study *cis* and *trans* contributions to differences in gene expression between strains or species is as follows: consider a gene with expression X_{P1} in a homozygous strain 1, X_{P2} in a homozygous strain 2, and expression X_{H1} for the haplotype from strain 1 in the F_1 cross of 1 and 2, and expression X_{H2} for the haplotype from strain 2 in the F_1 cross of 1 and 2. Let $R_P = \log_2\left(\frac{X_{P1}}{X_{P2}}\right)$ and $R_H = \log_2\left(\frac{X_{H1}}{X_{H2}}\right)$. That is, R_P corresponds to the log2 fold change difference in expression between the two parental strains, and R_H to the log2 fold change difference between the expression of the hybrid haplotypes in the F_1 offspring. The connection between R_P, R_H, and regulation is as follows: consider that a gene can be regulated via *cis*, *trans*, or both (Fig. 1A). Gene expression measurements in the parents and hybrid (Fig. 1B), can be used to infer the nature of regulation underlying the difference in expression in the parental strains (Fig. 1C). Specifically, amending the classification of (6), we have:

- ***conserved:*** No change in gene expression indicating there has been no change in regulation, i.e. $R_P = 0$ and $R_H = 0$, which implies $R_P - R_H = 0$.
- ***cis:*** The relative difference in gene expression between the parents is the same as between the haplotypes in the hybrid indicating that the difference in parents is due to local *cis* effects, i.e. $R_H \neq 0$, $R_P \neq 0$, and $R_P = R_H$ which implies $R_P - R_H = 0$, arises from changes only in cis-regulatory elements.
- ***trans:*** Gene expression from the two haplotypes in the hybrid is the same, indicating that differences between the parents resulted from non-local *trans* regulation, i.e. $R_H = 0$ and $R_P \neq 0$, which implies $R_P - R_H \neq 0$, arises from a change only in *trans*-regulatory elements.
- ***cis + trans:*** $R_H \neq 0$ and $R_P \neq R_H$ with $\text{sgn}(R_H) = \text{sgn}(R_P - R_H)$ arises as a result of change in both *cis*- and *trans*-regulatory elements with changes in *cis* and *trans* contributing to changes in gene expression between strains in the same direction.
- ***cis × trans:*** $R_H \neq 0$ and $R_P \neq R_H$ with $\text{sgn}(R_H) \neq \text{sgn}(R_P - R_H)$ arises as a result of compensatory change in both *cis*- and *trans*-regulatory elements with changes in *cis* and *trans* contributing to changes in gene expression between strains in the opposite direction.

This classification corrects (6), which fails to properly assign regulation differences to *cis* + *trans* when both $R_H < 0$ and $R_P < 0$. The relationships between R_P and R_H in general, can be visualized as lines and regions in a 2D plot (5), as illustrated in Fig. 1D(i). While this direct representation of R_P and R_H is useful, a quantitative assessment of the gene regulatory modalities reflected in R_P and R_H requires a biologically meaningful notion of distance between points in Fig. 1D(i). Consider, for example, the situation where $X_{P1} = 2X_{P2}$ and $X_{H1} = 2X_{H2}$, i.e. a 2-fold change in

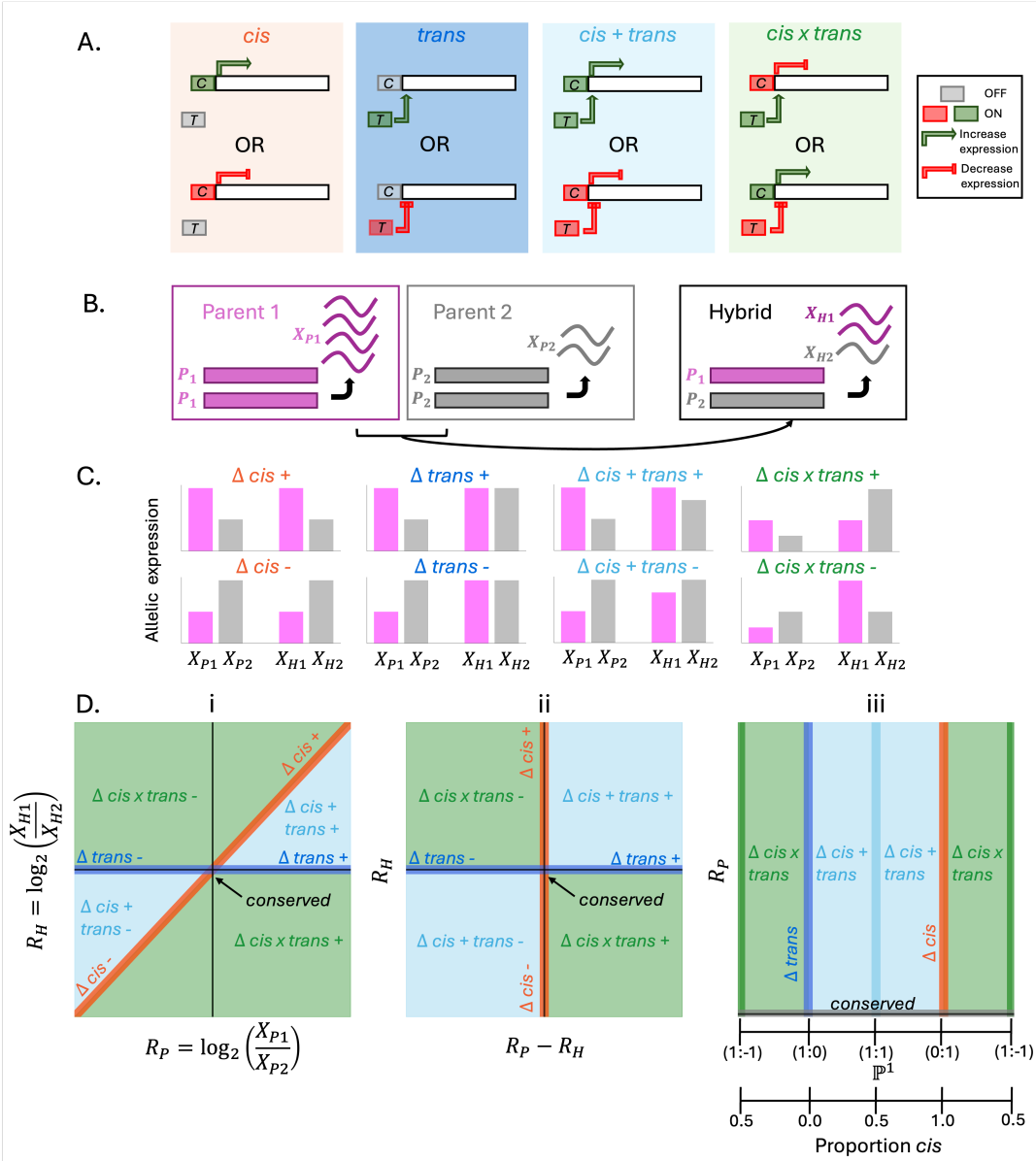

Fig. 1. Geometry of parental and hybrid expression ratios can be used to assess regulatory differences between homozygous crosses. A) Diagram of the *cis*, *trans*, *cis + trans*, and *cis* $\times$ *trans* types of regulatory differences. B) Homozygous parents with haplotypes P_1 and P_2, counts of RNA molecules for a gene (X_{P1} and X_{P2} respectively), and the haplotypes in an F_1 hybrid along with counts for the gene (X_{H1} and X_{H2}). C) Differences in regulation between the parents are reflected in distinct ratios between counts X_{P1}, X_{P2} and X_{H1}, X_{H2}. D. i) Illustration of how regulation differences emerge in log-fold changes R_P and R_H, D. ii) Linear transformation of $R_P = \log_2\left(\frac{X_{P1}}{X_{P2}}\right)$ and $R_H = \log_2\left(\frac{X_{H1}}{X_{H2}}\right)$ to yield orthogonal *cis* and *trans* coordinates, D. iii) Representation of proportion *cis* in real projective space $\mathbb{P}^1$.

gene expression between the parents due solely to *cis* regulation which corresponds to the point $(1,1)$ in Fig. 1D(i). The distance from this point to the origin (conserved), is $\sqrt{2}$, whereas the same 2-fold difference in gene expression in the parents due solely to *trans* regulation with $R_P = 1$ and $R_H = 0$ is distance 1 from the origin. This imbalance can be corrected via a linear transformation.

Results

Geometry. To decouple the effects of *cis* and *trans* regulation on R_P and R_H we begin by noting that if the difference in parental expression is solely due to *cis* regulation, then $R_P = R_H$, or equivalently $R_P - R_H = 0$ (orange vertical line in Fig. 1D(ii)). If the difference in parental expression is solely due to *trans* regulation, then $R_H = 0$ (blue horizontal line in Fig. 1D(ii)). Therefore, the transformation from Fig. 1D(ii) to Fig. 1D(i) is obtained by

$$\begin{pmatrix} R_P \\ R_H \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} R_P - R_H \\ R_H \end{pmatrix}. \tag{1}$$

Thus, the inverse transformation from the coordinate system in Fig. 1D(i) to the coordinate system in Fig. 1D(ii) is given by

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \tag{2}$$

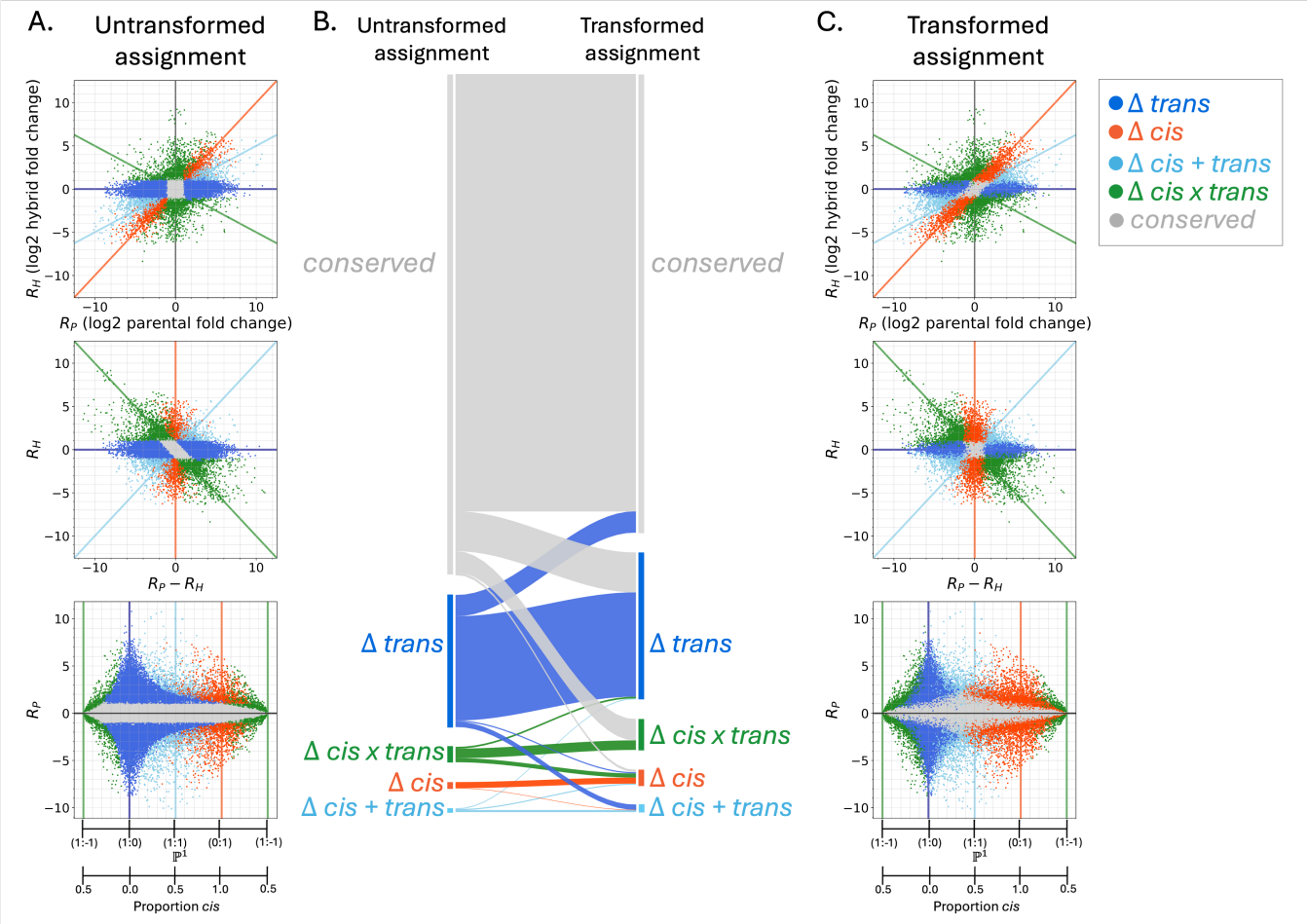

Fig. 2. Contributions of *cis* and *trans* regulation to differences in gene expression between yeast strains, with data from (9) (n = 285,777 gene-cross combinations from 179 unique yeast parent crosses). A) Untransformed ratios, transformed ratios, and log2 of parental fold change versus proportion *cis*, colored by regulatory assignments from (9). B) Comparison of the results of (9) to the assignment of genes based on the proposed hypothesis testing and geometric assignment procedure. C) Reanalysis of the 285,777 gene-cross expression data from (9) using the proposed hypothesis testing and geometric assignment procedure.

i.e., the transformation inverts the mapping of the vertical axis to the diagonal *cis* line.

The determination of whether a difference in parental gene expression is due to *cis* or *trans* can now be understood to be an assessment of whether the line passing through the origin and a point $(R_P - R_H, R_H)$ is a perturbation (due to noise in gene expression measurement) of the line Δ *trans*, the line Δ *cis* or sufficiently far away from the axes in Fig. 1D(ii) to merit a designation of Δ *cis* + *trans* or Δ *cis* × *trans* (the designations are enumerated in Supp. Table 1). In other words, the sufficient statistic is a point in real projective space $\mathbb{P}^1$ (Fig. 1D(iii)), and the proportion of the difference in gene expression between parents that can be attributed to *cis* can be understood to be a scaling of the angle of the line through the origin corresponding to the point in $\mathbb{P}^1$, i.e.,

$$\text{proportion } cis = \frac{2}{\pi}\left|\text{atan2}(R_H,\ R_P - R_H)\right|. \qquad \textbf{(3)}$$

Statistics. The determination of whether a measurement (R_P, R_H) reflects a difference in gene expression between parents due to *cis* or *trans* regulation, or both, requires a statistical assessment (10). Specifically, hypothesis tests can be used to reject a null hypothesis of a difference in gene expression being due solely to *trans* or solely to *cis*. Geometrically, as evident from the linear transformation on log-fold changes, these two tests correspond to testing whether one can reject the null hypotheses that $(R_P - R_H, R_H)$ is located on the x- and y- axes respectively.

We first developed a framework for hypothesis testing in this case where there are no replicates of the gene expression measurements, as in (9), which consists of bulk RNA-seq performed on hybrid and parental strains of genetically divergent *Saccharomyces cerevisiae* (see Methods, Supp. Figs. 1,2). Briefly, (9) generated hybrid crosses from 26 parental yeast isolates derived from diverse environmental conditions. In (9), genes were classified according to the representation shown in Fig. 1D(i). First, both R_P and R_H were tested for statistically significant differences from 0 and assigned "null" (our *conserved*) if both tests failed to reject the null hypothesis (one-sample allele-specific expression tests); then, changes between parental and hybrid allelic ratios were tested for significance (two-sample allele-specific expression tests). For genes that passed significance thresholds, regulatory assignments were made by dividing the untransformed 2D plane into cones (Fig. 2A). We reassigned genes based on our hypothesis tests as derived from the transformed coordi-

nate system (Fig. 2C), with one test with the null hypothesis of *cis* regulation and one with the null hypothesis of *trans* regulation, thereby putting the two regulation strategies on equal footing. Compared to the original study, we found major differences in assignment (Fig. 2B, Supp. Fig. 3).

Interestingly, whereas (9) conclude that "the transcriptome is globally buffered at the genetic level mainly due to trans-regulatory variation in the population", we find that a considerable amount of difference in gene expression can be attributed to *cis* (Supp. Fig. 3C), with the difference due to (9) deriving assignments in the untransformed coordinate system using a statistical testing procedure that treats *cis* and *trans* regulation asymmetrically. Specifically (9) report 57,253 cases where gene expression difference is due to *trans* regulation (Supp. Fig. 3A). We find 51,627 cases that can be assigned to *trans* (Supp. Fig. 3B). These numbers are similar; however, (9) report 2,804 cases assigned to *cis* (Supp. Fig. 3A), whereas we find *17,112* (Supp. Fig. 3B). Furthermore, we find 4,807 cases assigned to *cis* + *trans* (Supp. Fig. 3B) versus 1,727 in (9) (Supp. Fig. 3A). Overall, there is a marked difference between our results and those of (9) (Supp. Fig. 3C).

In the case where replicates have been obtained, a generalized linear model (GLM), as is commonly used for differential expression in bulk RNA-seq (11, 12), can be adapted to utilize gene expression variance estimates within parents and hybrids. We tested our approach (see Methods) on data from (13), in which two wild-derived mouse strains and their F1 hybrids were used to study the effect of genotype and environment (temperature) on gene expression divergence in rapidly evolving strains. Two inbred lines of house mice were derived from mice from a cold environment (Saratoga Springs, New York, USA) and a warm environment (Manaus, Amazonas, Brazil) (13), and two tissues, brown adipose tissue (BAT) and liver, were evaluated in each of the strains and their F1 hybrids, also raised at the two different temperatures. The experiment was performed in six replicates each of male and female parents, as well as six replicates each of male and female F1 hybrids (13) (Fig. 3A). The GLM framework allows the joint analysis of all samples (parents and hybrids) rather than requiring that separate tests be run on separate subsets of the data (e.g., differential expression between parents, differential expression between hybrid allelic expression, and a ratio test as in (13)). It further enables explicit modeling of specific biological hypotheses of *trans* regulatory action (see Methods, Supp. Figs. 4 – 6).

Restricting our analysis to a single tissue, temperature, and sex (brown adipose tissue collected from male mice reared in the cold temperature), we fit a GLM with the log additive hypothesis and classified genes based on the reduced models of no *cis* or no *trans* regulation (see Methods, Fig. 3B). Consistent with our single sample analyses, we noted an increase in the number of *cis* assigned genes from 478 in the original study (13) to 877 (Fig. 3C). This trend held in different tissue and temperatures (Supp. Fig. 7). We further fit the log additive model to all male samples, including environment (cold/warm) and organ (liver/BAT) specific *cis* and *trans* weights (Fig. 3D). We identified genes that exhibited significant overall *cis* and *trans* regulatory changes (Fig. 3E, top panel), as well as those that were only significantly different in one organ (Fig. 3E, middle panel) or environment (Fig. 3E, bottom panel). Among these genes were several with functions related to body fat and temperature. For example, *Coasy*, which was identified as having an overall significant *trans* regulatory difference between the strains (Fig. 3E, top panel), codes for CoA (coenzyme A) synthase. CoA is critical for energy production and fatty acid synthesis, and other proteins involved in its production pathway have been shown to regulate adaptive thermogenesis in mice (14). *Samd8*, identified as having tissue-specific regulatory patterns, codes for a protein that synthesizes sphingolipids, which act as key regulators in brown adipose tissue with accumulation linked to tissue aging and impaired lipid homeostasis (15). The gene that produces mitogen-activated protein kinase NIK, *Map3k14*, was found to have a significant temperature-dependent *cis* effect: related proteins in the MAPK family mediate heat-shock reactions in mice, with their inhibition leading to reduced protection against heat (16). These genes are plausible candidates for differential regulation between mouse strains adapting to life in cold and warm environments. The flexibility of the GLM framework facilitates identification of genes with unique regulatory patterns, such as the tissue and organ-specific effects, a key improvement over post-hoc comparisons of multiple model fits that would be otherwise required.

Proportion *cis*. In addition to naturally revealing the appropriate hypothesis tests to conduct for attribution of gene expression difference in parents to *cis* or *trans*, the linear transformation we propose leads directly to a meaningful measure of the proportion of difference in gene expression that can be attributed to *cis* regulation (or *trans*) (Equation 3). To illustrate this, we re-analyzed a dataset of gene expression from human-chimpanzee cell line hybrids (17), calculating the proportion *cis* according to Equation 3. In (17), the proportion *cis* was calculated using slope, as is natural to do when working in the untransformed coordinate framework, whereas the correct calculation (Equation 4) uses angle in the transformed coordinate system. While the absolute differences are small (Fig. 4A), with a maximum difference of 0.045 (see Methods), the relative difference is large when the proportion *cis* is small (Fig. 4B), and can be as high as 57% (see Methods). Moreover, our computation provides a biologically interpretable measure of proportion *cis*. We found several genes in (17) exhibiting high variance in proportion *cis* (Fig. 4C), and identified interesting differences between cell types (Supp. Fig. 8 – 11).

Discussion

The use of crosses between strains to identify the nature of differential regulation is a powerful tool for genetics studies that is particularly relevant now that single-cell RNA-seq can be used for cell type resolution. Moreover, while original studies were limited to a handful of genes, genome-wide

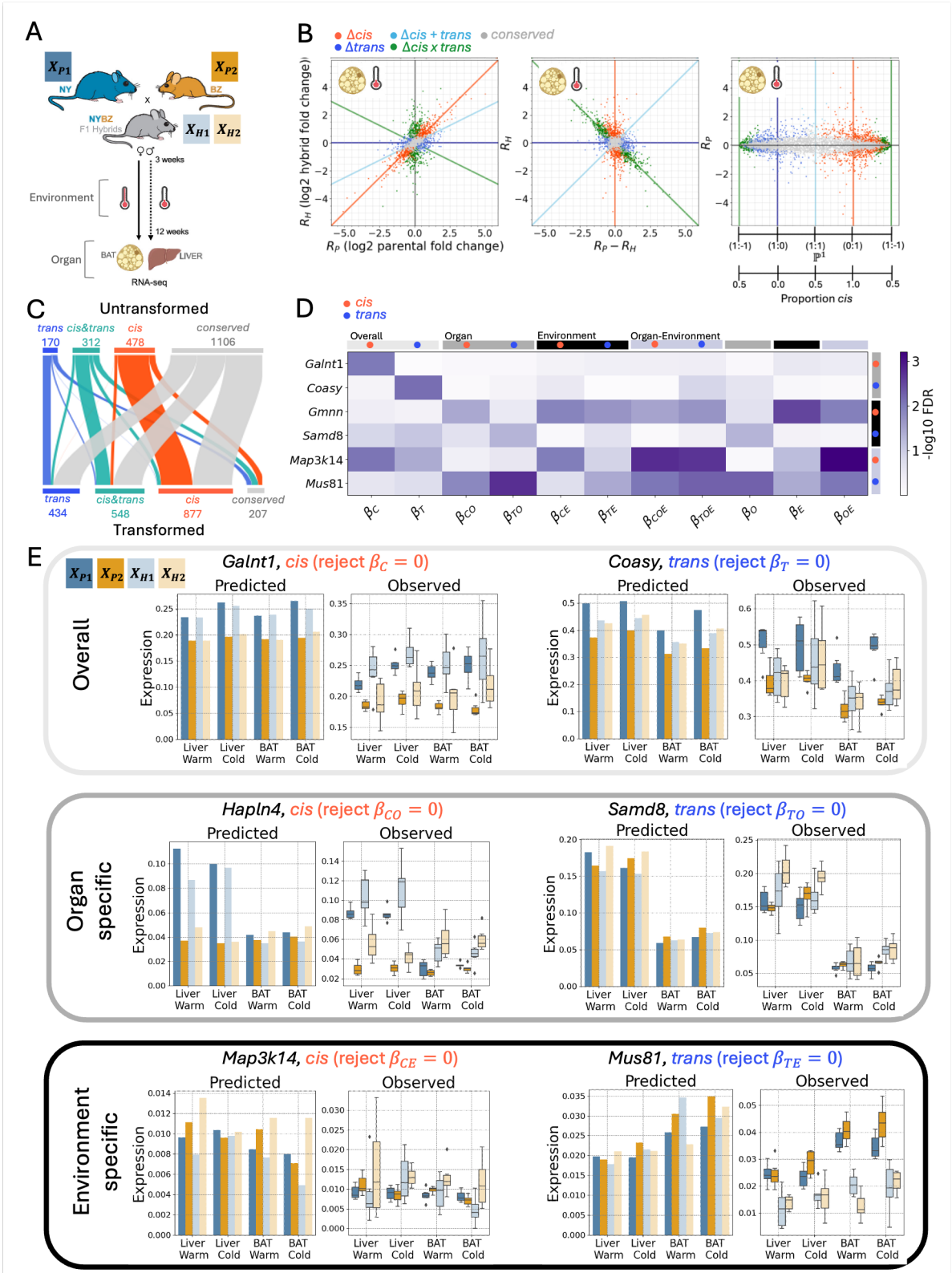

Fig. 3. Multi-sample regulatory assignment can be determined using generalized linear models and condition-specific hypothesis tests. A) Cartoon of experimental design set up from original study (adapted from Fig. 1C of (13)). B) Model predicted ratios colored by GLM determined regulatory assignments as described in Methods. The log additive model was fit to brown adipose tissue from male mice reared in the cold (n = 6 NY mice, n = 6 BZ mice, and n = 6 hybrid crosses, and n = 5,970 genes). C) Untransformed (original) regulatory assignments (reported in (13)) derived from three independent model fits and statistical tests differ from transformed regulatory assignments derived using the unified GLM framework (genes and assignments from B, excepting genes that were categorized as conserved by both the original and our testing procedure, which are not shown). D) The log additive model was fit to samples from all four environment/organ pairs (BAT/cold, BAT/warm, liver/cold, liver/warm, n = 6 NY/BZ/hybrid samples per condition) with weights for environment and organ-specific regulatory effects. The negative log10 of false discovery rates (FDR) highlights several genes that exhibit significant *cis* or *trans* behavior overall or only in an organ or environment-specific manner. E) For the same genes, GLM predicted (bar plots) and observed gene expression (box plots with counts normalized by the total count per sample for parents and per haplotype for hybrids) for parental and hybrid alleles (n = 6 samples per box, with the box showing values' interquartile ranges and a line at the mean and whiskers extending to min/max values).

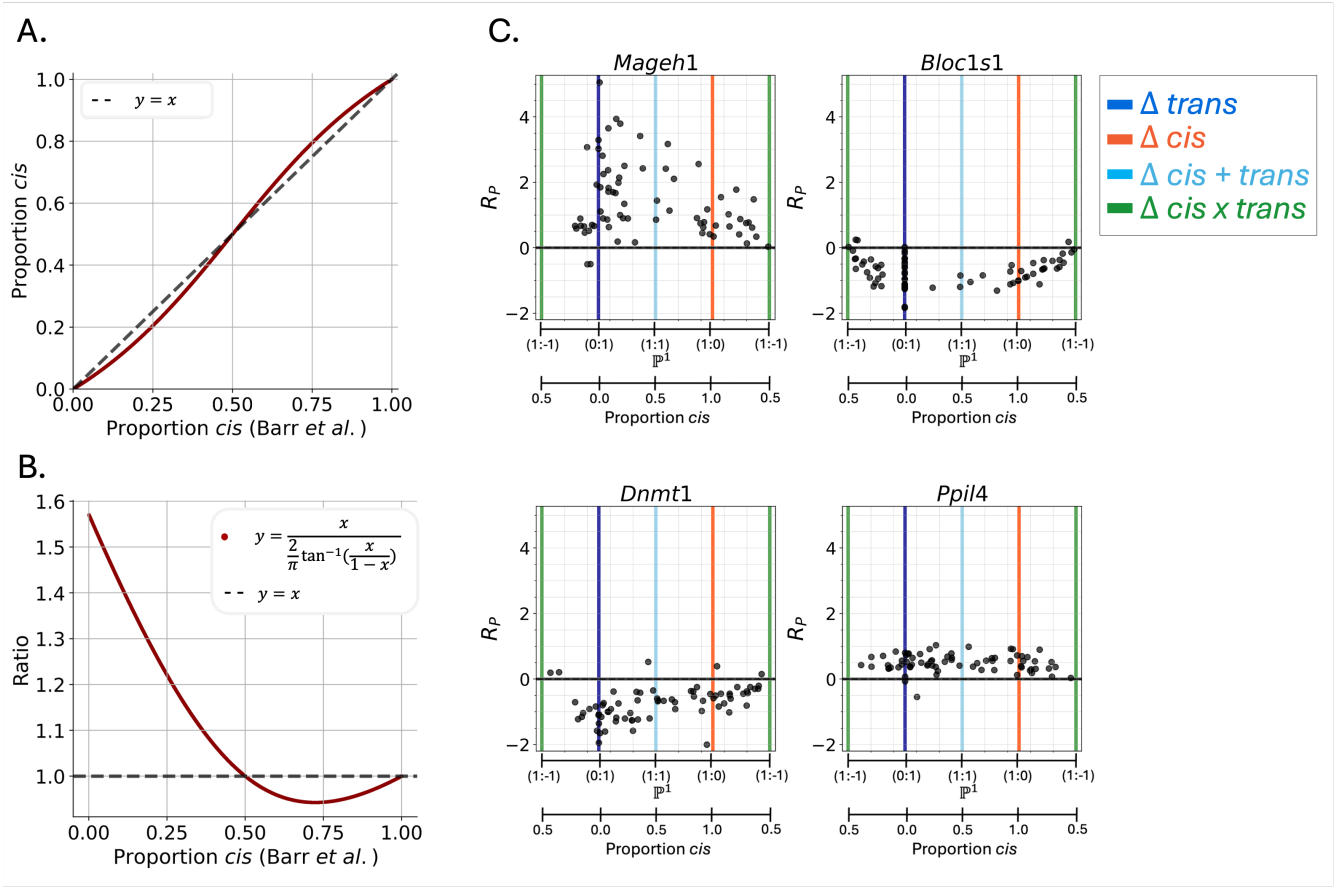

Fig. 4. In addition to broad regulatory categories, genes can display subtle differences in their proportion of *cis* regulation. A) Comparison of slope (as previously defined (17), x-axis) and angle (y-axis) for determining proportion *cis*. B) Ratio of slope and angle determination of proportion *cis* (dark red line, dashed line displays equality). Note that at small values of proportion *cis*, the slope can be over 50% the determined angle. C) Gene expression data from (17) (human-chimpanzee crossed cell lines) was used to determine cell type specific proportion *cis*. These four genes display high variance in proportion *cis* across cell types (black points are determined proportion *cis* in n = 72 cell types).

single-cell RNA-seq assays can complement genome-wide eQTL studies.

We have shown that geometric considerations reveal the need for applying a linear transformation prior to visualization. The linear transformation highlights independent axes that lead naturally to hypothesis tests for classifying genes according to the type of regulation underlying differences in gene expression between parents. While we have focused on explaining differences between parents that fall into five categories (conserved, *cis*, *trans*, *cis* + *trans*, *cis* × *trans*), our approach can be extended to finer classifications such as in (9). We note that in the hypothesis testing framework, referring to a gene as having differences in expression explained by *cis* or *trans* is technically incorrect. This is because the rejection of the *cis* null hypothesis only shows that the difference in gene expression is not due solely to *cis* regulation. This does not mean that the difference in gene expression can or should be attributed solely to *trans*. The same is the case for rejecting the *trans* hypothesis. In Fig. 2, our coloring of genes as *cis* or *trans* is therefore not precise, but we have done so to facilitate comparisons to previous work.

Our single sample statistical tests depend on the assumption that read counts are binomially distributed, which is standard in the absence of biological replicates. However, when replicate samples are available, we can better assess the extent of technical variation and use generalized linear models. Although our results are shown using commonly applied estimation techniques for negative binomially distributed counts (11), we also measure the extent to which differing estimates of variation change the results (Supplement) (11, 12). We encourage experimentalists to include enough samples in the experimental design such that heuristic estimations are unnecessary and the observed sample variation can be used for hypothesis testing.

Finally, we note that our framework is general and can be applied to more complex experimental designs, regulatory hypotheses, and phenotypes other than gene expression. For example, the versatility of our GLM framework allows modeling of *cis* or *trans* differences in specific conditions and in interactions between specific conditions. In terms of additional phenotypes, with single-cell RNA sequence data our approach could be used in conjunction with methods such as (18, 19) to assess the regulation mechanisms underlying differences in biophysical aspects of gene transcription, splicing and degradation. Such extensions will be particularly interesting to explore in conjunction with complementary modalities (20, 21).

Data and Code Availability

All code to download data and generate the main and supplementary figures is available at https://github.com/pachterlab/HCP_2024, and an R package (*XgeneR*) to implement the multi-sample tests at .

Acknowledgements

IH and LP were funded, in part, by NIH 5UM1HG012077-02. MC was funded by a NSF graduate research fellowship under Grant No. 2139433. This work was additionally supported by the Caltech Bioinformatics Resource Center.

References

1. François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.
2. Rita JS Phillips. A cis-trans position effect at the a locus of the house mouse. *Genetics*, 54 (2):485, 1966.
3. John S Kovach, Antonio O Ballesteros, Marilyn Meyers, Marco Soria, and Robert F Goldberger. A cis/trans test of the effect of the first enzyme for histidine biosynthesis on regulation of the histidine operon. *Journal of Bacteriology*, 114(1):351–356, 1973.
4. Christopher R Cowles, Joel N Hirschhorn, David Altshuler, and Eric S Lander. Detection of regulatory variation in mouse genes. *Nature genetics*, 32(3):432–437, 2002.
5. Patricia J Wittkopp, Belinda K Haerum, and Andrew G Clark. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88, 2004.
6. Christian R Landry, Patricia J Wittkopp, Clifford H Taubes, Jose M Ranz, Andrew G Clark, and Daniel L Hartl. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of drosophila. *Genetics*, 171(4):1813–1822, 2005.
7. Angela Goncalves, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro, Paul Flicek, Alvis Brazma, Duncan T Odom, and John C Marioni. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome research*, 22(12): 2376–2384, 2012.
8. Andreas Massouras, Sebastian M Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F Ayroles, Emmanouil T Dermitzakis, Eric A Stone, Jeffrey D Jensen, Trudy FC Mackay, et al. Genomic variation and its impact on gene expression in drosophila melanogaster. *PLoS genetics*, 8(11):e1003055, 2012.
9. Andreas Tsouris, Gauthier Brach, Joseph Schacherer, and Jing Hou. Non-additive genetic components contribute significantly to population-wide gene expression variation. *Cell Genomics*, 4(1), 2024.
10. C Joel McManus, Joseph D Coolon, Michael O Duff, Jodi Eipper-Mains, Brenton R Graveley, and Patricia J Wittkopp. Regulatory divergence in drosophila revealed by mrna-seq. *Genome research*, 20(6):816–825, 2010.
11. 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. doi: 10.1093/bioinformatics/btp616.
12. 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, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.
13. Mallory A. Ballinger, Katya L. Mack, Sylvia M. Durkin, Eric A. Riddell, and Michael W. Nachman. Environmentally robust cis-regulatory changes underlie rapid climatic adaptation. *Proceedings of the National Academy of Sciences*, 120(39):e2214614120, 2023. doi: 10.1073/pnas.2214614120.
14. Hye Won Kang, Cafer Ozdemir, Yuki Kawano, Katherine B. LeClair, Cecile Vernochet, C. Ronald Kahn, Susan J. Hagen, and David E. Cohen. Thioesterase superfamily member 2/acyl-coa thioesterase 13 (them2/acot13) regulates adaptive thermogenesis in mice. *The Journal of Biological Chemistry*, 288(46):33376–33386, November 2013. ISSN 0021-9258. doi: 10.1074/jbc.M113.481408.
15. Sabrina Gohlke, Vyacheslav Zagoriy, Alvaro Cuadros Inostroza, Michaël Méret, Carola Mancini, Lukasz Japtok, Fabian Schumacher, Doreen Kuhlow, Antonia Graja, Heike Stephanowitz, Markus Jähnert, Eberhard Krause, Andreas Wernitz, Klaus-Jürgen Petzke, Annette Schürmann, Burkhard Kleuser, and Tim J. Schulz. Identification of functional lipid metabolism biomarkers of brown adipose tissue aging. *Molecular Metabolism*, 24:1–17, April 2019. ISSN 2212-8778. doi: 10.1016/j.molmet.2019.03.011.
16. Demet Tekin, Lei Xi, Tingcun Zhao, M. Isabel Tejero-Taldo, Sumanth Atluri, and Rakesh C. Kukreja. Mitogen-activated protein kinases mediate heat shock-induced delayed protection in mouse heart. *American Journal of Physiology - Heart and Circulatory Physiology*, 281 (2):H523–H532, August 2001. ISSN 0363-6135. doi: 10.1152/ajpheart.2001.281.2.H523.
17. Kenneth A Barr, Katherine L Rhodes, and Yoav Gilad. The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. *Genome Biology*, 24(1):207, 2023.
18. Gennady Gorin, John J Vastola, and Lior Pachter. Studying stochastic systems biology of the cell with single-cell genomics data. *Cell Systems*, 14(10):822–843, 2023.
19. Maria Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. Biophysical modeling with variational autoencoders for bimodal, single-cell rna sequencing data. *Nature Methods*, 21(8):1466–1469, 2024.
20. Ibai Irastorza-Azcarate, Alexander Kukalev, Rieke Kempfer, Christoph J Thieme, Guido Mastrobuoni, Julia Markowski, Gesa Loof, Thomas M Sparks, Emily Brookes, Kedar Nath Natarajan, et al. Extensive folding variability between homologous chromosomes in mammalian cells. *bioRxiv*, pages 2024–05, 2024.
21. Zhenzhen Ma, Alexander Lorenzo Starr, David Gokhman, and Hunter Fraser. The causes and consequences of human-specific dna methylation. *bioRxiv*, pages 2026–01, 2026.
22. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
23. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Methods

Acquisition and preprocessing of data from yeast crosses. Processed RNA-seq data from (9) were obtained from http://1002genomes.u-strasbg.fr/files/Diallel_RNAseq/ASE in the file 'Datafile2_ase_sum_20230609.tab.' This file included reported values for allelic expression in parents and expression per parental allele in hybrids for 179 unique parent-hybrid trios and 285,777 unique gene-trio combinations. Details on the estimation of allelic expression are described in the original study (9). As integer value counts are required for the binomial and the two sample binomial ratio statistical tests, reported values were rounded to the nearest integer before statistical tests were performed.

In (9), 7 categories were considered (*reverse, attenuating, reinforcing, compensatory, cis only, trans only* and *null*). In order to maintain consistency with (6), we collapsed these into four categories: *cis, trans, cis + trans*, and *cis × trans* as follows: reverse → *cis × trans*, *cis only* → *cis*, *trans only* → *trans*, null → conserved, and attenuating, reinforcing and compensatory → *cis × trans* if $\text{sgn}(R_P - R_H) = \text{sgn}(R_H)$ and *cis + trans* otherwise.

Acquisition of data from human-chimpanzee crosses. The reported log2 fold changes between human and chimpanzee parental gene expression (R_P) and hybrid allelic log2 fold changes for human-chimpanzee hybrid cell lines (R_H) were obtained from (17) for 72 cell types and 14,487 genes per cell type. The reported log2 fold changes were used without modification to recalculate proportion *cis* as

$$\text{proportion } cis = \frac{2}{\pi}\left|\text{atan2}(R_H,\, R_P - R_H)\right|. \tag{4}$$

The provided data also included reported proportion *cis* per gene per cell type, defined as

$$\text{proportion } cis \text{ (17)} = \frac{|R_H|}{|R_H| + |R_P - R_H|}, \tag{5}$$

which were directly compared to Equation 4.

The maximum difference between $x := \frac{|R_H|}{|R_H|+|R_P-R_H|}$ and $\frac{2}{\pi}\left|\text{atan2}(R_H,\, R_P - R_H)\right|$ is 0.045 at $x = 0.239$ and $x = 0.761$. The maximum of the ratio is 1.57 and occurs at $x \approx 0.00006$ (Fig. 3B).

Acquisition of data from mouse crosses. (13) performed a study to assess regulatory strategies driving environmental adaptations in house mouse from two different climates. To do so, they created two inbred lines derived from wild house mice from Saratoga Springs, New York, USA (NY, cold climate) and Manaus, Amazonas, Brazil (BZ, warm climate) and crossed them. Mice from both inbred lines and their hybrid crosses were housed in warm (21 °C) and cold (5 °C) conditions, then liver and brown adipose tissue (BAT) samples from male and female mice were collected. RNA sequencing, alignment, and allele-specific expression (ASE) quantification for parental and hybrid tissue samples were performed as described in the original publication ((13)).

To assign genes' regulatory categories, they performed three statistical tests on the parental counts and ASE in the hybrids (using *DESeq2* (12)): first, they tested for differential expression between lines using only parental counts; second, they tested for differential expression between alleles in the hybrids using only hybrid ASE; and third, they tested for ratio differences between alleles in the parents and alleles in the hybrids using parental counts and ASE (details in the Methods of (13)). They assigned genes as *cis, trans, cis & trans* or *conserved or ambiguous* based on the outcomes of the three tests, as described in (13). For consistent comparison, we collapse the categories *cis + trans* and *cis x trans* to *cis & trans*.

We obtained parental counts, regulatory assignments for genes in the four separate environment (warm/cold) and organ (liver/BAT) combinations, and sample metadata from the GitHub repository `https://github.com/malballinger/BallingerMack_PNAS_2023/`. We obtained ASE for hybrids from the authors directly. We applied our regulatory assignment framework (see the subsection ***Generalized linear models for multiple samples***) to male mice (for each of the four environment-organ combination there were 6 SARA mice, 6 MANA mice, and 6 hybrids, for a total of 18 mice and 24 rows in the design matrix, as each hybrid contributes expression levels for two alleles) for each environment/organ condition separately (liver/warm, liver/cold, BAT/warm, BAT/cold); and for all male samples together (total of 72 mice and 96 rows of the design matrix). We used the log-additive *trans* hypothesis model for testing all samples and condition-specific effects and for (liver/warm, liver/cold, and BAT/warm) male samples. For all warm samples, there were 5,897 fit genes, and 5,970 for cold samples. We fit the log additive, dominant, and free models to the BAT/cold male samples, with comparisons of fit weights, significance values, and gene regulatory assignments shown in Supp. Figs. 4 – 6. All fits were implemented using the XgeneR package with edgeR version 4.0.16 (11).

Statistical tests.

Binomial test. To test the null hypothesis that there is no difference between expression of parental alleles in hybrids (or that the difference in regulation is purely *trans*) was performed on rounded integer counts from (9) using the function

`scipy.stats.binomtest` (22). This tests for the probability of having observed a value at least as extreme as k successes given probability p of success and a total of N trials by summing over binomial probabilities:

$$P(k;N,p) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{6}$$

If X_{H1} is the allelic expression in the hybrid of one parental allele and X_{H2} is the allelic expression in the hybrid of the other parental allele, the binomial test was performed with $k = X_{H1}, N = X_{H1} + X_{H2}$, $p = 0.5$ and a two-sided alternative hypothesis.

Two sample binomial ratio test. To test the null hypothesis that there is no difference in the ratio of expression of alleles in parents and allelic expression in hybrids (or that the difference in regulation between parents is purely cis), we performed a two sample binomial ratio test in which we assumed both X_{P1} and X_{H1} are sampled from a binomial distribution with the same probability of success p_s. Using

$$p_s = \frac{X_{P1} + X_{H1}}{X_{P1} + X_{P2} + X_{H1} + X_{H2}}, \tag{7}$$

we calculated the grid of probabilities over possible P_1 and H_1 values:

$$P(X_{P1} = x_{P1}, X_{H1} = x_{H1}; N_P, N_H, p_s) = \binom{N_P}{x_{P1}} p_s^{x_{P1}} (1-p_s)^{N_P - x_{P1}} \cdot \binom{N_H}{x_{H1}} p_s^{x_{H1}} (1-p_s)^{N_H - x_{H1}}, \tag{8}$$

where $N_P = X_{P1} + X_{P2}$ is the total number of counts from parents and $N_H = X_{H1} + X_{H2}$ is the total number of counts from the hybrid. We then calculated the probability of having observed values at least as extreme as the observed X_{P1} and X_{H1}.

To account for multiple testing, we corrected both binomial and two sample binomial significance values using the Benjamini-Hochberg correction for false discovery rates (23). We rejected the null hypothesis for tests with false discovery rates less than 0.05.

Generalized linear models for multiple samples. We propose a generalized linear model (GLM) framework for testing the null hypotheses of *no cis* and *no trans* regulation, both overall and condition-specific, when multiple samples per condition are available. Our framework differs from previous applications of GLMs for *cis/trans* distinction (13) by virtue of using one GLM for all data (parental counts, hybrid ASE values, with all conditions represented), rather than fitting three separate models to different subsets of the data (e.g., only parental counts for one test or only hybrid ASE values for another). We do this by developing a model with weights for overall and condition-specific *cis* and *trans* regulatory effects, as well as including condition intercepts to properly attribute observed allele expression patterns to regulation or condition. In addition, our approach allows us to propose different hypotheses about the nature of the *trans* regulatory effect: if it is log-additive, dominant, or free (described below). To properly account for the overdispersion and discreteness of RNA-seq counts, we fit a GLM with a negative binomial likelihood function and log link function.

One condition: In the simplest case, we set expression of the first parental allele undergoing only *cis* regulation to be the intercept β and include weights for the *cis* regulatory difference (β_C) and *trans* regulatory differences (β_T) between parental strains. This allows the normalized expression from the parental alleles (X_{P1}, X_{P2}) and normalized allelic expression from the hybrids (X_{H1}, X_{H2}) to be modeled under three different hypotheses of *trans* regulation.

The first hypothesis (**log-additive**) assumes that the *trans* expression difference between parental strains at each allele is influenced by both sets of parental chromosomes multiplicatively, and is thus reduced to the square root of the full parental effect in the hybrids. The design matrix is accordingly

$$\log \mathbb{E} \begin{bmatrix} X_{P1} \\ X_{P2} \\ X_{H1} \\ X_{H2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0.5 \\ 1 & 1 & 0.5 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_C \\ \beta_T \end{bmatrix}. \tag{9}$$

A second hypothesis (**dominant**) assumes that the *trans* expression difference between parental strains is the same regardless of if both sets or one set of parental chromosomes are present: it is the same in the parents and hybrids, and the design matrix is

$$\log \mathbb{E} \begin{bmatrix} X_{P1} \\ X_{P2} \\ X_{H1} \\ X_{H2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_C \\ \beta_T \end{bmatrix}. \tag{10}$$

Finally, and most analogous to previous hypothesis testing methods (13), *trans* expression differences between parental strains can have an unconstrained relationship (**free**) with their effects in the hybrids by allowing each hybrid to have individual specific effects (denoted below with the subscript i indexing the hybrid samples). The resulting design matrix is

$$\log \mathbb{E} \begin{bmatrix} X_{P1} \\ X_{P2} \\ X_{H1_i} \\ X_{H2_i} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_C \\ \beta_T \\ \beta_{H_i} \end{bmatrix}. \tag{11}$$

For all models, we can test the null hypothesis of *no cis* differences between parental strains by testing if $\beta_C = 0$ and the null hypothesis of *no trans* differences by testing if $\beta_T = 0$.

Multiple conditions: As more conditions are added to the experimental design, they can be naturally incorporated into the GLM framework, with condition-specific *cis* and *trans* weights as well as intercepts and interaction terms. For example, we fit data for liver and BAT samples for mice grown in warm and cold environments (four separate organ/environment) using the following model under the log-additive *trans* hypothesis:

$$\begin{aligned} \log \mathbb{E}[X] \sim \beta + \beta_C + \beta_T + \beta_{CO} + \beta_{TO} + \\ + \beta_{CE} + \beta_{TE} + \beta_{COE} + \beta_{TOE} + \beta_O + \beta_E + \beta_{OE}. \end{aligned} \tag{12}$$

The subscripts C, T, O and E refer respectively to differential *cis* regulation between strains, differential *trans* regulation between strains, organ, and environment effects, with combinations denoting their interactions (e.g., β_{CO} models the organ-specific *cis* effect). This is also compatible with any of the three previously described *trans* hypotheses (log-additive, dominant, and free).

Implementations: We implemented GLMs for the one condition in edgeR version 4.0.16 (11) and DESeq2 version 1.42.1 (12), showing differences in estimated gene variance, data likelihoods and hypothesis test significance values in Supp. Fig. 12. In particular, we tested significance using a likelihood ratio test (LRT) between the full and reduced models. Differences in significance values lead to differences in regulatory assignments, especially pertinent in this our case in which both *no cis* and *no trans* hypothesis testing results are important for regulatory categorization. As edgeR is the more conservative method, estimating higher variance and false discovery rates (Supp. Fig. 12), we used it for full model fits and regulatory assignments. We corrected LRT p-values for each hypothesis test for the number of genes tested using the Benjamini-Hochberg correction (23) to obtain false discovery rates, and rejected null hypotheses with false discovery rates less than 0.05.

Classifying genes based on statistical test results. The results of the two tests for single sample experiments or experiments with replicates are used to categorize genes as outlined in Table 1.

Single sample: Binomial test Replicates: $\beta_C = 0$ **(null: *no cis*)**	Single sample: Binomial ratio test Replicates: $\beta_T = 0$ **(null: *no trans*)**	**Classification**
Fail to reject	Fail to reject	conserved
Reject	Fail to reject	*cis*
Fail to reject	Reject	*trans*
Reject	Reject	*cis* & *trans* (*cis* + *trans* or *cis* × *trans* by geometry)

Table 1. Statistical assignments of gene regulatory differences. Genes assigned to *cis, trans, cis + trans* or *cis* × *trans* regulation based on two statistical tests, described above for single sample or replicate experiments. If both statistical tests are rejected, regulatory assignment can be made by locating the point $(R_P - R_H, R_H)$ (see Results Section, Geometry) in the transformed space and following the geometric assignments listed in Supp. Table 1.

Note that some care must be taken when interpreting a *cis* or *trans* assignment within this hypothesis testing framework. For example, a *cis* assignment is made when "no *cis*" can be rejected but "no *trans*" cannot be rejected. Failure to reject "no *trans*" does not imply that one can accept the null hypothesis, but we utilize the assignment of *cis* in this case for simplicity for users. However, this caveat should be taken into account when interpreting the assignments of *cis*, *trans*, and *cis* & *trans* assignments.