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.

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 to 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 $R_P = log\left(\frac{X_{P1}}{X_{P2}}\right)$ and $R_H = log\left(\frac{X_{H1}}{X_{H2}}\right)$. That is, R_P corresponds to the log fold-change difference in expression between the two parental strains, and R_H to the log-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 regulations 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 $R_P - R_H = 0$, which implies $R_P = R_H$, 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 $R_P - R_H \neq 0$, which implies $R_P \neq 0$, arises from a change only in *trans*-regulatory elements.

- ***cis + trans:*** $R_H \neq 0$ and $R_P - R_H \neq 0$ 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 $R_P - R_H \neq 0$ 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 fail to properly assign regulation differences to *cis* + *trans* when both $R_H < 0$ and $R_P < 0$. This issue, and 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 gene expression between the parents due to 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 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.

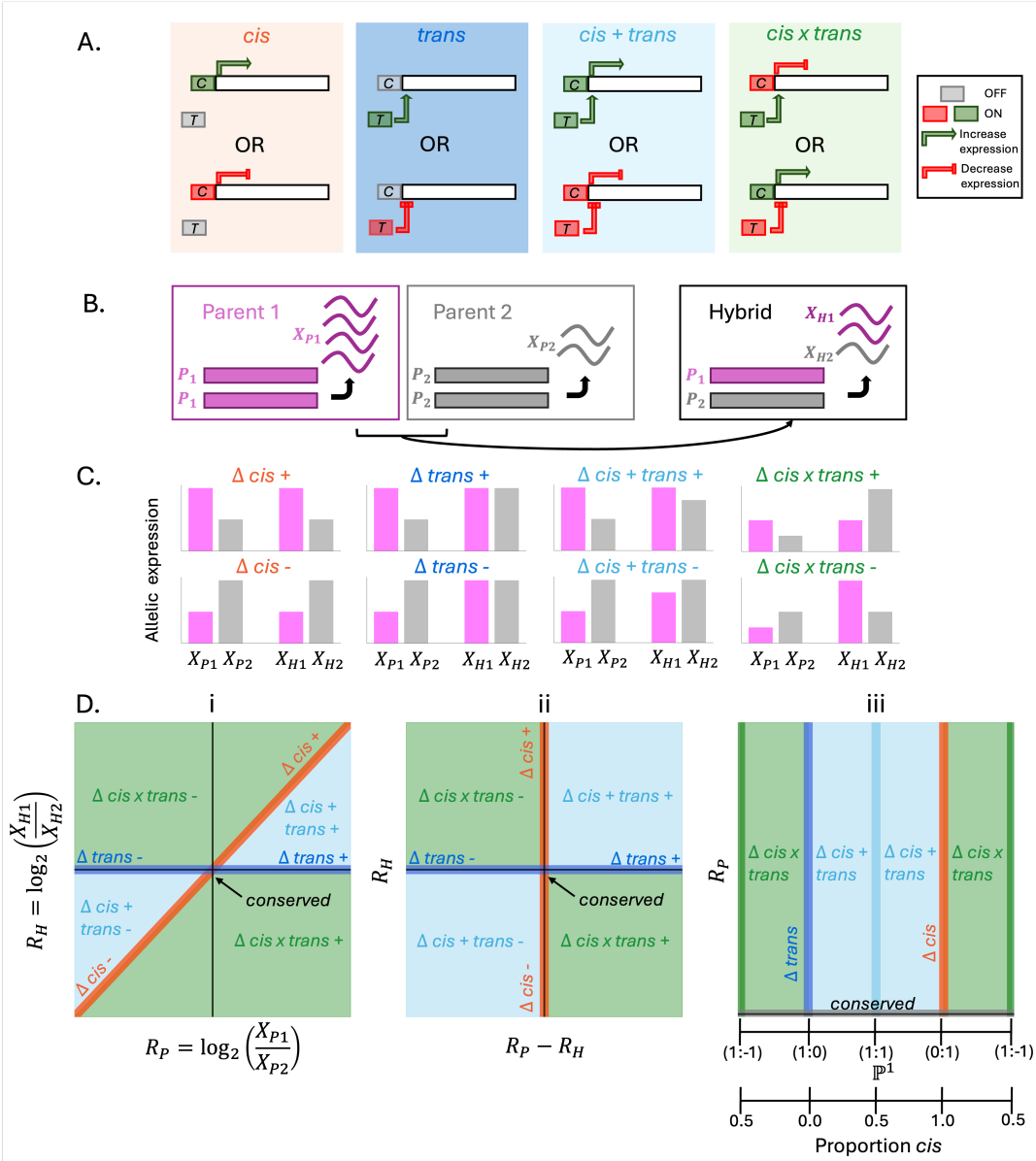

Fig. 1. Geometry of parental and hybrid expression ratios. A) The *cis*, *trans*, *cis + trans*, and *cis* × *trans* types of regulation. B) Homozygous parents with haplotypes P_1 and P_2, counts of RNA molecules for a genes (X_{P1} and X_{P2} respectively), and the haplotypes in an F_1 hybrid along with counts for the gene (X_{H1} and X_{H2}). C) Differences in regulation between the parents is 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 $\mathbb{P}^1$.

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

$$\begin{pmatrix} R_P \\ R_H \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} R_P - R_H \\ R_H \end{pmatrix}. \quad \textbf{(1)}$$

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

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad \textbf{(2)}$$

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 $(R_P, R_P - R_H)$ is a perturbation (due to noise) 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

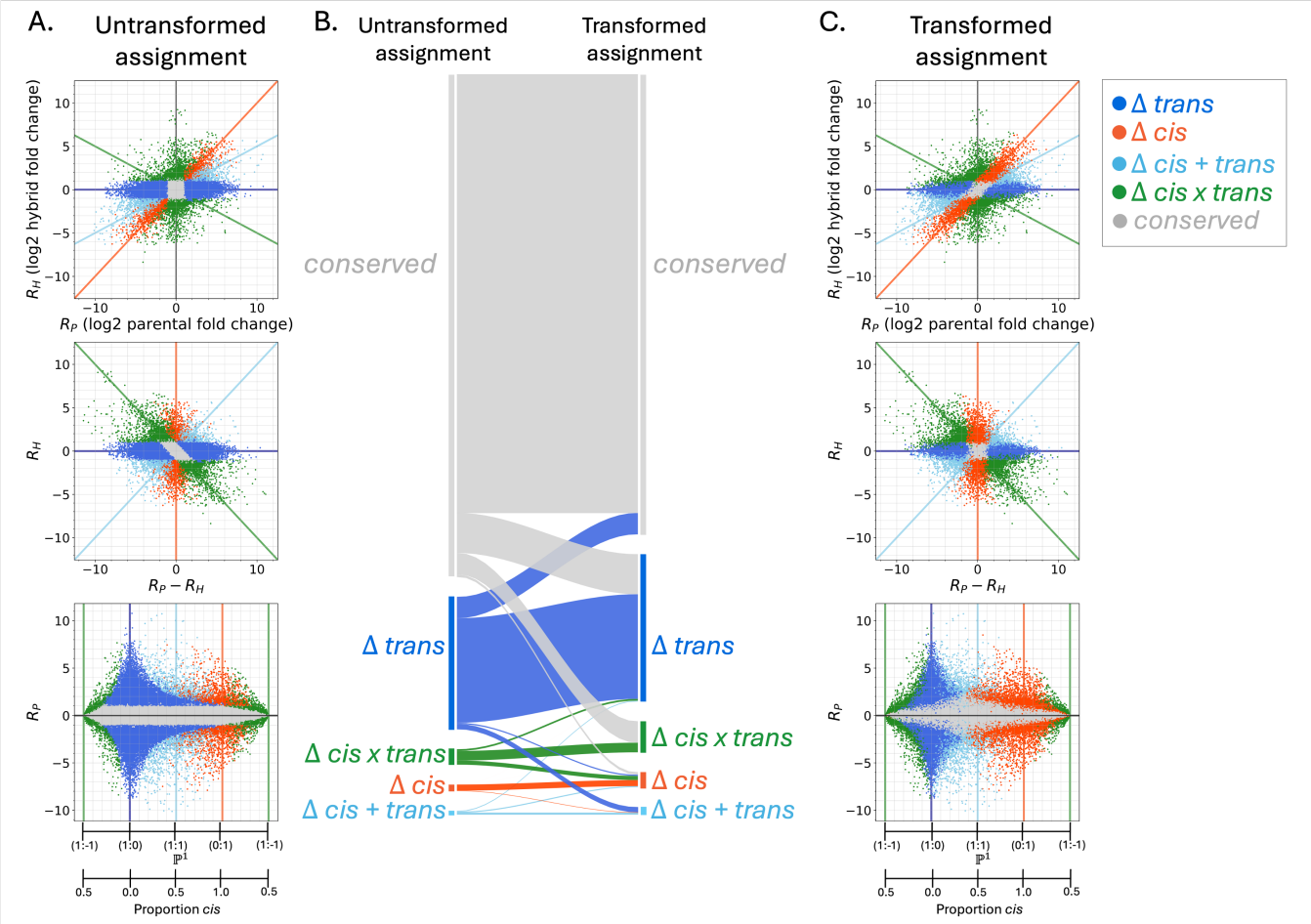

Fig. 2. Contributions of *cis* and *trans* regulation to differences in gene expression between yeast strains. A) Results of (9) B) Comparison of the results of (9) to the assignment of genes based on hypothesis testing, C) Reanalysis of the data from (9).

cis + *trans* or Δ *cis* $\times$ *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 correspond to the point in $\mathbb{P}^1$, i.e.

$$\text{proportion } cis = \left| \frac{2}{\pi} \tan^{-1}\left(\frac{R_H}{R_P - R_H} \right) \right|. \qquad \textbf{(3)}$$

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 $(R_P, R_P - R_H)$ is located on the x- and y- axes respectively.

We performed such hypothesis tests for data from (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 323 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; then, changes between parental and hybrid allelic ratios were tested for significance. 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 coordinate 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. 3A,3B, Supp. Table 2).

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

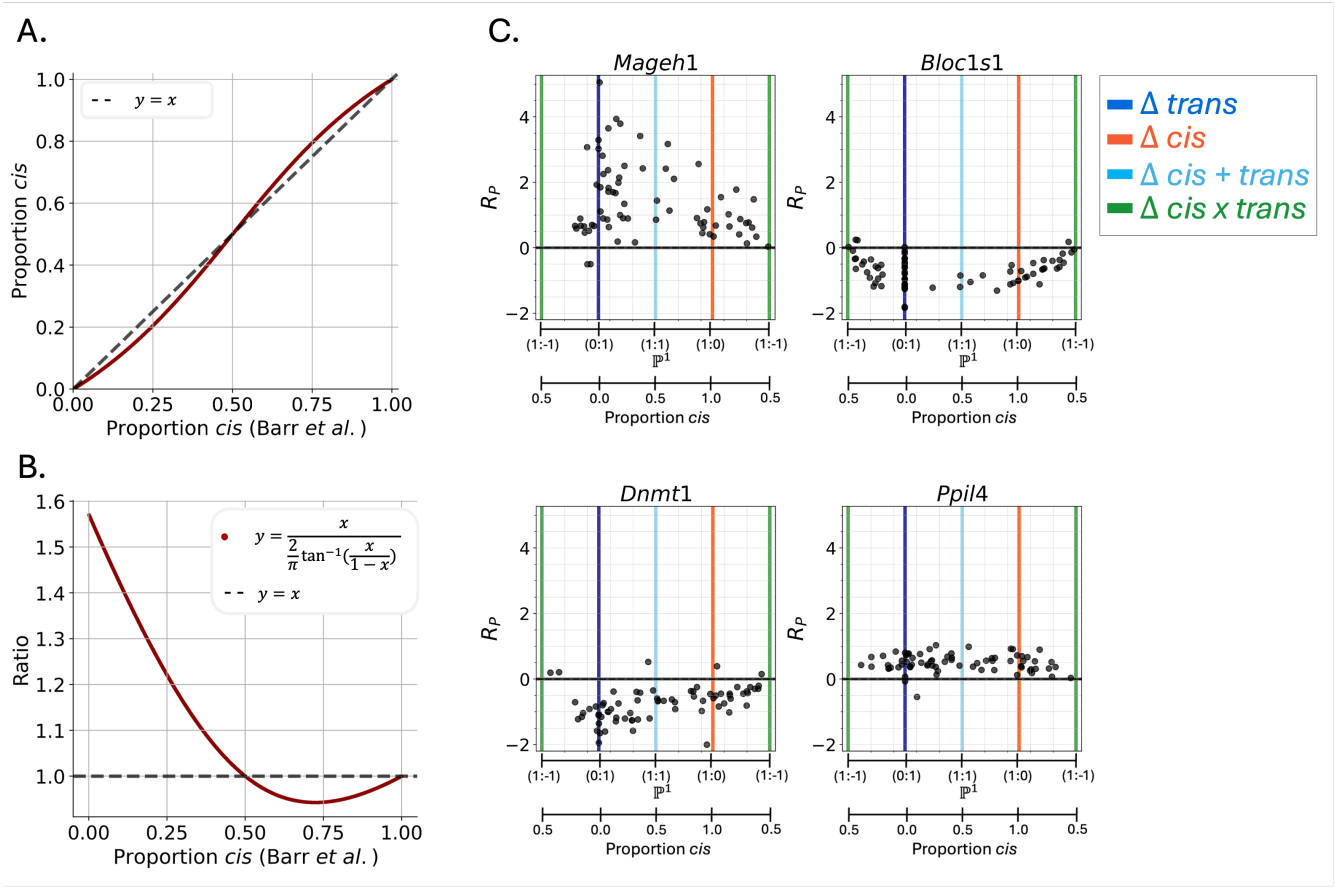

Fig. 3. Proportion *cis*. A) Comparison of slope vs. angle in determining proportion *cis*. B) Ratio between slope and angle. C) Examples of four genes displaying high variance in proportion *cis* across cell types.

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).

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*, as previously discussed 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 examined and re-analyzed a dataset of gene expression from human-chimpanzee hybrids (11), calculating the proportion *cis* according to Equation 3. In (11), 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. 3A), with a maximum difference of 0.045 (see Methods), the relative difference is large when the proportion *cis* is small (Fig. 3B), 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 (11) exhibiting high variance in proportion *cis* (Fig. 3C), and identified interesting differences between cell types (Supp. Fig. 4 – 7).

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, whereas original studies were limited by gene expression measurement technologies to a handful of genes, genome-wide 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 visualisation and data analysis. The linear transformation also 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. It 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 prior work.

Our statistical tests also depend on the assumption that read counts are binomially distributed. This assumption is standard in the absence of biological replicates, however in practice biological replicates should be obtained, and they can and should be used to better assess the extent of technical variation. An extension of this work to that case is possible and can be based on modeling counts with negative binomial distributions, as is done in methods such as WASP (12).

Finally, we note that our framework is general and can be applied to phenotypes other than gene expression. For example, with single-cell RNA-seq data it could be used to assess the regulation mechanisms underlying differences in variance in gene expression. Such extensions will be particularly interesting to explore in conjunction with complementary modalities (13).

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, executable in Google Colaboratory notebooks.

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. 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.
12. Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12 (11):1061–1063, 2015.
13. 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.
14. 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.
15. 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 323 unique parent-hybrid trios. 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* $\times$ *trans* as follows: reverse $\rightarrow$ *cis* $\times$ *trans*, *cis only* $\rightarrow$ *cis*, *trans only* $\rightarrow$ *trans*, null $\rightarrow$ conserved, and attenuating, reinforcing and compensatory $\rightarrow$ *cis* $\times$ *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 cell lines (R_P) and hybrid allelic $log2$ fold changes for human-chimpanzee hybrid cell lines (R_H) were obtained from (11). The reported $log2$ fold changes were used without modification to recalculate proportion *cis* as

$$\text{proportion } cis = \left| \frac{2}{\pi} \tan^{-1}\left(\frac{R_H}{R_P - R_H}\right) \right|. \tag{4}$$

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

$$\text{proportion } cis \text{ (11)} = \frac{|R_H|}{|R_H| + |R_P - R_H|}, \tag{5}$$

which were directly compared to Equation 4.

The maximum difference between $x := \frac{|R_H|}{|R_H|+|R_P-R_H|}$ and $\frac{2}{\pi}\tan^{-1}\left(\frac{R_H}{R_P-R_H}\right)$ is 0.045 at $x = 0.239$ and $x = 0.761$. The maximum of the ratio is 1.57 and occurs at $x \approx 0.00006$.

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` (14). 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 (15). We rejected the null hypothesis for tests with false discovery rates less than 0.05.