Genomic Studies of Gene Expression Errors and Their Evolutionary Ramifications

by

Mengyi Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2021

Doctoral Committee:
Professor Jianzhi Zhang, Chair
Professor Timothy James
Professor Alexey Kondrashov
Associate Professor Xiaoquan Wen
Professor Patricia Wittkopp

Mengyi Sun

mengysun@umich.edu

ORCID iD: 0000-0003-0842-8098

DEDICATION

To the endless forms most beautiful

ACKNOWLEDGEMENTS

In retrospect, I have to say that mentoring me, a student who was not quite well-polished, must have been very challenging and maybe annoying. So first and foremost, I would like to say thank you to my mentor, Professor George Zhang. George's great support and mentorship have enabled me to find my authentic niche in scientific research. His patience, calmness, and kindness have shown me a path to be a better person. But perhaps most importantly, he has taught me how to disagree respectfully.

I would like to thank Professor Misha Teplitskiy. It has been a pleasure to collaborate with you, and thank you very much for your help during my difficult moments of switching to a new research direction.

I would like to thank my committee members: Professor Patricia Wittkopp, Professor Timothy James, Professor Alexey Kondrashov, and Professor Xiaoquan Wen, for their critical and insightful feedback on my dissertation.

I would also like to thank my undergraduate mentors: Professor Xionglei He and Professor Yongjun Lu. They taught me the fundamentals of biology and helped me apply for graduate school.

I am indebted to all current members of the Zhang lab. Especially, I would like to express my gratitude to the experimental group with who I interacted most frequently (despite being a computational guy): Xukang Shen, Haiqing Xu, Piaopiao Chen, and Haoxuan Liu, for their invaluable advice in both work and life.

During my Ph.D., I have been very lucky to interact with many former members of the Zhang lab. Particularly, I would like to express my sincere appreciation to Wenfeng Qian,

Jianrong Yang, Xinzhu Wei, Nagarjun Vijay, Chuan Li, and Wei-Chin Ho, for those interesting and intellectually stimulating moments.

Although by no means painless, my Ph.D. has been a great journey. It would not be so without all my friends, in Ann Arbor and from all over the world. I can’t list them all but a few, in no particular order: Shang Zhang, Jia Li, Jiawen Zhang, Muyang Lv, Shuting Wu, Yuchen Zhao, Xiang Ji, Chao Shi, Jianliang Wu, Zilong Liang, Siliang Song, Erping Long and the aforementioned current and former Zhang lab members. I am most grateful for your tolerance to my impoliteness once upon a time. I wish I could have done better.

In the end, I would like to thank my parents for their unconditional love and support throughout these years. I am extremely fortunate to have parents who would buy me books without hesitation whenever I was eager to read---I could never forget the sparkling moment when I opened *"The Feynman’s Lectures on Physics"*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Gene expression produces biologically functional RNAs and proteins and is essential for life. Nevertheless, gene expression is subject to several types of errors that are generally harmful. Despite the prevalence and significant consequences of expression errors, their genome-wide patterns are not well characterized. Furthermore, the evolutionary ramifications of such errors are poorly understood. In my dissertation, I address the above questions using novel computational approaches. I focus on two types of gene expression errors: (i) stochastic gene expression, which leads to a variation of the expression level among isogenic cells in the same environment (gene expression noise), and (ii) mistranslation, which induces protein misfolding and can be toxic to the cells.

My thesis has three main chapters in addition to the introduction and conclusion chapters. First, in Chapter 2, I studied gene expression noises of individual genes. I decomposed noises of 3975 mouse genes into intrinsic noise and extrinsic noises and studied their biological

mechanisms and evolution consequences. Next, in Chapter 3, I move forward to consider gene expression noises for pairs of genes simultaneously. I discovered chromosome-wide co-fluctuation in expression for linked genes, which is partly due to chromatin co-accessibilities of linked loci attributable to three-dimensional proximity. I further found that genes encoding components of the same protein complex are more likely to become linked during evolution due to natural selection for intracellular among-component dosage balance. Thus, selection for mitigating the harm of expression noise drives the nonrandom genomic distributions of genes. Finally, in Chapter 4, I studied yet another kind of expression error: mistranslation. I focused on the relationship between mistranslation and codon usage. Specifically, I provide the first direct and global evidence for a prominent but unresolved hypothesis: preferred codons are translated more accurately. Furthermore, I showed that this proposition is generally true across three domains of life. Interestingly, the relative translational accuracies of synonymous codons vary drastically among species, which is mainly explained by the variation of tRNA compositions. Together with other information, these findings suggest that codon usage coevolves with the cellular tRNA pool to maximize translational accuracy and efficiency.

In conclusion, my dissertation documents the genome-wide patterns of gene expression errors and demonstrates their profound impacts on both molecular and phenotypic evolution. The knowledge gained has implications beyond expression errors because of the universality of molecular errors in cellular life.

Chapter 1: General Introduction

Primum non nocere (First, do no harm).

-Thomas Sydenham

Background introduction

Cellular life depends on chemical reactions, which are intrinsically stochastic and imprecise. As a result, many fundamental cellular processes are subject to errors. For instance, every step in the central dogma of molecular biology has errors: DNA replication has an error rate on the order of 10^{-10} per bp per replication, transcription has an error rate on the order of 10^{-5} per bp per transcription, and translation has an error rate on the order of 10^{-4} per amino acid per translation (Milo and Phillips, 2015). Besides, DNA, mRNA, and proteins, the key players of the central dogma, are all subjected to noisy modifications after being produced (Arber and Linn, 1969; Walsh, 2006; Zhao et al., 2017).

Errors in cellular processes have consequences. The vast majority of errors are deleterious or, at best, neutral (Zhang, 2018). Again, if we consider the molecular processes in central dogma: DNA replication errors (mutations) cause cancer (Moolgavkar and Knudson, 1981), transcription and translation errors cause protein misfolding that has been implicated in neurodegenerative diseases (Drummond and Wilke, 2009). Because of the burden of molecular error, many mechanisms have been evolved to reduce the error rate of molecular processes and/or minimize the cost of individual error events, such as homologous recombination DNA repair pathway (Li and Heyer, 2008), Nonsense-mediated decay of mRNAs containing premature stop codons (Chang et al., 2007), and kinetic proofreading in the process of charging tRNA with their corresponding amino-acids (Hopfield, 1974).

Nevertheless, despite not being the main focus of this dissertation, it is worth mentioning that errors can occasionally be beneficial (Tawfik, 2010). As a neat example, mutations are the ultimate source of adaptation (Sniegowski and Lenski, 1995). Either way, errors profoundly impact molecular and phenotypic evolution.

The fast development of omics techniques enables us to study the genome-wide patterns of different types of molecular errors, including but not limited to (i) genomic mutations (Liu and Zhang, 2019), (ii) stochastic initiation of transcription that results in gene expression level fluctuations (Faure et al., 2017), (iii) misincorporation of nucleotides in transcription (Gout et al., 2013), (iv) errors in mRNA processing such as splicing(Pickrell et al., 2010) and polyadenylation (Xu and Zhang, 2018), (v) errors in post-transcriptional modification (Liu and Zhang, 2018), (vi) misincorporation of amino acids in translation (Mordret et al., 2019), and (vii) stop-codon readthrough (Li and Zhang, 2019).

Interestingly, despite the universality and significance of molecular errors, most analysis on omics data assumes molecular diversity observed in the data is beneficial (Gruber and Zavolan, 2019; Modrek and Lee, 2002), perhaps due to the bias inherent in human cognition that favors adaptive storytelling (Gould and Lewontin, 2020). Consequently, numerous dubious ‘genome-wide adaptation’ has been found. For instance, it has been reported that there is a genome-wide convergent adaptation in echolocating mammals(Parker et al., 2013), despite that the same pattern could be found in cow (Thomas and Hahn, 2015; Zou and Zhang, 2015), a non-echolocating mammal. At the transcriptome level, it has been routinely assumed that alternative splicing creates functional diversity and plays an important role in gene expression regulation(Modrek and Lee, 2002). However, proteomics data and various other indirect evidence suggest that only one isoform is translated for the vast majority of the genes (Tress et al., 2017). Finally, the proteome is not an exception: despite some important cases of phosphorylation at particular sites (Rubin and Rosen, 1975), most phosphorylation sites are not conserved and are unlikely to be functional (Studer et al., 2016).

Given that a functional perspective on omics data often results in vague and elusive interpretations, I hypothesize that analyzing these data from the perspective of molecular error could provide a more coherent picture of the genome-wide patterns of molecular diversity. To this end, I analyzed molecular errors occurring in gene expression processes using multiple omics datasets (Mordret et al., 2019; Reinius et al., 2016). Specifically, in my dissertation, I study the mechanisms and consequences of two kinds of gene expression errors: gene expression noise (Blake et al., 2003) and mistranslation (Drummond and Wilke, 2008). Gene expression noise will be the focus of Chapter 2(Sun and Zhang, 2020) and Chapter 3 (Sun and Zhang, 2019), whereas mistranslation would be the focus of Chapter 4. Below, I will briefly summarize the content of each of the three main chapters.

Thesis overview

I first focus on the expression noise of individual genes in Chapter 2. The expression noise of a gene is the variation in the expression level of the gene among genetically identical cells in the same environment (Blake et al., 2003; Elowitz et al., 2002). Gene expression noise is often deleterious because it leads to imprecise cellular behaviors. For example, it may ruin the stoichiometric relationship among functionally related proteins (Veitia, 2004), which may further disrupt cellular homeostasis. However, under certain circumstances, gene expression noise can be beneficial. Prominent examples include bet-hedging strategies of microbes in fluctuating environments (Veening et al., 2008) and stochastic mechanisms for initiating cellular differentiation in multicellular organisms (Huang, 2009). Gene expression noise has extrinsic and intrinsic components (Elowitz et al., 2002). Extrinsic noise arises from cell-to-cell variation in cellular states such as different cell stages, whereas intrinsic noise is caused by the stochastic process of gene expression even under a given cell state. Dissecting expression noise into intrinsic noise and extrinsic noise has provided insights into the causes of expression noise (Raser and O'shea, 2005). However, the existing method for measuring the two noise components is laborious and slow (Elowitz et al., 2002). As a result, accurate knowledge about intrinsic and extrinsic noise is limited to only a few genes, and a general understanding of the pattern, regulation, and evolution of these two noise components is lacking. To address these questions, I designed a high-throughput method for estimating intrinsic and extrinsic expression noises by allele-specific single-cell RNA sequencing (Reinius et al., 2016). Using publicly available data, I estimated the two noise components of 3975 genes in mouse fibroblast cells. My analyses verified predicted influences of several factors such as the TATA-box and microRNA targeting on intrinsic or extrinsic noises and revealed gene function-associated noise trends implicating the action of natural selection. These findings unravel differential regulations, optimizations, and biological consequences of intrinsic and extrinsic noises and can aid the construction of desired synthetic circuits.

While Chapter 2 studies the expression noise of individual genes, no gene functions in isolation. In Chapter 3, I focus on the following questions: if every gene has expression noise, is there any relationship in the expression fluctuations of different genes, and will this relationship have functional and fitness consequences? I hypothesize that neighboring genes on the same chromosome co-fluctuate in expression because of their common chromatin dynamics (Raj et al., 2006). To test this linkage hypothesis, I analyzed the mouse allele-specific single-cell RNA sequencing data used in Chapter 2. Unexpectedly, the co-fluctuation exists not only for neighboring genes but also for genes over 60 million bases apart on the same chromosome. I provided evidence that this long-range effect arises in part from chromatin co-accessibilities of linked loci attributable to three-dimensional proximity, which is much closer intra-chromosomally than inter-chromosomally. Most importantly, I discovered that genes encoding components of the same protein complex tend to become chromosomally linked during evolution, which is likely an outcome of natural selection for intracellular among-component dosage balance (Veitia, 2010). Thus, natural selection mitigating the harm of expression noise has resulted in nonrandom genomic distributions of genes. These findings have implications for both the evolution of genome organization and the optimal design of synthetic genomes in the face of gene expression noise.

Finally, in Chapter 4, I shift gear to study protein mistranslation (Drummond and Wilke, 2008). In particular, I study how mistranslation impacts codon usage evolution (Akashi, 1994), a prominent question in molecular evolution. Analyzing proteomic data from *Escherichia coli* (Mordret et al., 2019), I provide direct, global support for the long-standing hypothesis that preferred codons are translated more accurately. Furthermore, I provide evidence for the generality of this hypothesis across three domains of life. Interestingly, the relative translational accuracies of synonymous codons vary drastically among species, and further analysis reveals a predominant role of the abundance of cognate tRNAs relative to that of near-cognate tRNAs in determining the relative translational accuracy of a codon (Kramer and Farabaugh, 2007). These findings, along with other information (Qian et al., 2012), suggest that codon usage coevolves with the cellular tRNA pool to maximize translational accuracy and efficiency.

In summary, I use novel computational approaches to study gene expression errors and their evolutionary ramifications in my dissertation research. This research is important because molecular and cellular errors are universal, and mitigating such errors is a major task in the evolution of cellular life.

References

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics *136*, 927-935.

Arber, W., and Linn, S. (1969). DNA modification and restriction. Annu Rev Biochem *38*, 467-500.

Blake, W.J., Kærn, M., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633.

Chang, Y.-F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem *76*, 51-74.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341-352.

Drummond, D.A., and Wilke, C.O. (2009). The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet *10*, 715-724.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Faure, A.J., Schmiedel, J.M., and Lehner, B. (2017). Systematic analysis of the determinants of gene expression noise in embryonic stem cells. Cell systems *5*, 471-484. e474.

Gould, S.J., and Lewontin, R.C. (2020). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme (Routledge).

Gout, J.-F., Thomas, W.K., Smith, Z., Okamoto, K., and Lynch, M. (2013). Large-scale detection of in vivo transcription errors. Proceedings of the National Academy of Sciences *110*, 18584-18589.

Gruber, A.J., and Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. Nat Rev Genet *20*, 599-614.

Hopfield, J.J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proceedings of the National Academy of Sciences *71*, 4135-4139.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. Development *136*, 3853-3862.

Kramer, E.B., and Farabaugh, P.J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA *13*, 87-96.

Li, C., and Zhang, J. (2019). Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. PLoS genetics *15*, e1008141.

Li, X., and Heyer, W.-D. (2008). Homologous recombination in DNA repair and DNA damage tolerance. Cell Res *18*, 99-113.

Liu, H., and Zhang, J. (2019). Yeast spontaneous mutation rate and spectrum vary with environment. Curr Biol *29*, 1584-1591. e1583.

Liu, Z., and Zhang, J. (2018). Most m6A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. Mol Biol Evol *35*, 666-675.

Milo, R., and Phillips, R. (2015). Cell biology by the numbers (Garland Science).

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet *30*, 13-19.

Moolgavkar, S.H., and Knudson, A.G. (1981). Mutation and cancer: a model for human carcinogenesis. JNCI: Journal of the National Cancer Institute *66*, 1037-1052.

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G.D., Cox, J., Geiger, T., Lindner, A.B., and Pilpel, Y. (2019). Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. Mol Cell *75*, 427-441. e425.

Parker, J., Tsagkogeorga, G., Cotton, J.A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S.J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. Nature *502*, 228-231.

Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet *6*, e1001236.

Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet *8*, e1002603.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol *4*, e309.

Raser, J.M., and O'shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.

Reinius, B., Mold, J.E., Ramsköld, D., Deng, Q., Johnsson, P., Michaëlsson, J., Frisén, J., and Sandberg, R. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet.

Rubin, C.S., and Rosen, O.M. (1975). Protein phosphorylation. Annu Rev Biochem *44*, 831-887.

Sniegowski, P.D., and Lenski, R.E. (1995). Mutation and adaptation: the directed mutation controversy in evolutionary perspective. Annu Rev Ecol Syst *26*, 553-578.

Studer, R.A., Rodriguez-Mias, R.A., Haas, K.M., Hsu, J.I., Viéitez, C., Solé, C., Swaney, D.L., Stanford, L.B., Liachko, I., and Böttcher, R. (2016). Evolution of protein phosphorylation across 18 fungal species. Science *354*, 229-232.

Sun, M., and Zhang, J. (2019). Chromosome-wide co-fluctuation of stochastic gene expression in mammalian cells. PLoS genetics *15*, e1008389.

Sun, M., and Zhang, J. (2020). Allele-specific single-cell RNA sequencing reveals different architectures of intrinsic and extrinsic gene expression noises. Nucleic Acids Res *48*, 533-547.

Tawfik, D.S. (2010). Messy biology and the origins of evolutionary innovations. Nat Chem Biol *6*, 692-696.

Thomas, G.W., and Hahn, M.W. (2015). Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. Mol Biol Evol *32*, 1232-1236.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Most alternative isoforms are not functionally important. Trends Biochem Sci *42*, 408-410.

Veening, J.-W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol *62*, 193-210.

Veitia, R.A. (2004). Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics *168*, 569-574.

Veitia, R.A. (2010). A generalized model of gene dosage and dominant negative effects in macromolecular complexes. The FASEB Journal *24*, 994-1002.

Walsh, C. (2006). Posttranslational modification of proteins: expanding nature's inventory (Roberts and Company Publishers).

Xu, C., and Zhang, J. (2018). Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. Cell systems *6*, 734-742. e734.

Zhang, J. (2018). Neutral theory and phenotypic evolution. Mol Biol Evol *35*, 1327-1331.

Zhao, B.S., Roundtree, I.A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. Nature reviews Molecular cell biology *18*, 31-42.

Zou, Z., and Zhang, J. (2015). No genome-wide protein sequence convergence for echolocation. Mol Biol Evol *32*, 1237-1241.

Chapter 2: Allele-Specific Single-Cell RNA Sequencing Reveals Different Architectures of Intrinsic and Extrinsic Gene Expression Noises

"Every moment happens twice: inside and outside, and they are two different histories."

-Zadie Smith

2.1 Abstract

Gene expression noise refers to the variation of the expression level of a gene among isogenic cells in the same environment, and has two sources: extrinsic noise arising from the disparity of the cell state and intrinsic noise arising from the stochastic process of gene expression in the same cell state. Due to the low throughput of the existing method for measuring the two noise components, the architectures of intrinsic and extrinsic expression noises remain elusive. Using allele-specific single-cell RNA sequencing, we here estimate the two noise components of 3975 genes in mouse fibroblast cells. Our analyses verify predicted influences of several factors such as the TATA-box and microRNA targeting on intrinsic or extrinsic noises and reveal gene function-associated noise trends implicating the action of natural selection. These findings unravel differential regulations, optimizations, and biological consequences of intrinsic and extrinsic noises and can aid the construction of desired synthetic circuits.

2.2 Introduction

Gene expression noise refers to the variation in gene expression level among genetically identical cells in the same environment (Raser and O'shea, 2005). Gene expression noise is often deleterious, because it leads to imprecise cellular behaviors. For example, it may ruin the stoichiometric relationship among functionally related proteins, which may further disrupt cellular homeostasis (Bahar et al., 2006; Batada and Hurst, 2007; Kemkemer et al., 2002; Lehner, 2008; Wang and Zhang, 2011; Xu et al., 2019). However, under certain circumstances, gene expression noise can be beneficial. Prominent examples include bet-hedging strategies of microbes in fluctuating environments (Veening et al., 2008; Zhang et al., 2009) and stochastic mechanisms for initiating cellular differentiation in multicellular organisms (Chang et al., 2008; Huang, 2009; Turing, 1952).

Gene expression noise has extrinsic and intrinsic components. The extrinsic noise arises from the among-cell variation in cell state such as the cell cycle stage or the concentrations of various transcription factors (TFs), while the intrinsic noise is due to the stochastic process of gene expression even under a given cell state such as the stochastic binding of a promoter to RNA polymerase (Hilfinger and Paulsson, 2011; Sharon et al., 2014; Swain et al., 2002). Note that our definitions of intrinsic and extrinsic noises are based on the source of the noise. Under these definitions, both intrinsic and extrinsic noises can vary among genes. For instance, the intrinsic expression noise of a gene is predicted to be negatively correlated with the mean expression level of the gene (Bar-Even et al., 2006), whereas the extrinsic noise can be different for genes belonging to different biological pathways (Raser and O'shea, 2005). Dissecting gene expression noise into the two components provides insights into its mechanistic basis (Raser and O'shea, 2004). Furthermore, the two noise components can have different biological consequences. For instance, genes regulating the cell cycle should ideally have high extrinsic noise but low intrinsic noise, because their expression levels should be variable among different cell states but stable under the same state. Dissecting the expression noise of a gene into intrinsic and extrinsic components requires a dual reporter assay typically performed in haploid cells by placing two copies of the same gene into the genome, each fused with a distinct reporter gene such as the yellow florescent protein (YFP) gene or cyan florescent protein (CFP) gene (Elowitz et al., 2002). This way, the intrinsic noise in protein concentration can be assessed by the difference between YFP and CFP concentrations within cells while the extrinsic noise can be measured by the covariation between YFP and CFP concentrations among cells. However, such experiments are laborious in strain construction and expression quantification, hindering the examination of many genes. Consequently, past genome-wide studies of gene expression noise measured only the total noise (Faure et al., 2017; Newman et al., 2006; Taniguchi et al., 2010; Zoller et al., 2015). Some authors attempted to focus on the intrinsic noise by limiting the analysis to cells of similar morphologies (Newman et al., 2006; Taniguchi et al., 2010). But because the extrinsic noise is not completely eliminated in the above experiments, the estimated intrinsic noise is inaccurate. Furthermore, these experiments could not study the extrinsic noise. As a result, accurate knowledge about intrinsic and extrinsic noise is limited to only a few genes (Elowitz et al., 2002; Stewart-Ornstein et al., 2012), and a general understanding of the pattern, regulation, and evolution of these two noise components is lacking.

Here we propose to use allele-specific single-cell RNA sequencing (scRNA-seq) to estimate the intrinsic and extrinsic expression noises at the mRNA level. When the two alleles of a gene are distinguished by their DNA sequences, the distinct sequences serve as dual reporters of mRNA concentrations in scRNA-seq. Our method is thus in principle similar to the classical dual reporter assay except that we study the intrinsic and extrinsic expression noises at the mRNA level whereas the classical assay studies them at the protein level. Because the protein noise is widely believed to arise primarily from the mRNA noise (Bar-Even et al., 2006; Sherman et al., 2015), findings about the latter will not only inform us the mRNA noise but also largely the protein noise. Because the dual reporters exist naturally at any heterozygous locus of the genotype investigated and because single-cell expression levels of all genes in the genome are measured simultaneously by scRNA-seq, our method can estimate the intrinsic and extrinsic expression noises at the genomic scale from one scRNA-seq experiment of a highly heterozygous genotype. Using publically available allele-specific scRNA-seq data from mouse fibroblast cells (Reinius et al., 2016), we estimate the intrinsic and extrinsic noises of 3975 genes, allowing depicting the architectures of the two noise components in mouse cells.

2.3 Materials and Methods

2.3.1 Intrinsic and extrinsic noise in diploid cells

Let *Y* be the expression level of a gene in a cell and let *X* describe the cell state. *Y* is a random variable that is a function of the random variable *X*. Gene expression noise is commonly measured by noise strength $\eta_{tot}^2 = Var(Y)/E^2(Y)$, where *Var* stands for variance and *E* stands for expectation. According to the law of total variance, $\frac{Var(Y)}{E^2(Y)} = \frac{E(Var(Y|X))}{E^2(Y)} + \frac{Var(E(Y|X))}{E^2(Y)}$, where the first term on the right-hand side of the equation describes the variation of *Y* given *X*, or intrinsic noise strength η_{int}^2, and the second term describes the variation of *Y* due to the variation of *X*, or extrinsic noise strength η_{ext}^2.

Most past studies of intrinsic and extrinsic expression noises of a gene were conducted in haploid cells by placing two copies of the gene (under the control of two identical, independent promoters) in the genome, each copy carrying a unique marker. Let the expression levels of the two gene copies be Y_1 and Y_2, respectively. It was found that the intrinsic noise of each gene copy can be expressed by $\eta_{int,H}^2 = \frac{E[(Y_1-Y_2)^2]}{2E(Y_1)E(Y_2)}$ and the extrinsic noise of each gene copy can be expressed by $\eta^2_{ext,H} = \frac{Cov(Y_1,Y_2)}{E(Y_1)E(Y_2)}$, where the subscript *H* indicates haploid and *Cov* indicates covariance (Swain et al., 2002).

Now let us consider a diploid cell in which the two alleles of the focal gene are controlled by two identical, independent promoters and have unique markers. We are interested in the noise of the total expression level of the two alleles. Because the expression levels of the two alleles are independent given the cell state, by definition, the intrinsic expression noise in diploid cells is $\eta^2_{int,D} = \frac{E(Var((Y_1+Y_2)|X))}{E^2(Y_1+Y_2)} = \frac{E((Var(Y_1)+Var(Y_2))|X)}{4E^2(Y_1)} = \frac{2E(Var(Y_1)|X)}{4E^2(Y_1)} = \eta^2_{int,H}/2$. Similarly, by definition, the extrinsic expression noise strength in diploid cells is $\eta^2_{ext,D} = \frac{Var(E(Y_1+Y_2)|X)}{E^2(Y_1+Y_2)} = \frac{Var(2E(Y_1)|X)}{4E^2(Y_1)} = \frac{Var(E(Y_1)|X)}{E^2(Y_1)} = \eta^2_{ext,H}$. Thus, we can adapt previously obtained formulas of intrinsic and extrinsic noise in haploid cells for the study of diploid cells.

2.3.2Allele-specific single-cell RNA-seq data and data preprocessing

The raw read counts of allele-specific scRNA-seq data (Reinius et al., 2016) were downloaded from https://github.com/RickardSandberg/Reinius_et_al_Nature_Genetics_2016?files=1 (mouse.c57.counts.rds and mouse.cast.counts.rds). We preprocessed the dataset by requiring that (i) all cells have the same genotype and (ii) there are spike-in standards in each cell. Two groups of cells satisfied our criteria: 60 cells from clone 7 and 75 cells from different clones or different individuals (IDs in the raw read-count dataset are 24-26, 28, 29, 31-35, 37-44, 46, 48-51, 53, 55, 58-60, and 124-170). Note that the latter group of cells are non-clonal and were isolated in different experiments; so they likely have larger variations in expression. Our analysis thus focused primarily on clone 7, although most results were also reproduced in the non-clonal cells. Because of the dual reporter design of our analysis, sex-linked genes were removed. For clone 7, we further removed genes on Chromosomes 3 and 4 due to aneuploidy. To ensure the relative reliability of our noise estimates, we limited the analysis to genes that have on average ≥5 reads mapped to each allele across cells. We then corrected the read counts mapped to each allele in each cell using spike-ins according to the following procedure. First, we obtained the number of reads mapped to spike-in molecules in each cell, yielding an array of 60 numbers, each specifying the number of reads mapped to spike-in molecules in one cell. Second, we divided each entry in the array by the largest number in the array, creating an array of 60 normalized factors that are all between 0 and 1. Third, we calibrated the number of reads mapped to each allele in each cell by dividing the original read number by the corresponding normalized factor in the array.

The noise decomposition requires the two reporters to have the same expression distribution. However, due to imprinting and polymorphisms in the regulatory regions, some genes might not have two alleles that are identically regulated. We thus performed a Kolmogorov–Smirnov test for the single-cell expression levels of the two alleles of each gene, and removed genes with $P < 0.05$ after multiple-testing correction (Benjamini-Hochberg correction). The data from the non-clonal cells were processed similarly. Some authors suggested normalizing single-cell expression levels of each reporter by its mean expression level to deal with unequal regulations between alleles (Fu and Pachter, 2016; Rhee et al., 2014). While this processing should allow analyzing more genes, the statistical properties of the normalization are not well understood. To be conservative, we chose to remove genes that do not satisfy the assumption of the dual reporter experiment instead of normalizing the expression levels.

2.3.3 Estimation of intrinsic and extrinsic noises

We estimated the intrinsic and extrinsic expression noises of haploids using an existing program (Fu and Pachter, 2016) and then converted them to the corresponding values in diploids using the formulas described above. We then derived noise estimates that are independent of the mean expression level and the mean read number, which is inversely correlated with the amount of technical noise (Grün et al., 2014). Because the exact forms of the above dependencies are unknown, we used a rank-based measure. Specifically, we performed robust linear regression of the rank of intrinsic (or extrinsic) noise on the rank of expression level and the rank of read number using the 'rlm' function of the 'MASS' package with default options in R; the residual from the regression, D_{int} (or D_{ext}), is the measurement of intrinsic (or extrinsic) noise. To obtain the intrinsic noise estimate of a gene that is also independent of its extrinsic noise, we regressed the rank of intrinsic noise on the rank of mean expression level, the rank of mean read number, and the rank of extrinsic noise simultaneously. The obtained residual is referred to as D'_{int}. We similarly obtained D'_{ext}. The procedure used to process the data and estimating the two noise components is summarized in **Fig. A1-1**.

2.3.4 Assessment of technical extrinsic noise using spike-in molecules

We assessed the extrinsic technical noise using spike-in molecules from clone 7 and non-clonal cells. First, we estimated the mean read number of each spike-in species from the corrected read number of each spike-in molecule in each cell. The correction procedure was the same as used for correcting allele-specific reads mapped to each gene. Second, we ordered the spike-in molecules by their mean read numbers and paired neighboring spike-in molecules whose mean read numbers are similar. For each pair of spike-in molecules, we used binomial sampling to down-sample in each cell the raw reads of the spike-in molecule whose mean read number is larger, according to the ratio between the mean read numbers of the two spike-in molecules. Finally, each pair of spike-in molecules was treated as two alleles of the same spike-in transcripts for estimating extrinsic noise. As in the analysis of actual genes, we filtered out spike-in molecules whose mean (raw) read numbers are smaller than 5.

2.3.5 Factors influencing intrinsic and extrinsic noise

Mouse genes with a TATA-box were downloaded from the Eukaryotic Promoter Database (EPD) (Dreos et al., 2016). Information of mouse miRNAs and their targets was downloaded from the RegNetwork database (Liu et al., 2015). Information about mouse *trans*-regulators and their target genes was also downloaded from RegNetwork (Liu et al., 2015). Note that miRNAs were considered *trans*-regulators in the database; so were they in our analysis. Some transcription factors target themselves. Because the total noise of a gene by definition correlates with the intrinsic and extrinsic noises of the gene, we removed the self-targeting pairs in the analysis of *trans*-regulators. This problem does not involve miRNAs because we have no miRNA noise measures.

To test the hypothesis that genes targeted by the same *trans*-regulator tend to have similar D_{ext}, we grouped genes that share a *trans*-regulator and computed the standard deviation (SD) of their D_{ext} within the group. We then computed the median SD across all groups. Because SD is undefined for groups containing only one gene, such groups were discarded. We also removed *trans*-regulators that have noise measures and are target genes, such that the regulators and targets have no overlaps.

To analyze the relationship between histone modifications and noise, we downloaded the computed modification peak position data from the Cistrome database (Liu et al., 2011). We focused on four types of histone modifications in mouse wild-type fibroblast cells: H3K4Me1 (Chronis et al., 2017), H3K4Me2 (Chronis et al., 2017), H3K4Me3 (Xie et al., 2017), and H3K27AC (Xie et al., 2017) . All four datasets used were of high quality and passed quality criteria of Cistrome. For each modification, we computed Spearman's correlation between the number of peaks overlapped with the core promoter region ((TSS − 200 bp, TSS + 100 bp), defined in (Faure et al., 2017)) and D'_{int} or D'_{ext}. The results are presented in **Fig. 2-2M.**

2.3.6 Noise comparison among genes of different functions

GO terms of mouse genes were downloaded from Ensembl BioMart (GRC38m.p5) (Aken et al., 2016). Genes functioning in the mitochondrion are associated with the GO cellular component term of "mitochondria", whereas cell cycle genes are associated with the GO biological process term of "cell cycle". Mouse protein complex data were downloaded from the CORUM database (http://mips.helmholtz-muenchen.de/corum/) (Ruepp et al., 2009).

To evaluate if a group of genes with a certain function (i.e., focal genes) are enriched/deprived with the TATA-box or miRNA targeting, we compared the group with other genes (i.e., non-focal genes) after controlling mean expression levels across 13 mouse tissues (Söllner et al., 2017). Specifically, we ranked the focal genes by the mean expression level and divided them into 50 equal-size bins. We then obtained non-focal genes falling into each of these expression bins and identified the smallest number (*m*) of non-focal genes of all bins. We randomly picked *m* non-focal genes per bin and used this set of non-focal genes to compare with the focal genes. As expected, the non-focal genes showed similar expression levels as the corresponding focal genes (P = 0.28 for genes functioning in the mitochondrion, P = 0.37 for genes encoding protein complex members, and P = 0.45 for cell cycle genes; Mann-Whitney U test). The non-focal genes are referred to as the "expression stratified control genes".

DAVID GO web server with default options was used to perform the GO term enrichment analysis (Huang et al., 2008), in which all genes with estimated D_{int} and D_{ext} were used as the background. The web server returned the *P*-value after Benjamini-Hochberg correction for multiple testing. We ranked the GO terms by the significance level and reported the three most significant GO terms for each group of genes with specific noise properties, if more than three GO terms were significantly enriched.

2.4 Results

2.4.1 High-throughput estimation of intrinsic and extrinsic expression noises

The expression noise of a gene is commonly measured by the noise strength η^2, which is the among-cell variance in expression level divided by the squared mean expression level. On the basis of previously derived formulas of intrinsic and extrinsic noises in haploids (Swain et al., 2002), we derived formulas for estimating intrinsic (η_{int}^2) and extrinsic (η_{ext}^2) noises in diploids (see Materials and Methods). Let the expression levels of the two alleles of a gene in a diploid cell be Y_1 and Y_2, respectively. If the two alleles are controlled by two independent, identical promoters, $\eta_{int}^2 = \frac{E[(Y_1-Y_2)^2]}{4E(Y_1)E(Y_2)}$ and $\eta_{ext}^2 = \frac{Cov(Y_1,Y_2)}{E(Y_1)E(Y_2)}$, where E and Cov respectively stand for expectation and covariance. Graphically, when the expression levels of the two alleles in each cell are respectively plotted on the x-axis and y-axis of a dot plot, extrinsic noise is represented by the spread of dots along the diagonal line of $y = x$, whereas the intrinsic noise is represented by the spread of dots along the direction perpendicular to the diagonal (left panel in **Fig. 2-1A**). As an example, single-cell expression levels of the gene *Tcof1* are plotted (right panel in **Fig. 2-1A**).

To estimate intrinsic and extrinsic gene expression noises, we used the scRNA-seq data of mouse fibroblast cells from an F1 hybrid of two mouse strains (Reinius et al., 2016). Note that scRNA-seq data are subject to large technical noises, which may also be decomposed into intrinsic and extrinsic technical noises (Grün et al., 2014). The intrinsic technical noise is primarily caused by the low capturing efficiency of cellular transcripts and can result in a high variance and high dropout rate in estimating the mRNA expression level. The intrinsic technical noise artificially increases the level of the estimated intrinsic expression noise. The extrinsic technical noise is mainly due to tube-to-tube variability in capturing efficiency and artificially increases the level of the estimated extrinsic expression noise. Imputation, which substitutes the observed expression level of a gene in a cell by its expected expression level, is often used to deal with technical noises in scRNA-seq-based cell classification (Wagner et al., 2016). But, imputation cannot be used in our study because it leads to underestimation of gene expression noise. Therefore, we only used spike-in control molecules to normalize expression levels in individual cells (see Materials and Methods).

Our analysis focused on clone 7 (derived from the hybrid of CAST/EiJ male × C57BL/6J female) in the data, because (1) the number of sequenced cells (n = 60) is the largest in this clone, and (2) all sequenced cells from this clone have spike-in control molecules, permitting accurate read count estimation. Upon the removal of genes whose two alleles show significantly different among-cell expression distributions and other steps of data processing (**Fig. A1-1**; see Materials and Methods), we obtained the intrinsic and extrinsic expression noises of 3975 genes. To assess the precision of our noise estimates, we randomly separated the cells of clone 7 into two 30-cell groups. We found that the estimates of the intrinsic noise of a gene from the two subsamples are highly correlated (Pearson's r = 0.79, $P < 1\times10^{-300}$; Spearman's ρ = 0.79, $P < 1\times10^{-300}$; **Fig. 2-1B**), while those of extrinsic noise are moderately correlated (r = 0.42, $P = 2.3\times10^{-151}$; ρ = 0.44, $P = 3.8\times10^{-185}$; **Fig. 2-1C**). Note that the above correlations demonstrate the precision rather than the accuracy of our measurements. The accuracy of our measurements depends on technical noises, which can in principle be estimated using spike-in molecules, because they have no biological variation among cells. However, two factors render the technical noises of spike-in molecules not directly comparable with those of natural transcripts. First, spike-in molecules provide information of the technical noise in sample preparation steps after the addition of spike-in molecules, so the technical noises associated with earlier steps are unknown (Wagner et al., 2016). Second, spike-in molecules have much lower capturing efficiencies (Svensson et al., 2017) than natural transcripts. Nonetheless, it can be shown that, after normalization by spike-in molecules (see Materials and Methods), extrinsic noises disappear for spike-in molecules (red dots in **Fig. A1-2**), whereas extrinsic noises for natural transcripts remain substantial (black dots in **Fig. A1-2**), indicating that the tube-to-tube variation in sample preparation steps after the addition of spike-in molecules has been corrected. Because the magnitudes of technical noises cannot be estimated in our dataset and because the measurements of intrinsic and extrinsic noises are subject to different technical noises, it is not possible to directly compare the contributions of intrinsic noise and extrinsic noise to the total noise in the data analyzed. Nevertheless, with proper statistical processing, we can compare extrinsic or intrinsic noise among genes.

In addition to clone 7, there is another group of cells with $n = 75$ that fulfill the above two criteria (see Materials and Methods), but this group of cells are non-clonal and were isolated in different experiments, so may be more heterogeneous in cell state and subject to larger technical variabilities. Our analysis thus focused primarily on clone 7, although most results were also reproduced in the non-clonal cells. While the precision of the intrinsic noise estimates is similarly high in the non-clonal cells ($r = 0.80$, $P < 1\times10^{-300}$; $\rho = 0.79$, $P < 1\times10^{-300}$; **Fig. A1-3A**) when compared with that in the clonal cells (**Fig. 2-1B**), the estimates of the extrinsic noise are much less precise in the non-clonal cells ($r = 0.31$, $P = 1.25\times10^{-102}$; $\rho = 0.24$, $P = 6.9\times10^{-65}$; **Fig. A1-3B**) than in the clonal cells (**Fig. 2-1C**), probably for the aforementioned reasons. The assessment of technical noise in non-clonal cells (**Fig. A1-3C**) yielded similar results as in clone 7 cells (**Fig. A1-2**).

In theory, the intrinsic expression noise of a gene should decrease with the mean expression level of the gene (Bar-Even et al., 2006; Hornung et al., 2012a), whereas no such relationship is expected for the extrinsic noise. We confirmed that our estimate of the intrinsic noise is indeed strongly negatively correlated with the mean expression level (Spearman's $\rho = -0.81$, $P < 1.0\times10^{-300}$; **Fig. 2-1D**). A similar trend was observed from the non-clonal cells (**Fig. A1-3D**). Intriguingly, we also found a weak, but significant negative correlation between the extrinsic noise and mean expression level (ρ = -0.083, P = 1.9×10^{-7}; **Fig. 2-1E**). Because the extrinsic noise is the normalized covariance between Y_1 and Y_2, and because the normalized covariance tends to be underestimated for lowly expressed genes due to larger sampling errors, the estimated extrinsic noise is expected to be positively correlated with the mean expression level for technical reasons. To assess the impact of the technical noise on extrinsic expression noise, we correlated across genes the extrinsic noise with the mean allele-specific read number, because the mean read number is not normalized by gene length so contains more information about the technical variation when compared with the mean expression level. Indeed, a positive correlation is observed between the estimated extrinsic noise and mean allele-specific read number instead of expression level (ρ = 0.06, P = 3.4×10^{-5}). Thus, the observed negative correlation between extrinsic noise and expression level is likely biological. The trend observed in the non-clonal cells is similar to that in the clonal cells (**Fig. A1-3E**).

It is preferable to remove the correlation between a noise measure and the mean expression level in order to identify factors that impact intrinsic or extrinsic noise not simply due to their influences on the mean expression level. In addition, because technical noise in scRNA-seq decreases with mean read number (Grün et al., 2014), it would be important to further remove the impact of the mean read number on our expression noise measures. To this end, we used robust linear regressions to remove the covariations with the mean expression level and mean read number in our measures of intrinsic and extrinsic noise (see Materials and Methods), which are referred to as D_{int} and D_{ext}, respectively. Note that D_{int} and D_{ext} are residuals in the regressions of expression noise ranks so have values potentially from -3975 to 3975. We used ranks instead of raw noise estimates because we do not know the exact relationship between the noises and the mean expression level or read number, because the expression noise estimates contain contributions from technical noises, and because rank statistics are robust to outliers. As expected, D_{int} is correlated with neither the mean expression level (ρ = -0.003, P = 0.85) nor the mean read number (ρ = -0.004, P = 0.82). Similarly, D_{ext} is correlated with neither the mean expression level (ρ = -0.002, P = 0.89) nor the mean read number (ρ = -0.0005, P = 0.98). To assess the precision of these new noise measures, we plotted the correlation between the estimates from two subsamples of clone 7 for D_{int} (**Fig. 2-1F**) and D_{ext} (**Fig. 2-1G**), respectively. We found the rank correlation of D_{int} from the two subsamples (r = 0.44, P = 1.7×10^{-180}; ρ = 0.40, P = 2.4×10^{-149}) similar to that of D_{ext} from the two subsamples (r = 0.44, P = 1.3×10^{-182}; ρ = 0.44, P =1.7×10^{-183}). Because our subsequent statistical analyses of D_{int} and D_{ext} are all rank-based, the measurement precision of D_{int} and D_{ext} can be treated as comparable. Compared with those in the clonal cells, the precision of D_{int} is similar (r = 0.48, P = 6.1×10^{-272}; ρ = 0.40, P = 1.3×10^{-188}; **Fig. A1-3F**) but that of D_{ext} is lower (r = 0.24, P = 8.8×10^{-66}; ρ = 0.23, P = 2.7×10^{-64}; **Fig. A1-3G**) in the non-clonal cells.

Interestingly, we observed a weak, but significant positive correlation between D_{int} and D_{ext} (ρ = 0.11, P = 3.8×10^{-12}; **Fig. 2-1H**). Similar results were obtained from the non-clonal cells (ρ = 0.047, P = 0.0008; **Fig. A1-3G**). Although previous theoretical studies predicted a dependency of intrinsic noise on extrinsic noise, the direction of the correlation was unpredicted (Hilfinger and Paulsson, 2011; Shahrezaei et al., 2008; Sherman et al., 2015). Because of this observed correlation, we further acquired an intrinsic noise estimate that is independent of the extrinsic noise by regressing the rank of intrinsic noise on the rank of mean expression level, the rank of mean read number, and the rank of extrinsic noise simultaneously. The obtained rank residual, referred to as D'_{int}, is correlated with none of the mean expression level (ρ = -0.002, P = 0.88), mean read number (ρ = -0.002, P = 0.90), and extrinsic noise (ρ = -0.003, P = 0.85). We similarly obtained D'_{ext}, which is correlated with none of the mean expression level (ρ = -0.005, P = 0.76), mean read number (ρ = -0.002, P = 0.91), and intrinsic noise (ρ = 0.005, P = 0.72). Finally, we used the “scran” package to divide the cells from clone 7 into G1 and G2–M cell cycle stages based on the total reads of each gene in each cell (Lun et al., 2016). We then computed D'_{int} and D'_{ext} of each gene in each stage. We found that both D'_{int} and D'_{ext} are similar between the stages (**Fig. A1-3I** and **J**, which can be compared with **Fig. 2-1F** and **G**, respectively), indicating that the adjusted noise is a robust property of a gene across cell cycle stages.

2.4.2 The TATA-box is associated with elevated intrinsic and extrinsic noises

Our estimates of D_{int} and D_{ext} for thousands of mouse genes allow testing the potential impacts of several factors on the two noise components. We focused on three factors with prior theoretical predictions of their effects. The first factor is the presence/absence of the TATA-box in the promoter region. The TATA-box has been predicted to increase the intrinsic noise because it enlarges the burst size in bursty gene expression through interacting with nucleosomes (Blake et al., 2006; Hornung et al., 2012a). In addition, the TATA-box can increase intrinsic noise by reducing the number of states in promoter cycles (Zoller et al., 2015). Indeed, D_{int} is significantly higher for genes with the TATA-box in the promoter than those without (**Fig. 2-2A**). The same is true for D'_{int}, which is independent of D_{ext} (**Fig. 2-2A**). Similar results were obtained from the non-clonal cells (**Fig. A1-4A**).

The presence of the TATA-box sensitizes the promoter to *trans*-regulation (Hornung et al., 2012b; Tirosh and Barkai, 2008) so should also increase the susceptibility of the promoter to cell state changes (Paulsson, 2004; Pedraza and van Oudenaarden, 2005). Hence, we predict that the TATA-box also raises the extrinsic noise. Supporting this prediction, genes with the TATA-box show significantly higher D_{ext} and D'_{ext} than those without (**Fig. 2-2B**). Similar patterns were observed in the non-clonal cells (**Fig. A1-4B**).

Because the above analyses of the TATA-box are based on correlations, they do not prove causality. Nevertheless, the only other known property of the TATA-box on gene expression is to increase the mean expression level (Kim et al., 1993), which has already been controlled in our D_{int} and D_{ext} estimates. Our observations, coupled with manipulative experiments showing increased (total) expression noise conferred by the TATA-box (Blake et al., 2006; Murphy et al., 2010; Raser and O'shea, 2004), suggests that the influences of the TATA-box on both intrinsic and extrinsic noise revealed here is causal.

2.4.3 Opposing effects of microRNAs on the intrinsic and extrinsic noise of target genes

A microRNA (miRNA) regulates the expressions of its target genes by degrading their mRNAs and/or suppressing their translations (Bartel, 2018). Combining mathematical modeling and experimental validation, Schmiedel et al. showed that a gene would have an elevated extrinsic protein expression noise if it is targeted by a miRNA than when it is not, because the miRNA concentration varies among cells (Schmiedel et al., 2015). For the same reason, we expect that miRNA targeting increases the extrinsic mRNA expression noise. Schmiedel et al. also showed that the protein intrinsic noise of a gene is reduced when it is targeted by a miRNA than when it is not (Schmiedel et al., 2017). This is because, under the assumption that the mean mRNA concentration is unaltered, being targeted by a miRNA means a reduction in mRNA half-life and a compensatory increase in transcription. Even though the magnitude of the fluctuation of the mRNA concentration in a cell may be unaltered (see below), the frequency of the fluctuation is higher, which leads to a lower protein intrinsic noise. However, the impact of miRNA targeting on the mRNA intrinsic noise depends on the mechanism underlying the compensatory increase in transcription. If the increased transcription is caused by a higher burst frequency in transcriptional initiation, mRNA intrinsic noise will be reduced. Alternatively, if it is caused by a greater burst size, mRNA intrinsic noise will be increased. It is also possible that the increased transcription is due to a combination of the two mechanisms. We thus explore the following three questions.

First, do genes targeted by miRNAs have lower or higher D_{int} and D'_{int} than those not targeted by miRNAs? Second, do genes targeted by more miRNA species have lower or higher D_{int} and D'_{int}? Third, do genes targeted by miRNAs have higher D_{ext} and D'_{ext} than those not targeted by miRNAs? We obtained relationships between miRNAs and their targets from the RegNetwork database (Liu et al., 2015) (see Materials and Methods). We found that genes targeted by miRNAs have significantly lower D_{int} and D'_{int} than genes not targeted by miRNAs (**Fig. 2-2C**). Furthermore, D_{int} (**Fig. 2-2D**) and D'_{int} (**Fig. 2-2E**) of a gene are significantly negatively correlated with the number of miRNA species targeting the gene. Regarding the extrinsic noise, D_{ext} and D'_{ext} are significantly higher for genes targeted by miRNAs than those not targeted by miRNAs (**Fig. 2-2F**). Similar results were obtained from the non-clonal cells (**Fig. A1-4C-F**), except that the results on D_{ext} and D'_{ext} are statistically non-significant (**Fig. A1-4F**), probably due to the aforementioned lower precision of extrinsic noise estimates in the non-clonal cells. Because the only other known function of miRNAs is to regulate the mean expression levels of their targets (Bartel, 2018), which are uncorrelated with our noise measures, it is likely that the effects observed here are causal.

2.4.4 Similar extrinsic noises of genes regulated by the same *trans*-regulator

According to the definitions of intrinsic and extrinsic noises, we predict that, if gene A *trans*-regulates gene B, the extrinsic but not intrinsic noise of gene B should rise with the expression noise of gene A. To test this prediction, we obtained the relationship between *trans*-regulators and their target genes from RegNetwork (Liu et al., 2015). Because both η_{int}^2 and η_{ext}^2 of the *trans*-regulator affect the extrinsic noise of the target genes, we need a measure of the *trans*-regulator noise that takes into account both η_{int}^2 and η_{ext}^2. For each *trans*-regulator that has estimated η_{int}^2 and η_{ext}^2, we computed its $\eta_{tot}^2 = \eta_{int}^2 + \eta_{ext}^2$. Here, we gave equal weights to the measured η_{int}^2 and η_{ext}^2, because of the lack of knowledge of the relative measurement accuracy of η_{int}^2 and η_{ext}^2. We then computed the average D_{int} and average D_{ext} of all the targets of the *trans*-regulator, respectively, after excluding the *trans*-regulator itself if it self-regulates, because the extrinsic noise of a gene is by definition correlated with its total noise irrespective of the validity of our hypothesis. In support of our hypothesis, we found a positive correlation between the mean target D_{ext} and η_{tot}^2 of their *trans*-regulator ($\rho = 0.27$, $P = 0.0024$; **Fig. 2-2G**). The same is true for D'_{ext} ($\rho = 0.25$, $P = 0.0047$; **Fig. 2-2H**). By contrast, although the mean D_{int} of the targets and η_{tot}^2 of their *trans*-regulator are correlated ($\rho = 0.20$, $P = 0.031$; **Fig. 2-2I**), the correlation becomes non-significant for D'_{int} ($\rho = 0.15$, $P = 0.091$; **Fig. 2-2J**). In the above, we considered η_{tot}^2 because it is the total noise of the regulator regardless of its source that influences the target extrinsic noise.

It can be further predicted that genes regulated by the same *trans*-regulator should have more similar D_{ext} values but not necessarily more similar D_{int} values, when compared with genes that are not co-regulated by a *trans*-regulator. To test this prediction, we grouped all target genes of each *trans*-regulator, followed by calculation of the standard deviation (SD) of D_{int} and that of D_{ext} within the group. We then computed the median SD of D_{int} and median SD of D_{ext} across all *trans*-regulators. As a comparison, we randomized the targets of each regulator, requiring only that the number of targets of each regulator remained unaltered (see Materials and Methods). We then similarly computed the median SD of D_{int} and median SD of D_{ext} across all *trans*-regulators. This randomization was repeated 10,000 times. We found that the observed median SD of D_{ext} is significantly lower than that from each of the 10,000 randomizations (i.e., $P < 0.0001$; **Fig. 2-2K**). By contrast, the observed median SD of D_{int} is smaller than that in only 25% of the 10,000 randomizations (i.e., $P = 0.75$; **Fig. 2-2L**). Together, our results confirm the theoretical prediction that the expression noise of *trans*-regulators primarily affects the extrinsic but not intrinsic expression noise of their targeted genes. We also performed the same analyses in the non-clonal cells. Although the trends exist, they are not statistically significant (**Fig. A1-4G-J**), likely due to the less precise estimation of expression noise in the non-clonal cells.

2.4.5 Differential effects of histone modification on intrinsic and extrinsic noises

In addition to the above factors, correlations between several histone modifications and gene expression noise has been reported (Chen and Zhang, 2016; Wu et al., 2017). Prompted by these studies, we respectively examined correlations between histone modification and intrinsic and extrinsic expression noises. To this end, we collected histone modification peak data from Cistrome (Liu et al., 2011), and computed the correlation between histone modification strength in the core promoter and D'_{int} or D'_{ext}. We found H3K4Me1 modification to be significantly positively correlated with D'_{int} but not significantly correlated with D'_{ext} (**Fig. 2-2M**). The same can be said for H3K4Me2 (**Fig. 2-2M**). By contrast, H3K4Me3 modification is significantly negatively correlated with both D'_{int} and D'_{ext}, but the correlation with D'_{ext} is much stronger than that with D'_{int} (**Fig. 2-2M**). H3K27Ac modification is significantly negatively correlated with D'_{int} but not significantly correlated with D'_{ext} (**Fig. 2-2M**). These observations suggest that histone modification often differentially impacts intrinsic and extrinsic expression noises.

The genome-wide finding that (i) the TATA-box increases both D_{int} and D_{ext}, (ii) miRNAs decrease the D_{int} but increase the D_{ext} of its targets, (iii) the D_{ext} but not D_{int} of a gene is impacted by the expression noise of its *trans*-regulator and (iv) histone modification differentially impacts D'_{int} and D'_{ext} not only reveals mechanisms responsible for the variations of intrinsic and extrinsic expression noises among genes, but also demonstrates that our high-throughput estimation of intrinsic and expression noises is reliable. Because the above analyses were all based on rank statistics, the absolute effect sizes are unknown and hence it is hard to answer whether the above findings are biologically relevant. In the following section, we address this question by examining patterns of D_{int} and D_{ext} among genes of various functions and testing if the two noise components have been subject to differential natural selection.

2.4.6 Genes with mitochondrial functions show lowered extrinsic expression noise

Previous studies found that the variation in mitochondrial function among cells is a primary source of global extrinsic noise of gene expression, because protein synthesis requires ATP, which is largely produced by the mitochondrion (Das Neves et al., 2010; Johnston et al., 2012). We thus predict that natural selection should have minimized the expression noise of (nuclear) genes that function in the mitochondrion in order to reduce the gene expression noise globally. Indeed, one source of the protein level noise of proteins localized to the mitochondrion is the partition of mitochondria during the cell division, and recent work showed that this partition is tightly regulated presumably to ensure equal partitions (Jajoo et al., 2016). To achieve a low expression noise at the mRNA level for nuclear genes with mitochondrial functions, selection could have reduced the intrinsic noise, extrinsic noise, or both. However, for highly expressed genes, the extrinsic noise is the main contributor to expression noise, because the intrinsic noise is naturally low when the mean expression is high (Schmiedel et al., 2015; Taniguchi et al., 2010). We noticed in our data that nuclear genes of mitochondrial functions are highly expressed relative to other nuclear genes ($P = 1.9\times10^{-15}$, Mann–Whitney U test). Because D_{int} and D_{ext} are independent of the mean expression level, we predict that genes functioning in the mitochondrion should have reduced D_{ext} but not necessarily reduced D_{int}. Indeed, D_{ext} is significantly lower for nuclear genes functioning in the mitochondrion when compared with other nuclear genes (**Fig. 2-3A**), and this disparity remains for D'_{ext} (**Fig. 2-3A**). By contrast, D_{int} is not significantly different between the two groups of genes (**Fig. 2-3B**), whereas D'_{int} is even slightly larger for genes functioning in the mitochondrion than other genes (**Fig. 2-3B**). Similar results were obtained from the non-clonal cells (**Fig. A1-5**).

What are the underlying molecular mechanisms responsible for the reduction of D_{ext} of genes functioning in the mitochondrion? Based on the earlier results (**Fig. 2-2**), possible mechanisms include the underrepresentation of the TATA-box in genes functioning in the mitochondrion, underrepresentation of miRNA targeting, and preferential regulation by quiet *trans*-regulators. Because our noise data do not include many *trans*-regulators, we focused on the first two mechanisms. Indeed, compared with other genes, those functioning in the mitochondrion are depleted of the TATA-box ($P = 4.6\times10^{-5}$, Fisher's exact test; **Fig. 2-3C**) and are less targeted by miRNAs ($P = 0.036$, Fisher's exact test; **Fig. 2-3D**). To explore whether the depletion of TATA-box and miRNA targeting can fully account for the reduction in extrinsic noise of nuclear genes functioning in the mitochondrion, we regressed D_{ext} as a linear function of the presence/absence of TATA-box and miRNA targeting. The residual of the above regression provided an extrinsic noise measure upon the control for TATA-box and miRNA targeting. We found that the difference in extrinsic noise between nuclear genes that function in the mitochondrion and other genes remains significant (D_{ext}: $P = 0.001$, Mann–Whitney U test; D'_{ext}: $P = 0.00065$, Mann-Whitney U test). Thus, depletions of the TATA-box and miRNA targeting are only part of the mechanisms responsible for the selective reduction of the D_{ext} of genes functioning in the mitochondrion.

2.4.7 Genes encoding protein complex members have lowered intrinsic expression noise

Because dosage balance is important for protein complex members (Birchler and Veitia, 2012; Papp et al., 2003) and because as long as members of the same protein complex are co-regulated in expression, extrinsic noise does not create dosage imbalance (Stewart-Ornstein et al., 2012), we predict that protein complex members have reduced intrinsic noise but not necessarily reduced extrinsic noise. An early yeast study showed that, compared with other proteins, protein complex members have lowered protein level noises measured in morphologically similar cells, suggesting that they have reduced intrinsic noise (Lehner, 2008). In our data where intrinsic and extrinsic noises are explicitly separated, we found D_{int} significantly lower for genes encoding protein complex members than other genes (**Fig. 2-4A**). The same is true for D'_{int} (**Fig. 2-4A**). By contrast, although D_{ext} is significantly lower for genes encoding protein complex members than other genes (**Fig. 2-4B**), this disparity becomes non-significant for D'_{ext} (**Fig. 2-4B**). Similar patterns were observed in the non-clonal cells (**Fig. A1-6**).

Potential mechanisms underlying the D_{ext} difference between genes encoding protein complex members and other genes can include a depletion of the TATA-box and an enrichment of miRNA targeting in the former group. Indeed, compared with other genes, those encoding protein complex members tend not to use the TATA-box (**Fig. 2-4C**), tend to be targeted by miRNAs (**Fig.2-4D**), and tend to be targeted by more miRNA species (**Fig. 2-4E**). The difference between genes encoding protein complex members and other genes in intrinsic noise after adjusting the presence/absence of TATA-box and the number of miRNA species targeting the gene by linear regression remains significant for both D_{int} ($P = 0.017$, Mann–Whitney U test) and D'_{int} ($P = 0.031$, Mann–Whitney U test), suggesting that other mechanisms also contribute to the lowered intrinsic noise of protein complex members.

2.4.8 Cell cycle genes have low intrinsic but high extrinsic noise

Cell cycle genes are those that control the cell cycle and hence should express differently at different cell cycle stages (Cho et al., 1998). However, within a cell that is at a cellular stage, cell cycle genes should preferably show consistent expressions. Thus, we predict that cell cycle genes have been selected to have low D_{int} but high D_{ext}. Indeed, compared with other genes, cell cycle genes show significantly lower D_{int} and D'_{int} (**Fig. 2-5A**), but significantly higher D_{ext} and D'_{ext} (**Fig. 2-5B**). This finding echoes the recent report that the genetic circuit underlying the biological clock often has an architecture to buffer the harmful internal fluctuation of signals while responding to the variation of the functional external stimuli (Pittayakanchit et al., 2018). The analysis of the non-clonal cells yielded similar results (**Fig. A1-7**).

Given the noise features of the cell cycle genes, we predict that they should be preferentially targeted by miRNAs, because miRNA targeting lowers the intrinsic noise but raises the extrinsic noise. In addition, we know that the impact of miRNAs on the intrinsic noise (but not necessarily the extrinsic noise) of a target rises with the number of miRNA species targeting the gene (**Fig. 2-2C**). We found that the fraction of genes targeted by miRNAs is not significantly higher for cell cycle genes than other genes (P = 0.30, Fisher's exact test; **Fig. 2-5C**), but the median number of miRNA species targeting a gene is significantly higher for cell cycle genes than other genes (P = 0.0071, Mann–Whitney U test; **Fig. 2-5D**). These observations suggest that miRNA targeting is not responsible for cell cycle genes' high D_{ext} but is responsible for their low D_{int}. Notwithstanding, we cannot rule out the possibility that the non-significant result in Fig. 5C is due to the relatively small sample size of cell cycle genes (n = 570, as opposed to 935 for genes encoding protein complex members and 1603 for genes functioning in the mitochondrion). After adjusting the number of miRNA species targeting a gene, we found that cell cycle genes still have lower D_{int} (P = 0.0057, Mann–Whitney U test) and D'_{int} (P = 0.0013, Mann–Whitney U test) than other genes, suggesting the existence of other factors contributing to the low intrinsic noise of cell cycle genes.

2.4.9 Other genes with exceptionally high or low extrinsic or intrinsic noise

To learn more about the biological implications of intrinsic and extrinsic noise, we performed gene ontology (GO) analysis on genes with extreme D_{ext} and/or D_{int} values. We first defined high D_{ext} genes as those genes whose D_{ext} values are in the highest 10% of all 3975 genes and low D_{ext} genes as those whose D_{ext} values are in the lowest 10% of all 3975 genes. We similarly defined high D_{int} genes and low D_{int} genes. These genes show enrichments of various functional categories (**Table 2-1**). For instance, both the high D_{ext} group and high D_{int} group are enriched with genes encoding secreted proteins and extracellular proteins. Secreted and extracellular proteins synthesized from many individual cells are mixed together and function outside the cells, so there is no need to reduce their expression noise at the mRNA level. Thus, their high noise likely reflects a lack of selection minimizing their noise. By contrast, the low D_{ext} group are enriched with genes whose products interact with RNAs, whereas the low D_{int} group are enriched with genes encoding phosphoproteins and proteins with coiled coil structure, again indicating that the biological implications of extrinsic noise and intrinsic noise can be different. Similar results were found for the non-clonal cells (**Table A1-1**).

We further examined genes with different combinations of extreme extrinsic and intrinsic noises (**Table 2-1** and **Table A1-1**). Specifically, we identified genes with both high D_{ext} and high D_{int}, high D_{ext} but low D_{int}, low D_{ext} but high D_{int}, and both low D_{ext} and low D_{int}, respectively. Here, a gene is considered to have high (or low) noise if its noise is ranked in the top (or bottom) 25% among the 3975 genes. As expected, the group with both high D_{ext} and high D_{int} is enriched with genes encoding secreted and extracellular proteins, while the group with high D_{ext} but low D_{int} is enriched with cell cycle genes. The group with low D_{ext} but high D_{int} is not enriched with any GO category. Finally, the group with both low D_{ext} and low D_{int} is enriched with genes encoding RNA-interacting proteins and phosphoproteins. The identification of genes with extreme noise values can help further understand the biological significance and constraints of intrinsic and extrinsic gene expression noises.

2.5 Discussion

Using allele-specific scRNA-seq, we performed the first genomic estimation of intrinsic and extrinsic expression noises of any species. The mRNA noise estimates obtained allowed us to evaluate the predicted effects of various factors. In particular, we found that (i) the presence of the TATA-box in the promoter of a gene increases both the intrinsic and extrinsic expression noise of the gene, (ii) miRNAs lower the intrinsic noise but increase the extrinsic noise of their target genes, (iii) the extrinsic noise of a gene increases with the total expression noise of its *trans*-regulator, and (iv) genes regulated by the same *trans*-regulator have more similar extrinsic expression noises than genes not co-regulated. Considering gene functions, we formulated hypotheses on natural selection for lowered or elevated intrinsic and/or extrinsic noise of groups of genes, and were able to find evidence supporting our hypotheses. Specifically, we predicted and then demonstrated that (nuclear) genes functioning in the mitochondrion have reduced extrinsic noise, genes encoding protein complex members have decreased intrinsic noise, and cell cycle genes have lowered intrinsic noise but elevated extrinsic noise.

It is valuable to compare our results with previous genome-wide studies of total protein or total mRNA expression noise. For example, a study in yeast showed that nuclear genes functioning in the mitochondrion have unusually high protein noise, presumably due to the random partition of mitochondria during cell division (Newman et al., 2006). Multiple studies reported that expression noise of nuclear genes functioning in the mitochondrion can result in large, presumably harmful among-cell variation in global gene expression (Das Neves et al., 2010; Dhar et al., 2019; Johnston et al., 2012). It was thus unclear whether the gene expression noise of nuclear genes functioning in mitochondrion has been subject to selective minimization. Our results on the mRNA expression noise of nuclear genes functioning in the mitochondrion provide clear evidence for the minimization. Our ability to detect this signal is likely because mRNAs are located in the cytoplasm so are not subject to the problem of block partition of mitochondrial proteins. Regarding genes encoding protein complex members, a previous study (Lehner, 2008) suggested that their low noise may be explained by one or more of the following reasons. First, protein complex members are enriched for essential genes and essential genes tend to have low noise. Second, protein complex members are more dosage-sensitive due to the requirement for dosage balance among members of the same complex. Third, the low noise of protein complex members is a by-product of their short protein half-lives. Our results do not support the first or third reason, because the first reason would predict both low extrinsic noise and low intrinsic noise, contrasting our observation of reduction in D_{int} but not D_{ext}, while the third reason would predict no reduction in the mRNA expression noise, contradictory to our observation of lowed D_{int}. With respect to cell cycle genes, no previous research has ever found them to have low expression noise despite the suggestion that cell cycle should be robust to biochemical noise (Li et al., 2004; Vilar et al., 2002). This is possibly because previous studies did not separate intrinsic from extrinsic noise, while cell cycle genes are expected to and indeed have low D_{int} but high D_{ext}. Regarding mRNA expression noise, several previous studies used scRNA-seq data. For instance, Wu et al. analyzed how histone modification independently modulates expression noise and mean expression level (Wu et al., 2017). Morgan et al. reported a correlation between CpG island and expression noise (Morgan and Marioni, 2018). In particular, allele-specific scRNA-seq has been used to characterize the technical noise versus biological noise (Kim et al., 2015) and estimate expression noise-related quantities such as transcriptional burst size and frequency (Jiang et al., 2017). Nevertheless, none of the previous studies used scRNA-seq to decompose expression noise into intrinsic and extrinsic noises.

Our analyses have several caveats that are worth discussion. First, although many of our statistical results are highly significant, the effect sizes of some factors appear small. This may be due to the high technical noises of scRNA-seq-based expression level measures (Marinov et al., 2014), which is further exacerbated in allele-specific scRNA-seq, because only reads containing information of the allele of origin, which constitute a small fraction of all reads, are useful to our analysis. The high technical noise introduces both random errors and systematic errors in our estimation of expression noise. Random errors are not expected to create spurious results in large samples (Hedge et al., 2018). By contrast, systematic errors may create spurious results. In our analysis, we removed known systematic errors from technical noises (Grün et al., 2014) by controlling for the number of reads per gene. Thus, the remaining errors in our estimation of intrinsic noise and extrinsic noise should be largely random, and these random errors have likely caused underestimation of effect sizes in our study. Furthermore, whether an effect is evolutionarily important depends on whether it is detectable by natural selection. Our observation of differential uses of various molecular mechanisms such as the TATA-box and miRNA targeting in the optimization of intrinsic and extrinsic noise levels demonstrates that the detected effects are important. Second, previous theoretical studies showed that noise decomposition using the dual reporter system is accurate under static environments but may not be accurate under dynamic environments; in the latter case, noise decomposition may not reveal the underlying mechanism (Hilfinger and Paulsson, 2011; Shahrezaei et al., 2008; Sherman et al., 2015). Notwithstanding, we found that the intrinsic and extrinsic noises estimated in this study largely follow expectations. More importantly, intrinsic and extrinsic noises do have different biological meanings and hence are differentially tuned evolutionarily. Hence, the noise decomposition appears biologically meaningful and useful. Third, a central topic about noise decomposition is the absolute magnitudes of intrinsic and extrinsic noises (Bar-Even et al., 2006; Elowitz et al., 2002; Raser and O'shea, 2004). As mentioned, because of the relatively large size of the technical noise from allele-specific scRNA-seq and different impacts of the technical noise on measures of intrinsic and extrinsic noises, it is impossible to compute and compare the absolute magnitudes of intrinsic and extrinsic noises. This limitation forced us to use rank-based statistics, which made it difficult to estimate absolute effect sizes of various factors..

Fourth, our study focused on mRNA expression noise, but one might argue that mRNA noise does not directly correspondent to protein noise. We believe that this should not be an issue, because of substantial evidence that mRNA noise is the major source of protein noise (Bar-Even et al., 2006; Batada and Hurst, 2007; Fraser et al., 2004; Raj et al., 2006; Sherman et al., 2015). Finally, to obtain reliable noise estimates, we filtered out genes with low average read counts. Therefore, our conclusions mainly apply to genes with moderate or high expression levels. Because the expressions of lowly expressed genes are impacted most by noise (Bar-Even et al., 2006), it will be important to study intrinsic and extrinsic noises of lowly expressed genes in the future.

In sum, our study performed the first genome-scale estimation of intrinsic and extrinsic gene expression noise at the mRNA level. We demonstrated the general reliability of our noise estimates and illustrated the utility of these estimates for understanding the mechanisms controlling and selections on the two noise components. Our findings may have implications for synthetic biology, where one often needs to design genetic circuits that have robust yet dynamic behaviors. For example, the detailed mechanisms that cells employ to allow cell cycle genes to have high extrinsic noise but low intrinsic noise may provide insights for designing oscillators that are sensitive to different cell states yet are robust to intrinsic noise (Elowitz and Leibler, 2000; Fung et al., 2005; Potvin-Trottier et al., 2016).

2.6 References

Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., and Clapham, P. (2016). Ensembl 2017. Nucleic Acids Res *45*, D635-D642.

Bahar, R., Hartmann, C.H., Rodriguez, K.A., Denny, A.D., Busuttil, R.A., Dollé, M.E., Calder, R.B., Chisholm, G.B., Pollock, B.H., and Klein, C.A. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature *441*, 1011.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. Nat Genet *38*, 636.

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20-51.

Batada, N.N., and Hurst, L.D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet *39*, 945.

Birchler, J.A., and Veitia, R.A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proceedings of the National Academy of Sciences, 201207726.

Blake, W.J., Balázsi, G., Kohanski, M.A., Isaacs, F.J., Murphy, K.F., Kuang, Y., Cantor, C.R., Walt, D.R., and Collins, J.J. (2006). Phenotypic consequences of promoter-mediated transcriptional noise. Mol Cell *24*, 853-865.

Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature *453*, 544.

Chen, X., and Zhang, J. (2016). The genomic landscape of position effects on protein expression level and noise in yeast. Cell systems *2*, 347-354.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., and Lockhart, D.J. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular cell *2*, 65-73.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442-459. e420.

Das Neves, R.P., Jones, N.S., Andreu, L., Gupta, R., Enver, T., and Iborra, F.J. (2010). Connecting variability in global transcription rate to mitochondrial variability. PLoS Biol *8*, e1000560.

Dhar, R., Missarova, A.M., Lehner, B., and Carey, L.B. (2019). Single cell functional genomics reveals the importance of mitochondria in cell-to-cell phenotypic variation. eLife *8*.

Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., and Bucher, P. (2016). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic acids research *45*, D51-D55.

Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature *403*, 335.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Faure, A.J., Schmiedel, J.M., and Lehner, B. (2017). Systematic analysis of the determinants of gene expression noise in embryonic stem cells. Cell systems *5*, 471-484. e474.

Fraser, H.B., Hirsh, A.E., Giaever, G., Kumm, J., and Eisen, M.B. (2004). Noise minimization in eukaryotic gene expression. PLoS biology *2*, e137.

Fu, A.Q., and Pachter, L. (2016). Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. Stat Appl Genet Mol Biol *15*, 447-471.

Fung, E., Wong, W.W., Suen, J.K., Bulter, T., Lee, S.-g., and Liao, J.C. (2005). A synthetic gene–metabolic oscillator. Nature *435*, 118.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nature methods *11*, 637.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav Res Methods *50*, 1166-1186.

Hilfinger, A., and Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proceedings of the National Academy of Sciences *108*, 12167-12172.

Hornung, G., Bar-Ziv, R., Rosin, D., Tokuriki, N., Tawfik, D.S., Oren, M., and Barkai, N. (2012a). Noise–mean relationship in mutated promoters. Genome research.

Hornung, G., Oren, M., and Barkai, N. (2012b). Nucleosome organization affects the sensitivity of gene expression to promoter mutations. Molecular cell *46*, 362-368.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research *37*, 1-13.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. Development *136*, 3853-3862.

Jajoo, R., Jung, Y., Huh, D., Viana, M.P., Rafelski, S.M., Springer, M., and Paulsson, J. (2016).

Accurate concentration control of mitochondria and nucleoids. Science *351*, 169-172.
Jiang, Y., Zhang, N.R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biol *18*, 74.

Johnston, I.G., Gaal, B., das Neves, R.P., Enver, T., Iborra, F.J., and Jones, N.S. (2012).

Mitochondrial variability as a source of extrinsic cellular noise. PLoS Comput Biol *8*, e1002416.

Kemkemer, R., Schrank, S., Vogel, W., Gruler, H., and Kaufmann, D. (2002). Increased noise as an effect of haploinsufficiency of the tumor-suppressor gene neurofibromatosis type 1 in vitro. Proceedings of the National Academy of Sciences *99*, 13783-13788.

Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nature communications *6*, 8687.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512.

Lehner, B. (2008). Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol Syst Biol *4*, 170.

Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed. Proceedings of the National Academy of Sciences *101,* 4781-4786.

Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., and Lei, Y. (2011). Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol *12,* R83.

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database *2015*.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research *5*.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res *24*, 496-510.

Morgan, M.D., and Marioni, J.C. (2018). CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. Genome Biol *19*, 81.

Murphy, K.F., Adams, R.M., Wang, X., Balazsi, G., and Collins, J.J. (2010). Tuning and controlling gene expression noise in synthetic gene networks. Nucleic Acids Res *38*, 2712-2726.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.

Papp, B., Pal, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194.

Paulsson, J. (2004). Summing up the noise in gene networks. Nature *427*, 415.

Pedraza, J.M., and van Oudenaarden, A. (2005). Noise propagation in gene networks. Science *307*, 1965-1969.

Pittayakanchit, W., Lu, Z., Chew, J., Rust, M.J., and Murugan, A. (2018). Biophysical clocks face a trade-off between internal and external noise resistance. eLife *7*, e37624.

Potvin-Trottier, L., Lord, N.D., Vinnicombe, G., and Paulsson, J. (2016). Synchronous long-term oscillations in a synthetic gene circuit. Nature *538*, 514.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol *4*, e309.

Raser, J.M., and O'shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811-1814.

Raser, J.M., and O'shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.

Reinius, B., Mold, J.E., Ramsköld, D., Deng, Q., Johnsson, P., Michaëlsson, J., Frisén, J., and Sandberg, R. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet.

Rhee, A., Cheong, R., and Levchenko, A. (2014). Noise decomposition of intracellular biochemical signaling networks using nonequivalent reporters. Proceedings of the National Academy of Sciences *111*, 17330-17335.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2009). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic acids research *38*, D497-D501.

Schmiedel, J., Marks, D.S., Lehner, B., and Bluthgen, N. (2017). Noise control is a primary function of microRNAs and post-transcriptional regulation. bioRxiv, 168641.

Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D.S., and van Oudenaarden, A. (2015). MicroRNA control of protein expression noise. Science *348*, 128-132.

Shahrezaei, V., Ollivier, J.F., and Swain, P.S. (2008). Colored extrinsic fluctuations and stochastic gene expression. Mol Syst Biol *4*, 196.

Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z., and Segal, E. (2014). Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res *24*, 1698-1706.

Sherman, M.S., Lorenz, K., Lanier, M.H., and Cohen, B.A. (2015). Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. Cell systems *1*, 315-325.

Söllner, J.F., Leparc, G., Hildebrandt, T., Klein, H., Thomas, L., Stupka, E., and Simon, E.

(2017). An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Scientific data *4*, 170185.

Stewart-Ornstein, J., Weissman, J.S., and El-Samad, H. (2012). Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol Cell *45*, 483-493.

Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nature methods *14*, 381.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proceedings of the National Academy of Sciences *99*, 12795-12800.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533-538.

Tirosh, I., and Barkai, N. (2008). Two strategies for gene regulation by promoter nucleosomes. Genome research.

Turing, A.M. (1952). The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences *237*, 37-72.

Veening, J.-W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol *62*, 193-210.

Vilar, J.M., Kueh, H.Y., Barkai, N., and Leibler, S. (2002). Mechanisms of noise-resistance in genetic oscillators. Proceedings of the National Academy of Sciences *99*, 5988-5992.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol *34*, 1145.

Wang, Z., and Zhang, J. (2011). Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proceedings of the National Academy of Sciences *108*, E67-E76.

Wu, S., Li, K., Li, Y., Zhao, T., Li, T., Yang, Y.-F., and Qian, W. (2017). Independent regulation of gene expression level and noise by histone modifications. PLoS Comput Biol *13*, e1005585.

Xie, W., Nagarajan, S., Baumgart, S.J., Kosinsky, R.L., Najafova, Z., Kari, V., Hennion, M., Indenbirken, D., Bonn, S., and Grundhoff, A. (2017). RNF40 regulates gene expression in an epigenetic context-dependent manner. Genome Biol *18*, 32.

Xu, H., Liu, J.-J., Liu, Z., Li, Y., Jin, Y.-S., and Zhang, J. (2019). Synchronization of stochastic expressions drives the clustering of functionally related genes. Science Advances *in press*.

Zhang, Z., Qian, W., and Zhang, J. (2009). Positive selection for elevated gene expression noise in yeast. Mol Syst Biol *5*, 299.

Zoller, B., Nicolas, D., Molina, N., and Naef, F. (2015). Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol Syst Biol *11*.

Table 2-1 Significantly enriched GO terms among genes with extreme intrinsic and/or extrinsic expression noise in clone 7. The three most significant terms are presented if more than three terms are significantly enriched.

GO terms	Corrected *P*-value
High extrinsic noise	
Secreted	3.5×10^{-12}
Extracellular region	1.5×10^{-11}
Signal peptide	7.3×10^{-9}
Low extrinsic noise	
Poly (A) RNA binding	6.7×10^{-7}
RNA binding	1.2×10^{-6}
rRNA processing	1.8×10^{-6}
High intrinsic noise	
Extracellular region	1.4×10^{-8}
Signal peptide	3.1×10^{-8}
Disulfide bond	1.0×10^{-7}
Low intrinsic noise	
Phosphoprotein	1.9×10^{-6}
Coiled coil	4.6×10^{-6}
High extrinsic noise and high intrinsic noise	
Signal peptide	1.2×10^{-13}
Secreted	4.9×10^{-13}
Extracellular region	2.1×10^{-12}
High extrinsic noise and low intrinsic noise	
Cell cycle	0.01
Low extrinsic noise and low intrinsic noise	
Poly (A) RNA binding	1.7×10^{-24}
Nucleus	1.1×10^{-8}
Nucleolus	1.9×10^{-8}

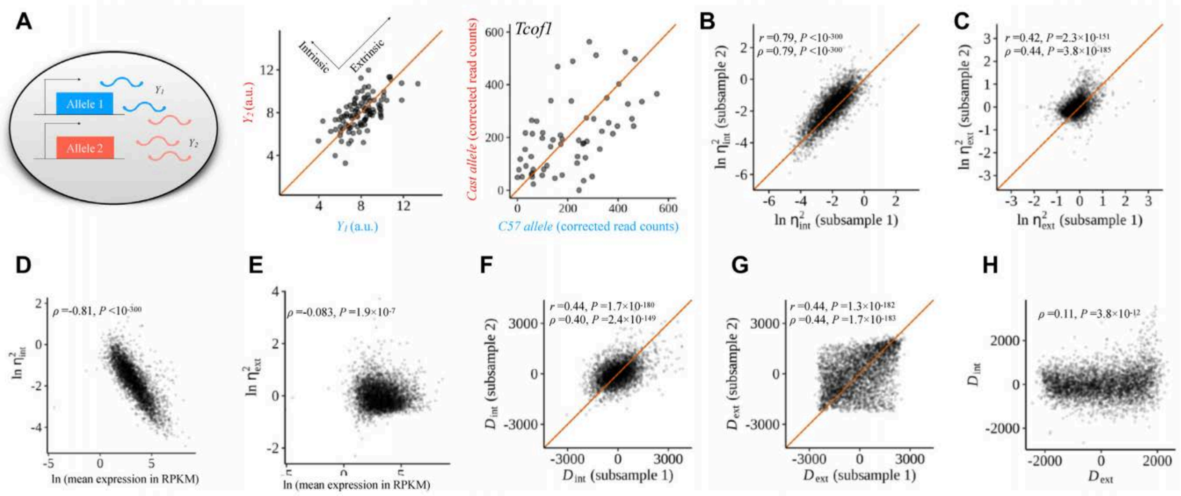

Fig. 2-1 Decomposition of gene expression noise into intrinsic and extrinsic noise. **(A)** Gene expression noise can be decomposed to its intrinsic and extrinsic components by the dual reporter assay, where two reporters represented respectively by the blue and orange boxes are controlled by independent, identical promoters. When plotting the expression level of one reporter against that of the other in each cell, the spread along the diagonal represents extrinsic noise, whereas the spread orthogonal to the diagonal represents intrinsic noise. Y_1 and Y_2 are the expression levels of the two reporters, respectively. The left plot shows hypothetical data from a gene, whereas the right plot presents the spike-in adjusted read-counts of two alleles of *Tcof1* from individual cells. **(B)** Intrinsic noises (η^2_{int}) estimated from two sub-samples of clone 7 are highly correlated with each other. Ln-transformed η^2_{int} is shown. Each dot is a gene. The orange line shows the diagonal. **(C)** Extrinsic noises (η^2_{ext}) estimated from two sub-samples of clone 7 are moderately correlated with each other. Ln-transformed η^2_{ext} is shown. Each dot is a gene. The orange line shows the diagonal. **(D)** The intrinsic expression noise of a gene is strongly negatively correlated with the mean expression level of the gene. Expression level is measured by Reads Per Kilobase of transcript per Million mapped reads (RPKM). **(E)** The extrinsic expression noise of a gene is weakly negatively correlated with the mean expression level of the gene. Because the extrinsic noise could be negative (see Materials and Methods), we added a small value (0.1 - the minimum of computed extrinsic noise) to all η^2_{ext} values before taking the natural log. **(F)** Intrinsic noise estimates adjusted for mean expression level and technical noise (D_{int}) are significantly correlated between two sub-samples of clone 7. The orange line shows the diagonal. **(G)** Extrinsic noise estimates adjusted for mean expression level and technical noise (D_{ext}) are significantly correlated between two sub-samples of clone 7. The orange line shows the diagonal. **(H)** D_{int} and D_{ext} are positively correlated.

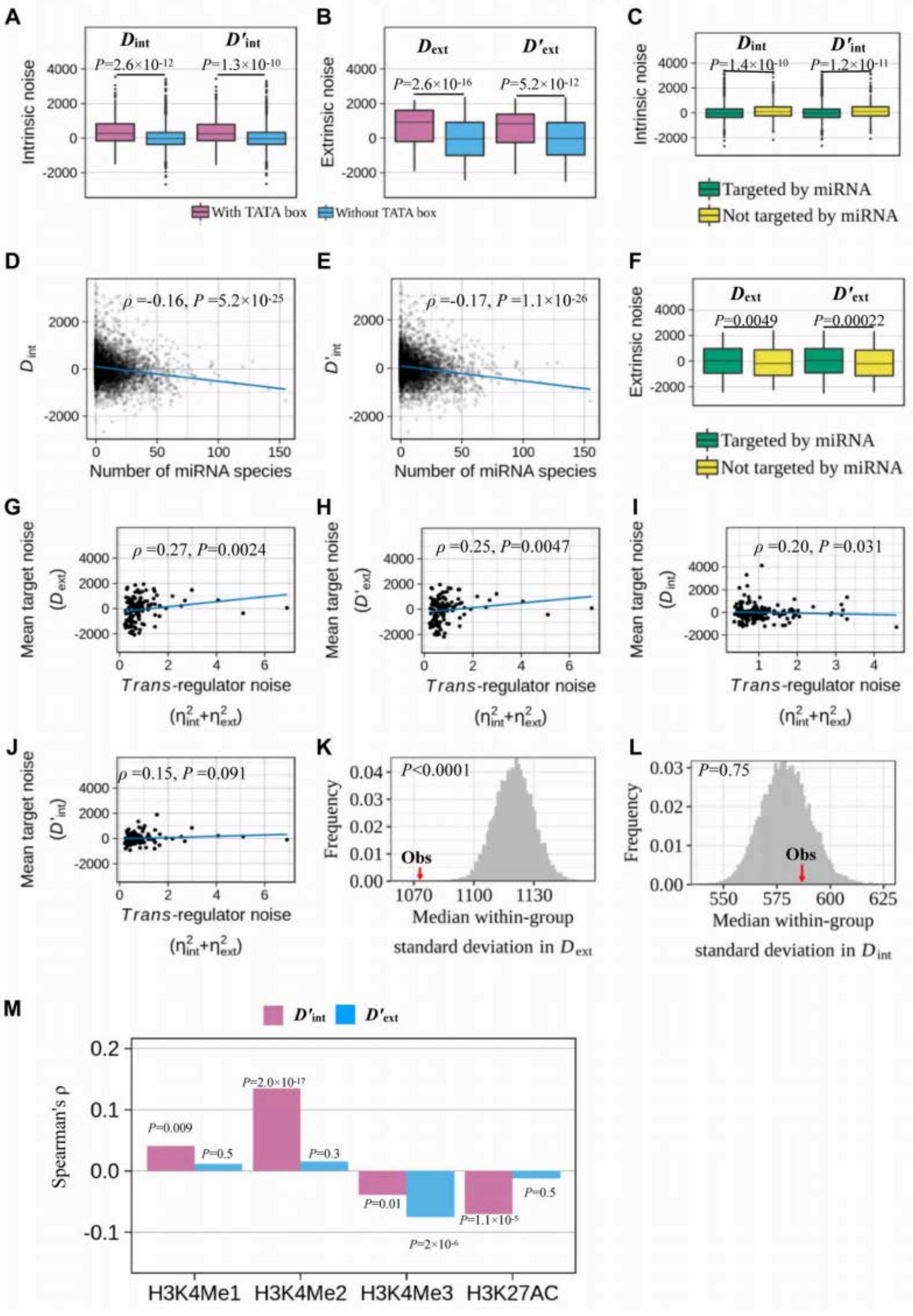
A
Intrinsic noise
D_{int}
D'_{int}
$P=2.6\times10^{-12}$
$P=1.3\times10^{-10}$
4000
2000
0
-2000
With TATA box
Without TATA box
B
Extrinsic noise
D_{ext}
D'_{ext}
$P=2.6\times10^{-16}$
$P=5.2\times10^{-12}$
C
Intrinsic noise
D_{int}
D'_{int}
$P=1.4\times10^{-10}$
$P=1.2\times10^{-11}$
Targeted by miRNA
Not targeted by miRNA
D
D_{int}
$\rho=-0.16$, $P=5.2\times10^{-25}$
Number of miRNA species
0
50
100
150
E
D'_{int}
$\rho=-0.17$, $P=1.1\times10^{-26}$
Number of miRNA species
F
Extrinsic noise
D_{ext}
D'_{ext}
$P=0.0049$
$P=0.00022$
Targeted by miRNA
Not targeted by miRNA
G
Mean target noise (D_{ext})
$\rho=0.27$, $P=0.0024$
Trans-regulator noise
($\eta^2_{int}+\eta^2_{ext}$)
H
Mean target noise (D'_{ext})
$\rho=0.25$, $P=0.0047$
Trans-regulator noise
($\eta^2_{int}+\eta^2_{ext}$)
I
Mean target noise (D_{int})
$\rho=0.20$, $P=0.031$
Trans-regulator noise
($\eta^2_{int}+\eta^2_{ext}$)
J
Mean target noise (D'_{int})
$\rho=0.15$, $P=0.091$
Trans-regulator noise
($\eta^2_{int}+\eta^2_{ext}$)
K
Frequency
$P<0.0001$
Obs
0.04
0.03
0.02
0.01
0.00
1070
1100
1130
Median within-group standard deviation in D_{ext}
L
Frequency
$P=0.75$
Obs
0.03
0.02
0.01
0.00
550
575
600
625
Median within-group standard deviation in D_{int}
M
D'_{int}
D'_{ext}
Spearman's ρ
0.2
0.1
0.0
-0.1
$P=0.009$
$P=0.5$
$P=2.0\times10^{-17}$
$P=0.3$
$P=0.01$
$P=2\times10^{-6}$
$P=1.1\times10^{-5}$
$P=0.5$
H3K4Me1
H3K4Me2
H3K4Me3
H3K27AC

Fig. 2-2 Factors influencing intrinsic and/or extrinsic gene expression noise. **(A)** Genes with a TATA-box in the promoter (pink) have significantly higher intrinsic noise (D_{int}) than genes without a TATA-box (blue). The same is true when intrinsic noise is measured by D'_{int}, which is uncorrelated with extrinsic noise. The lower and upper edges of a box represent the first (qu_1) and third (qu3) quartiles, respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Genes with a TATA-box in the promoter (pink) have significantly higher extrinsic noise (D_{ext}) than genes without a TATA-box (blue). The same is true when extrinsic noise is measured by D'_{ext}, which is uncorrelated with intrinsic noise.**(C)** Genes targeted by miRNA (green) have significantly lower intrinsic noise (D_{int} and D'_{int}) than genes not targeted by miRNA (yellow). **(D)** Genes targeted by more miRNA species have lower D_{int}. The blue line displays the linear regression of D_{int} of a target gene on the number of miRNA species targeting it. **(E)** Genes targeted by more miRNA species have lower D'_{int}. The blue line displays the linear regression of D'_{int} of a target gene on the number of miRNA species targeting it. **(F)** Genes targeted by miRNA (green) have significantly higher extrinsic noise (D_{ext} and D'_{ext}) than genes not targeted by miRNA (yellow). **(G)** The mean extrinsic noise (D_{ext}) of genes targeted by the same *trans*-regulator is significantly positively correlated with the total noise ($\eta_{\mathrm{int}}^2 + \eta_{\mathrm{ext}}^2$) of the *trans*-regulators. **(H)** The mean extrinsic noise (upon the control for intrinsic noise) (D'_{ext}) of genes targeted by the same *trans*-regulator is significantly positively correlated with the total noise ($\eta_{\mathrm{int}}^2 + \eta_{\mathrm{ext}}^2$) of the *trans*-regulators. **(I)** The mean intrinsic noise (D_{int}) of genes targeted by the same *trans*-regulator is significantly positively correlated with the total noise ($\eta_{\mathrm{int}}^2 + \eta_{\mathrm{ext}}^2$) of the *trans*-regulator. **(J)** The mean intrinsic noise (upon the control for extrinsic noise) (D'_{int}) of genes targeted by the same *trans*-regulator is not significantly positively correlated with the total noise ($\eta_{\mathrm{int}}^2 + \eta_{\mathrm{ext}}^2$) of the *trans*-regulator. **(K)** The observed median standard deviation of D_{ext} among genes regulated by the same *trans*-regulator (red arrow) is significantly smaller than the random expectation (histograms). **(L)** The observed median standard deviation of D_{int} among genes regulated by the same *trans*-regulator is not significantly different from the random expectation (histograms). **(M)** Spearman's correlation between the number of histone modification peaks that overlap the core promoter and intrinsic or extrinsic noise.

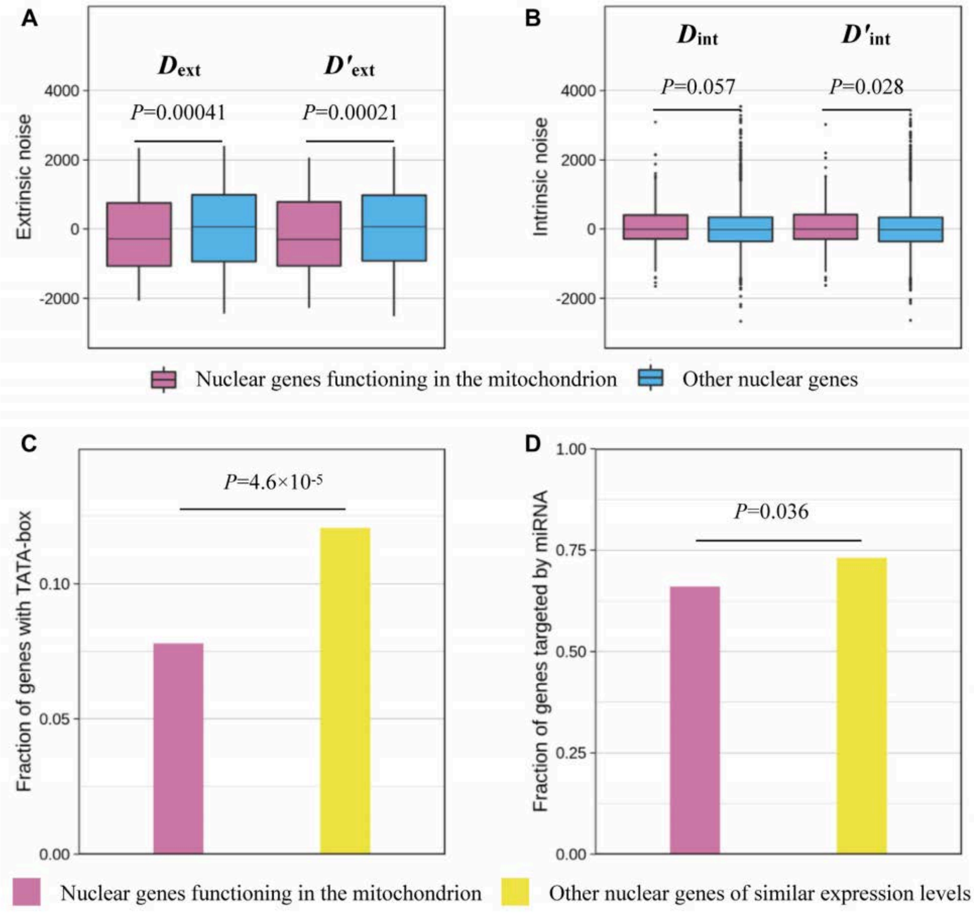

Fig. 2-3 Nuclear genes functioning in the mitochondrion have lower extrinsic noise but not lower intrinsic noise when compared with other genes. **(A)** Nuclear genes functioning in the mitochondrion (pink) have significantly lower extrinsic noise (D_{ext} and D'_{ext}) than other genes (blue). The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Nuclear genes functioning in the mitochondrion (pink) do not have significantly lower intrinsic noise D_{int} and even have significantly higher D'_{int} than other genes (blue). **(C)** TATA-box is underrepresented in the promoters of nuclear genes functioning in the mitochondrion (pink) when compared with other genes of similar expression levels (yellow). **(D)** Nuclear genes functioning in the mitochondrion (pink) are less targeted by miRNAs than other genes with similar expression levels (yellow).

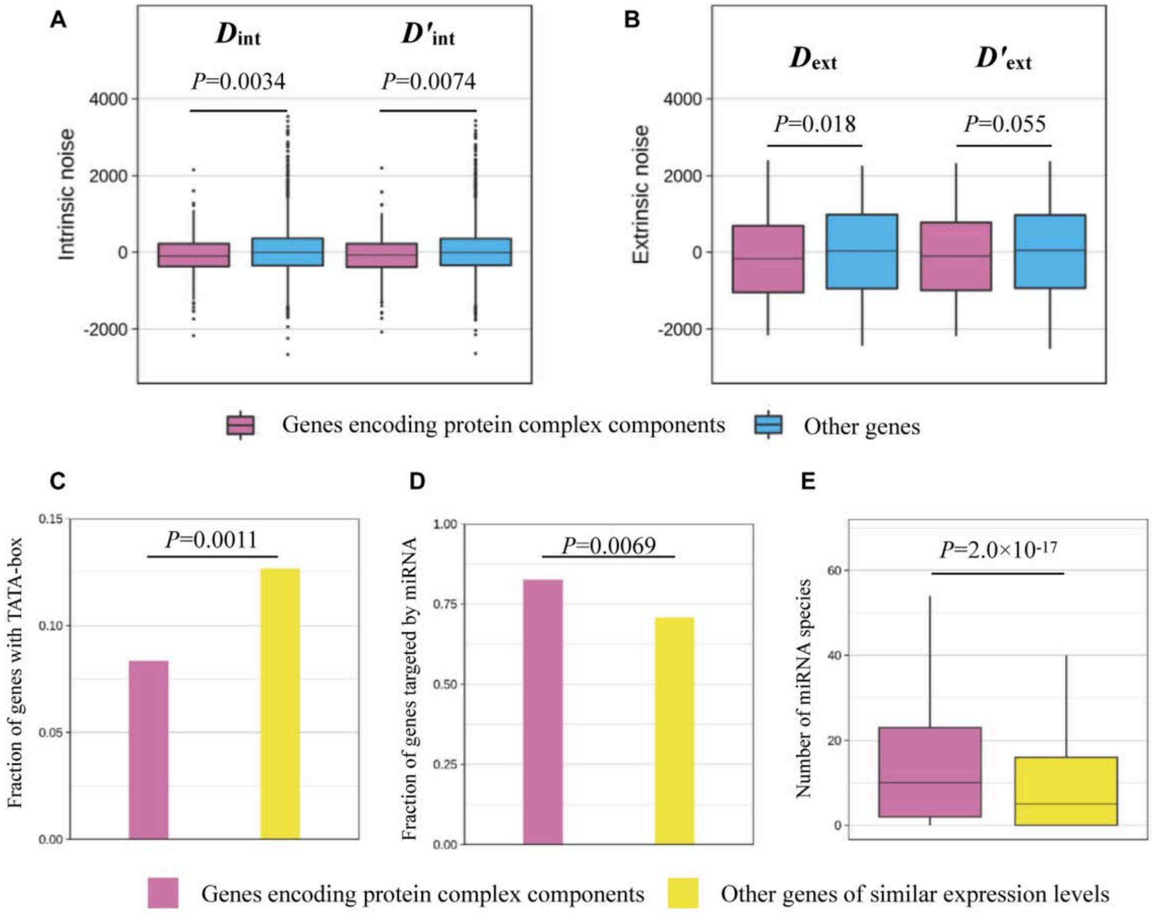

Fig. 2-4 Genes encoding protein complex components have lower intrinsic noise but not lower extrinsic noise than other genes. **(A)** Genes encoding protein complex components (pink) have significantly lower intrinsic noise (D_{int} and D'_{int}) than other genes (blue). The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Genes encoding protein complex components (pink) have significantly lower D_{ext} but not significantly lower D'_{ext} than other genes (blue). **(C)** TATA-box is underrepresented in the promoters of genes encoding protein complex components (pink) when compared with other genes of similar expression levels (yellow). **(D)** Genes encoding protein complex components (pink) are more likely to be targeted by miRNAs when compared with other genes of similar expression levels (yellow). **(E)** Genes encoding protein complex components (pink) tend to be targeted by more miRNA species when compared with other genes of similar expression levels (yellow).

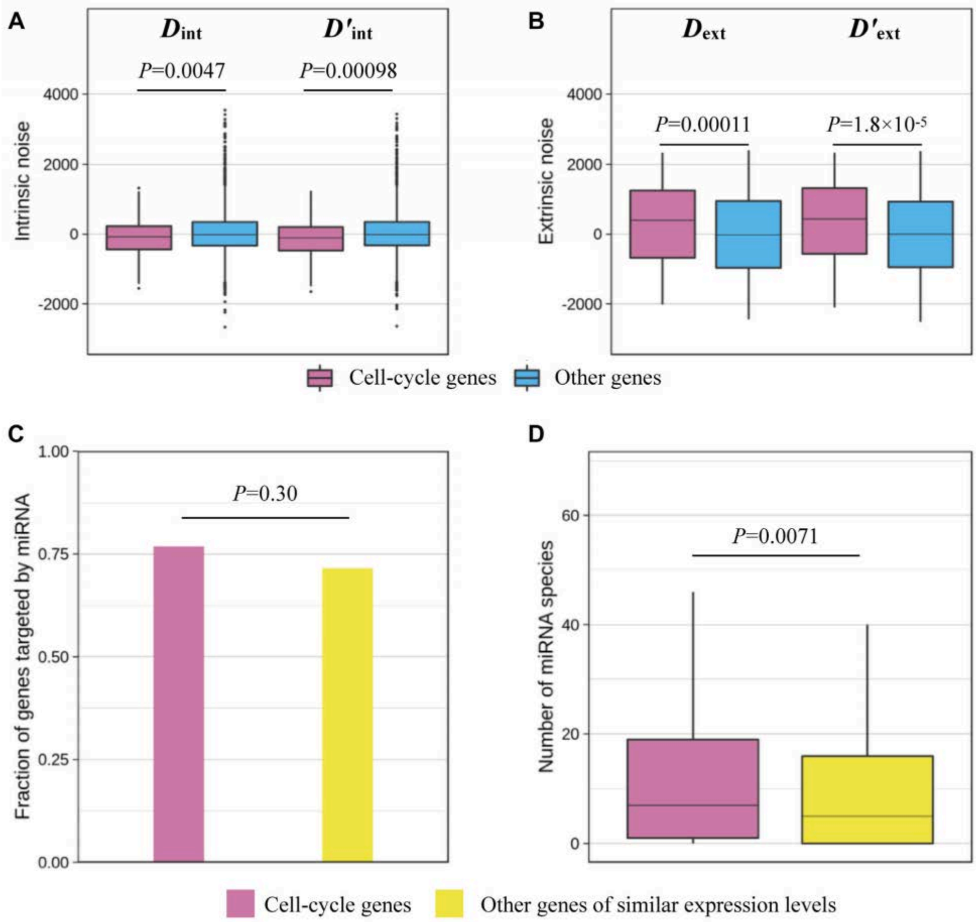

Fig. 2-5 Cell cycle genes have lower intrinsic noise but higher extrinsic noise than other genes. **(A)** Cell cycle genes (pink) have significantly lower intrinsic noise (D_{int} and D'_{int}) when compared with other genes (blue). The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Cell cycle genes (pink) have significantly higher extrinsic noise (D_{ext} and D'_{ext}) when compared with other genes. **(C)** Fraction of genes targeted by miRNAs is not significantly different between cell cycle genes (pink) and other genes of similar expression levels (yellow). **(D)** Cell cycle genes (pink) tend to be targeted by more miRNA species than other genes of similar expression levels (yellow).

Chapter 3: Chromosome-Wide Co-Fluctuation of Stochastic Gene Expression in Mammalian Cells

No man is an island.

-John Donne

3.1 Abstract

Gene expression is subject to stochastic noise, but to what extent and by which means such stochastic variations are coordinated among different genes are unclear. We hypothesize that neighboring genes on the same chromosome co-fluctuate in expression because of their common chromatin dynamics, and verify it at the genomic scale using allele-specific single-cell RNA-sequencing data of mouse cells. Unexpectedly, the co-fluctuation extends to genes that are over 60 million bases apart. We provide evidence that this long-range effect arises in part from chromatin co-accessibilities of linked loci attributable to three-dimensional proximity, which is much closer intra-chromosomally than inter-chromosomally. We further show that genes encoding components of the same protein complex tend to be chromosomally linked, likely resulting from natural selection for intracellular among-component dosage balance. These findings have implications for both the evolution of genome organization and optimal design of synthetic genomes in the face of gene expression noise.

3.2 Introduction

Gene expression is subject to considerable stochasticity that is known as expression noise, formally defined as the expression variation of a given gene among isogenic cells in the same environment (Blake et al., 2003; Elowitz et al., 2002; Raser and O'shea, 2005). Gene expression noise is a double-edged sword. On the one hand, it can be deleterious because it leads to imprecise controls of cellular behavior, including, for example, destroying the stoichiometric relationship among functionally related proteins and disrupting homeostasis (Bahar et al., 2006; Batada and Hurst, 2007; Kemkemer et al., 2002; Lehner, 2008; Wang and Zhang, 2011). On the other hand, gene expression noise can be beneficial. For instance, unicellular organisms may exploit gene expression noise to employ bet-hedging strategies in fluctuating environments (Veening et al., 2008; Zhang et al., 2009), whereas multicellular organisms can make use of expression noise to initiate developmental processes (Chang et al., 2008; Huang, 2009; Turing, 1952).

By quantifying protein concentrations in individual isogenic cells cultured in a common environment, researchers have measured the expression noise for thousands of genes in the bacterium *Escherichia coli* (Taniguchi et al., 2010) and unicellular eukaryote *Saccharomyces cerevisiae* (Newman et al., 2006). Nevertheless, because genes are not in isolation, one wonders whether and to what extent expression levels co-vary among genes at a steady state, which unfortunately cannot be studied by the above data. By simultaneously tagging two genes with different florescent markers, Stewart-Ornstein et al. discovered strong co-fluctuation of the concentrations of some functionally related proteins in yeast such as those involved in the Msn2/4 stress response pathway, amino acid synthesis, and mitochondrial maintenance, respectively(Stewart-Ornstein et al., 2012), and the expression co-fluctuation of these genes is facilitated by their sharing of transcriptional regulators (Stewart-Ornstein et al., 2013).

Here we explore yet another mechanism for expression co-fluctuation. We hypothesize that, due to the sharing of chromatin dynamics (Raj and van Oudenaarden, 2008), a key contributor to gene expression noise (Brown et al., 2013; Raj and van Oudenaarden, 2008; Sanchez et al., 2013), genes that are closely linked on the same chromosome should exhibit a stronger expression co-fluctuation when compared with genes that are not closely linked or unlinked (**Fig.3-1**). We refer to this potential influence of chromosomal linkage of two genes on their expression co-fluctuation as the linkage effect. The linkage-effect hypothesis is supported by two pioneering studies demonstrating that the correlation in expression level between two reporter genes across isogeneic cells in the same environment is much higher when they are placed next to each other on the same chromosome than when they are placed on separate chromosomes (Becskei et al., 2005; Raj et al., 2006). However, neither the generality of the linkage effect nor the chromosomal proximity required for this effect are known. Furthermore, the biological significance of the linkage effect and its potential impact on genome organization and evolution have not been investigated. In this study, we address these questions by analyzing allele-specific single-cell RNA-sequencing (RNA-seq) data from mouse cells (Reinius et al., 2016). We demonstrate that the linkage effect is not only general but also long-range, extending to gene pairs that are tens of millions of bases apart. We provide evidence that three-dimensional (3D) chromatin proximities are responsible for the long-range co-fluctuation through mediating chromatin accessibility covariations. Finally, we show theoretically and empirically that the linkage effect has likely impacted the evolution of the chromosomal locations of genes encoding members of the same protein complex.

3.3 Results

3.3.1 Linkage effect on gene expression co-fluctuation is general and long-range

Let us consider two genes *A* and *B* each with two alleles respectively named 1 and 2 in a diploid cell. When A and B are chromosomally linked, without loss of generality, we assume that A_1 and B_1 are on the same chromosome whereas A_2 and B_2 are on its homologous chromosome (**Fig.3-2A**). Expression co-fluctuation between one allele of A and one allele of B (e.g., A_1 and B_2) is measured by Pearson's correlation (r_e, where the subscript "e" stands for expression) between the expression levels of the two alleles across isogenic cells under the same environment. Among the four possible pairs of alleles A_1-B_1, A_2-B_2, A_1-B_2, and A_2-B_1, the former two pairs are physically linked whereas the latter two pairs are unlinked. The linkage-effect hypothesis asserts that, at a steady state, expression correlations between linked alleles (*cis*-correlations) are greater than those between unlinked alleles (*trans*-correlations). That is, $\delta_{\mathrm{e}} = [r_{\mathrm{e}}(A_1, B_1) + r_{\mathrm{e}}(A_2, B_2) - r_{\mathrm{e}}(A_1, B_2) - r_{\mathrm{e}}(A_2, B_1)]/2 > 0$. Note that this formulation is valid regardless of whether the two alleles of the same gene have equal mean expression levels. While each of the four correlations could be positive or negative, in the large data analyzed below, they are mostly positive and show approximately normal distributions across gene pairs examined.

To verify the above prediction about δ_{e}, we analyzed a single-cell RNA-seq dataset of fibroblast cells derived from a hybrid between two mouse strains (CAST/EiJ × C57BL/6J) (Reinius et al., 2016). Single-cell RNA-seq profiles the transcriptomes of individual cells, allowing quantifying stochastic gene expression variations among isogenic cells in the same environment (Hashimshony et al., 2016; Macosko et al., 2015; Picelli et al., 2014). DNA polymorphisms in the hybrid allow estimation of the expression level of each allele for thousands of genes per cell. The dataset includes data from seven fibroblast clones and some non-clonal fibroblast cells of the same genotype. We focused our analysis on clone 7 (derived from the hybrid of CAST/EiJ male × C57BL/6J female) in the dataset, because the number of cells sequenced in this clone is the largest (n = 60) among all clones. We excluded from our analysis all genes on Chromosomes 3 and 4 due to aneuploidy in this clone and X-linked genes due to X inactivation. To increase the sensitivity of our analysis and remove imprinted genes, we focused on the 3405 genes that have at least 10 RNA-seq reads mapped to each of the two alleles. These genes form 3404×3405/2 = 5,795,310 gene pairs, among which 377,584 pairs are chromosomally linked.

For each pair of chromosomally linked genes, we computed their δ_e by treating the allele from CAST/EiJ as allele 1 and that from C57BL/6J as allele 2 at each locus. The fraction of gene pairs with $\delta_e > 0$ is 0.61 (**Fig.3-2B**). As shown by the 95% confidence intervals, this trend is significantly higher than null expectation. Because a gene can appear in multiple gene pairs, the δ_e from all pairs might not be fully independent. To be conservative, we further applied binomial test in a subset of gene pairs where each gene appears only once. Specifically, we randomly shuffled the orders of all genes on each chromosome and considered from one end of the chromosome to the other end non-overlapping consecutive windows of two genes. The result is still significantly exceeding the null expectation of 0.5 ($P < 2.4\times10^{-16}$, binomial test). That most gene pairs exhibit $\delta_e > 0$ holds in each of the 17 chromosomes examined, with the trend being statistically significant in 6 chromosomes even using the very conservative test as described above (nominal $P < 0.05$; **Fig.3-2C**). As a negative control, we analyzed gene pairs located on different chromosomes, treating alleles the same way as described above. As expected, this time the fraction of gene pairs with $\delta_e > 0$ is not significantly different from 0.5 ($P = 0.25$; **Fig.3-2B**). The fraction of gene pairs with $\delta_e > 0$ appears to vary among chromosomes (**Fig.3-2C**). To assess the significance of this variation, we compared the fraction of independent gene pairs with $\delta_e > 0$ between every two chromosomes by Fisher's exact test. After correcting for multiple testing, we found no significant difference between any two chromosomes.

To examine the generality of the findings from clone 7, we also analyzed clone 6 (derived from the hybrid of C57BL/6J female × CAST/EiJ male), which has 38 cells with

RNA-seq data. In the supplementary material of reference 23(Reinius et al., 2016) , the authors mentioned that 10 cells of clone 6 are aneuploidy for different chromosomes. We therefore removed these 10 cells. Similar results were obtained (**Fig.A2-1A** and **A2-1B**). Because clone 6 was from a male whereas clone 7 was from a female, our results apparently apply to both sexes. We also analyzed 47 non-clonal fibroblast cells with the same genetic background (cell IDs from 124 to 170, derived from the hybrid of C57BL/6J female × CAST/EiJ male), and obtained similar results (**Fig.A2-1C** and **Fig.A2-1D**). These findings establish that the linkage effect on expression co-fluctuation is neither limited to a few genes in a specific clone nor an epigenetic artifact of clonal cells, but is general. The linkage effect on co-fluctuation (and the decrease of the effect with genomic distance shown below) is robust to the definition of δ_e, because similar results are obtained when correlation coefficients are replaced with squares of correlation coefficients in the definition of δ_e.

We next investigated how close two genes need to be on the same chromosome for them to co-fluctuate in expression. We divided all pairs of chromosomally linked genes into 100 equal-interval bins based on the genomic distance between genes, defined by the number of nucleotides between their transcription start sites (TSSs). The median δ_e in a bin is found to decrease with the genomic distance represented by the bin (**Fig.3-2D**). Furthermore, even for the unbinned data, δ_e for a pair of linked genes correlates negatively with their genomic distance (Spearman's ρ = -0.029). To assess the statistical significance of this negative correlation, we randomly shuffled the genomic coordinates of genes within chromosomes and recomputed the correlation. This was repeated 1000 times and none of the 1000 ρ values were equal to or more negative than the observed ρ. Hence, the linkage effect on expression co-fluctuation of two linked genes weakens significantly with their genomic distance ($P <$ 0.001).

Surprisingly, however, median δ_e exceeds 0 for every bin except when the genomic distance exceeds 150 Mb (**Fig.3-2D**). Hence, the linkage effect is long-range. To statistically verify the potentially chromosome-wide linkage effect, we focused on linked gene pairs that are at least 63 Mb apart, which is one half the median size of mouse chromosomes. The median δ_e for these gene pairs is 0.017, or 68% of the median δ_e for the left-most bin in **Fig.3-2D**. We randomly shuffled the genomic positions of all genes and repeated the above analysis 1000 times. In none of the 1000 shuffled genomes did we observe the median δ_e greater than 0.017 for linked genes of distances >63 Mb, validating the long-range expression co-fluctuation in the actual genome. The above observations are not clone-specific, because the same trend is observed for cells of clone 6 (**Fig.A2-1B**).

Notably, a previous experiment in mammalian cells (Raj et al., 2006) detected a linkage effect for chromosomally adjacent reporter genes (δ_e = 0.834) orders of magnitude stronger than what is observed here. This is primarily because expression levels estimated using single-cell RNA fluorescence in situ hybridization in the early study (Raj et al., 2006) are much more precise than those estimated using allele-specific single-cell RNA-seq (Raj et al., 2008) here. We thus predict that the linkage effect detected will be more pronounced as the expression level estimates become more precise. As a proof of principle, we gradually raised the required minimal number of reads per allele in our analysis, which should increase the precision of expression level estimation but decrease the number of genes that can be analyzed. Indeed, as the minimal read number rises, the fraction of chromosomally linked gene pairs with a positive δ_e (**Fig.3-2E**), median δ_e for all chromosomally linked gene pairs (**Fig.3-2F**), and median δ_e for the left-most bin (**Fig.3-2F**) all increase. Further more,through simulation that incorparates known parameters with regard to our dataset(see Methods), we can estimate a lower bound for $\delta_{\mathrm{e}.}$ As shown in **Fig.A2-6A**, if we take into account the low capturing efficiency of single-cell RNA-seq, we will have the relationship:

$$\text{estimated } \delta_e \ = \ 0.13 \times \text{true } \delta_e + 0.0009$$

The median estimated δ_e in our data set is 0.020, therefore, the true δ_e is estimated to be $(0.02 - 0.0009)/0.13 \approx 0.15$. This estimation is strictly a lower bound, since we only considered the transcript capturing loss in the reverse transcription step whose magnitude we have empirical knowledge about. The true δ_e can only be larger.

Because what matters to a cell is the total number of transcripts produced from the two alleles of a gene instead of the number produced from each allele, we also calculated the pairwise correlation in expression level between genes using either the total number of reads mapped to both alleles of a gene or normalized expression level of the gene. We similarly found a long-range linkage effect (**Fig.A2-2**), with trends and effect sizes close to the observations based on allele-specific expressions.

Previous studies reported that the relative transcriptional orientations of neighboring genes influence their expression co-fluctuation (Yan et al., 2016). This impact, however, is unobserved in our study (**Fig.A2-4**), which may be due to the limited precision of the expression estimates and the fact that only 422 pairs of neighboring genes satisfy the minimal read number requirement.

3.3.2 Shared chemical environment for transcription results in the long-range linkage effect

What has caused the chromosome-wide expression co-fluctuation of linked genes? Individual chromosomes in mammalian cells are organized into territories with a diameter of 1~2 µm (Dekker and Mirny, 2016), whereas the diameter of the nucleus is ~8 µm (Dekker and Mirny, 2016). Thus, the physical distance between chromosomally linked genes is below 1~2 µm, whereas that between unlinked genes is usually > 1~2 µm and can be as large as ~8 µm. Because it takes time for macromolecules to diffuse in the nucleus, linked genes tend to have similar chemical environments and hence similar transcriptional dynamics (i.e., promoter co-accessibility and/or co-transcription) when compared with unlinked genes. We thus hypothesize that the linkage effect is fundamentally explained by the 3D proximity of linked genes compared with unlinked genes (**Fig.3-3A**). Below we provide evidence for this model.

We started by comparing the 3D distances between linked alleles with those between unlinked alleles. The 3D distance between two genomic regions can be approximately measured by Hi-C, a high-throughput chromosome conformation capture method for quantifying the number of interactions between genomic loci that are nearby in 3D space (Belton et al., 2012). The smaller the 3D distance between two genomic regions, the higher the interaction frequency between them(Dekker et al., 2013). It is predicted that the interaction frequency between the physically linked alleles of two genes (*cis*-interaction) is greater than that between the unlinked alleles of the same gene pair (*trans*-interaction). To verify this prediction, we analyzed the recently published allele-specific 500kb-resolution Hi-C interaction matrix (Giorgetti et al., 2016) of mouse neural progenitor cells (NPC). For any two linked loci *A* and *B* as depicted in the left diagram of **Fig.3-2A**, we computed $\delta_\mathrm{i} = [F(A_1, B_1) + F(A_2, B_2) - F(A_1, B_2) - F(A_2, B_1)]/2$, where *F* is the interaction frequency between the two alleles in the parentheses and the subscript "i" refers to interaction. We found that 99% of pairs of linked loci have a positive δ_i ($P < 2.2\times10^{-16}$, binomial test on independent locus pairs; **Fig.3-3B**). By contrast, among unlinked gene pairs, the fraction with a positive δ_i is not significantly different from that with a negative δ_i ($P = 0.90$, binomial test on independent locus pairs; **Fig.3-3B**). In the analysis of unlinked loci, we treated all alleles from one parental species of the hybrid as alleles 1 and all alleles from the other parental species of the hybrid as alleles 2 in the above formula of δ_i. These results clearly demonstrate the 3D proximity of genes on the same chromosome when compared with those on two homologous chromosomes.

To examine if the above phenomenon is long-range, we plotted δ_i as a function of the distance (in Mb) between two linked loci considered. Indeed, even when the distance exceeds 63 Mb, one half the median size of mouse chromosomes, almost all locus pairs still show positive δ_i (**Fig.3-3C**). Similar to the phenomenon of the linkage effect on gene expression co-fluctuation, we observed a negative correlation between the genomic distance between two linked loci and δ_i ($\rho = -0.81$ for unbinned data). This correlation is statistically significant ($P < 0.001$), because it is stronger than the corresponding correlation in each of the 1000 negative controls where the genomic positions of all genes are randomly shuffled within chromosomes.

As mentioned, 3D proximity should synchronize the transcriptional dynamics of linked alleles. Based on the bursty model of gene expression (Phillips et al., 2012), transcription involves two primary steps. In the first step, the promoter region switches from the inactive state to the active state such that it becomes accessible to the transcriptional machinery. In the second step, RNA polymerase binds to the activated promoter to initiate transcription. In principle, the synchronization of either step can result in co-fluctuation of mRNA concentrations. Because the accessibility of promoters can be detected using transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2015a) in a high-throughput manner, we focused our empirical analysis on promoter co-accessibility.

To verify the potential long-range linkage effect on chromatin co-accessibility, we should ideally use single-cell allele-specific measures of chromatin accessibility. However, such data are unavailable. We reason that, the accessibility covariation of genomic regions among cells may be quantified by the corresponding covariation among populations of cells of the same type cultured under the same environment. In fact, it can be shown mathematically that, under certain conditions, chromatin co-accessibility of two genomic regions among cells equals the corresponding chromatin co-accessibility across cell populations (see Methods). Based on this result, we analyzed a dataset collected from allele-specific ATAC-seq in 16 NPC cell populations (Xu et al., 2017). We first removed sex chromosomes and then required the number of reads mapped to each allele of a peak to exceed 50 for the peak to be considered. This latter step removed imprinted loci and ensured that the considered peaks are relatively reliable. About 3500 peaks remained after the filtering. This sample size is comparable to the number of genes used in the analysis of expression co-fluctuation. For each pair of ATAC peaks, we computed $\delta_\mathrm{a} = [r_\mathrm{a}(A_1, B_1) + r_\mathrm{a}(A_2, B_2) - r_\mathrm{a}(A_1, B_2) - r_\mathrm{a}(A_2, B_1)]/2$, where r_a is the correlation in ATAC-seq read number between the alleles specified in the parentheses (following the left diagram in **Fig.3-2A**) across the 16 cell populations and the subscript "a" refers to chromatin accessibility. The fraction of peak pairs with a positive δ_a is significantly greater than 0.5 for linked peak pairs but not significantly different from 0.5 for unlinked peak pairs (binomial test on independent peak pairs; **Fig.3-3D**). Furthermore, after grouping ATAC peak pairs into 100 equal-interval bins according to the genomic distance between peaks, we observed a clear trend that δ_a decreases with the genomic distance between peaks (ρ = -0.05 for unbinned data, $P < 0.001$, within-chromosome shuffling test; **Fig.3-3E**). In addition, even for linked peak pairs with a distance greater than 63 Mb, their median δ_a is significantly greater than that of unlinked peak pairs ($P < 0.001$, among-chromosome shuffling test). Together, these results demonstrate a long-range linkage effect on chromatin co-accessibility. Similar to the δ_e, the observed δ_a is small in our dataset. As already shown in **Fig.A2-5C~D**, this is likely also due to the low capturing efficiency in high-throughput sequencing technique. Through simulation that incorporates known parameters of our dataset, we have (**Fig.A2-6B**):

$$estimated\ \delta_a = 0.04 \times \mathrm{true}\ \delta_a + 0.002$$

Given that the median δ_a is 0.0036, the true δ_a is at least $(0.0036 - 0.002)/0.04 \approx 0.03$, an order of magnitude larger. Again, this estimation is strictly a lower bound since we only consider capturing loss due to reverse transcription.

Because we hypothesize that the linkage effect on expression co-fluctuation is via 3D chromatin proximity that leads to chromatin co-accessibility (**Fig.3-3A**), we should verify the relationship between 3D proximity and chromatin co-accessibility for unlinked genomic regions to avoid the confounding factor of linkage. To this end, we converted ATAC-seq read counts to a 500kb resolution by summing up read counts for all allele-specific chromatin accessibility peaks that fall within the corresponding Hi-C bin, because the resolution of the Hi-C data is 500kb. Because alleles from different parents are unlinked in the hybrid used for ATAC-seq, for each pair of bins, we computed the mean correlation in chromatin accessibility between the alleles derived from different parents among the 16 cell populations, or *trans*-r_a = $r_a(A_1, B_2)/2 + r_a(A_2, B_1)/2$. For the same reason, we computed the sum of Hi-C contact frequency between the alleles derived from different parents, *trans*-F = $F(A_1, B_2) + F(A_2, B_1)$. Because interaction frequencies in Hi-C data are generally low for unlinked regions, we separated all pairs of bins into two categories, contacted (i.e., *trans*-F > 0) and uncontacted (i.e., *trans*-F = 0). We found that *trans*-r_a values for contacted bin pairs are significantly higher than those for uncontacted bin pairs ($P < 0.0001$; **Fig.3-3F**), consistent with our hypothesis that 3D chromatin proximity induces chromatin co-accessibility. The above statistical significance was determined by performing a Mantel test using the original *trans*-r_a matrix of the aforementioned allele pairs and the corresponding *trans*-F matrix. Corroborating our finding, a recent study of single-cell (but not allele-specific) chromatin accessibility data also found that the co-accessibility of two loci rises with their 3D proximity (Buenrostro et al., 2015b).

To test the hypothesis that chromatin co-accessibility leads to expression co-fluctuation (even for unlinked alleles) (**Fig.3-3A**), we analyzed the allele-specific ATAC-seq data and single-cell allele-specific RNA-seq data together. Although these data were generated from different cell types in mouse, we reason that, because the 3D chromosome conformation is highly similar among tissues (Dixon et al., 2012), chromatin co-accessibility, which is affected by 3D chromatin proximity (**Fig.3-3F**), may also be similar among tissues. Hence, it may be possible to detect a correlation between chromatin co-accessibility and expression co-fluctuation. To this end, we used unbinned ATAC-peak data to compute *trans*-r_a but limited the analysis to those peaks with at least 10 reads per allele. We used the allele-specific RNA-seq data to compute *trans*-$r_\mathrm{e} = r_\mathrm{e}(A_1, B_2)/2 + r_\mathrm{e}(A_2, B_1)/2$ for pairs of linked genes. We then assigned each gene to its nearest ATAC peak and averaged *trans*-r_e among gene pairs assigned to the same pair of ATAC peaks. We subsequently grouped ATAC peak pairs into 100 equal-interval bins according to their co-accessibilities, and observed a clear positive correlation between median *trans*-r_a and median *trans*-r_e across the 100 bins (**Fig.3-3G**). For unbinned data, *trans*-r_a and *trans*-r_e also show a significant, positive correlation ($\rho = 0.021$, $P = 0.027$, Mantel test).

Heretofore we showed qualitatively shared chemical enviroment due to 3D proximity can result in chromatin co-accessibility, which leads to expression co-fluctuation. In order to visualize the relationship between 3D proximity and expression co-fluctuation quantitatively, we analyzed Hi-C contact frequency and gene expression co-fluctuation together. Notice, for vast majority of unlinked genomic regions, the Hi-C contact frequency is zero, which means their 3D proximity information is lost. Therefore, we only consider genomic region pairs that are linked in this analysis. For Hi-C contact requency of each genomic region pair, we sum up all four interaction frequencies as a measure of total interaction frequencies for that pair. Next, we assigned each gene in our dataset to their nearest Hi-C bin. For each gene pair, we compute the correlation as the average of the four allelic pair expression correlations computed previously. And for gene pairs that assigned to the same genomic region pairs defined by the Hi-C bins, we computed the average correlations of all pairs. We subsequently grouped the region pairs into 100 bins according to their Hi-C total contact frequency. For each bin, we computed the median Hi-C contact frequency and the median expression co-fluctuation. Because the dynamic range of Hi-C contact frequency is large, we converted median Hi-C contact frequency in each bin into log scale. We found a clear positive trend between Hi-C total contact frequencies and expression co-fluctuation (**Fig.3-3H**). To assess the significance of this trend, we first ordered the genomic region pairs by their Hi-C contact frequencies in a descending order. We then went through all the genomic region pairs and recorded the newly encounter genomic regions. If we encounter a genomic region pair that contains a genomic region that already recorded before, we removed that genomic pair. This operation allows us to obtain a set of independent pairs for which every genomic region only appears once. The reason that we ordered our pairs first is to ensure that genomic region pairs with high Hi-C measurement accuracy are more likely to be retained, since the measurement for low Hi-C contact frequency values is inaccurate due to small number effect. We further controlled 1D distance using partial correlation. We found that the correlation between Hi-C contact frequency and expression co-fluctuation is significant (Partial r=0.14, P=0.004). We then explored whether or not co-accessibility mediated by physical proximity can fully account for the positive correlation. To obtain co-accessibility measure for each independent genomic region pair, we used the *trans*-r_a computed previously and computed *cis*-r_a similarly. We used the average of all the r_a values as our co-accessibility measure for the independent genomic pairs that we kept.We found that the correlation remains significant after controlling for co-accessibility (Partial r=0.13, P=0.01). The above results give quantitative support for our model demonstrated in **Fig.3-3A**.

The above results support our hypothesis that, compared with unlinked genes, linked genes have a shared chemical environment due to their 3D proximity and hence chromatin co-accessibility, which leads to their expression co-fluctuation (**Fig.3-3A**). However, 3D proximity can lead to promoter co-accessibility by several means, which have been broadly summarized into three categories of mechanisms (Dekker and Mirny, 2016): 1D scanning, 3D looping, and 3D diffusion. 1D scanning refers to the spread of chromatin states along an entire chromosome. However, 1D scanning is rare, with only a few known examples such as X-chromosome inactivation (Dekker and Mirny, 2016). Hence, 1D scanning is unlikely to be the mechanism responsible for the broad linkage effect discovered here. 3D looping refers to the phenomenon that a chromosome often forms loops to bring far-separated loci into contact, whereas 3D diffusion refers to chromosome communication by local diffusion of transcription-related proteins. For tightly linked loci, our data do not allow a clear distinction between 3D looping and 3D diffusion in causing the linkage effect discovered here. But 3D diffusion seems more likely for the long-range effect, because the range of 3D looping seems limited to loci separated by no more than 200 kb simply due to the rapid decrease of the contact frequency with the physical distance between two loci (Hahn and Kim, 2013), evident in **Fig.3-3C** (note the log scale of the Y-axis). It has been estimated that loci separated by 10 Mb behave essentially the same as two loci that are on different chromosomes in terms of the contact frequency (Dekker and Mirny, 2016), and any contact-based mechanism is unlikely to be long-range (e.g., topologically associating domains) (Dixon et al., 2012). Therefore, the most likely cause of our observed long-range linkage effect is 3D diffusion.

In the 3D diffusion mechanism, which molecule is most likely responsible for the observed long-range linkage effect on expression co-fluctuation? If the chemical influencing transcription has a diffusion time in the nucleus much shorter than the interval between transcriptional bursts, two genes have essentially the same environment with respect to that chemical regardless of their 3D distance (Mahmutovic et al., 2012) and hence no linkage effect is expected (top cell in **Fig.3-3I**). On the contrary, if the chemical diffuses too slowly to even distribute evenly in a chromosomal territory in a time comparable to the interval between transcriptional bursts, the linkage effect will be local (Mahmutovic et al., 2012) and hence cannot be chromosome-wide (bottom cell in **Fig.3-3I**). Therefore, the diffusion rate of the chemical responsible for the long-range linkage effect cannot be too low or too high such that they become evenly distributed in a chromosome territory but not the whole nucleus in a time comparable to the interval between transcriptional bursts (middle cell in **Fig.3-3I**). The typical transcriptional burst interval is 18-50 minutes in mammalian cells (Dar et al., 2012; Suter et al., 2011). The time for a chemical to distribute evenly in a given volume with radius *R* is on the order of R^2/D, where *D* is the diffusion coefficient of the chemical (Phillips et al., 2012). Most molecules in the nucleus are rapidly diffused. For example, transcription factors typically have a diffusion coefficient of 0.5-5 $\mu m^2/s$ in the nucleus (Hager et al., 2009; Phillips et al., 2012), meaning that they can diffuse across the whole nucleus in ~3~30 seconds. By contrast, core histone proteins such as H2B proteins diffuse extremely slowly due to their tight binding to DNA. They are usually considered immobilized because diffusion is rarely observed during the course of an experiment (Hager et al., 2009; Lever et al., 2000). Therefore, none of these molecules are responsible for the long-range linkage effect observed. Interestingly, linker histones, which include five subtypes of H1 histones in mouse that play important roles in chromatin structure and transcription regulation (Fyodorov et al., 2018), have a diffusion coefficient of ~0.01$\mu m^2/s$ (Bernas et al., 2014). Thus, it takes H1 proteins 25-100 seconds to diffuse through a chromosome territory, but ~30 minutes to diffuse across the whole nucleus. The former time but not the latter is much smaller than the typical transcriptional burst interval. Hence, it is possible that H1 diffusion in the nucleus is the ultimate cause of the linkage effect. We provide empirical evidence for this hypothesis in a later section.

3.3.3 Beneficial linkage of genes encoding components of the same protein complex

Our finding that chromosomal linkage leads to gene expression co-fluctuation implies that linkage between genes could be selected for when expression co-fluctuation is beneficial.

Due to the complexity of biology, it is generally difficult to predict whether the expression co-fluctuation of a pair of genes is beneficial, neutral, or deleterious. However, the expression co-fluctuation of genes encoding components of the same protein complex is likely advantageous. To see why this is the case, let us consider a dimer composed of one molecule of protein A and one molecule of protein B; the heterodimer is functional but monomers are not. We denote the concentration of dissociated protein A as [A], the concentration of dissociated protein B as [B], and the concentration of protein complex AB as [AB]. At the steady state, [AB] = K[A][B], where K is the association constant (Veitia, 2010). Furthermore, the total concentration of protein A, $[A]_t$, equals [A] + [AB], and the total concentration of protein B, $[B]_t$, equals [B] + [AB]. Based on these relationships, we simulated 10,000 cells, where the mean and coefficient of variation (CV) are respectively 1 and 0.2 for both $[A]_t$ and $[B]_t$ (see Methods). We assumed $K = 10^5$ based on empirical K values of protein complexes (Milo et al., 2009). We found that, as the correlation between $[A]_t$ and $[B]_t$ increases, mean [AB] of the 10,000 cells rises (**Fig.3-4A**). If we assume that fitness rises with [AB], the co-fluctuation of $[A]_t$ and $[B]_t$ is beneficial, compared with independent fluctuations of $[A]_t$ and $[B]_t$. Furthermore, because mean [A] and mean [B] must decrease with the rise of mean [AB], the co-fluctuation of $[A]_t$ and $[B]_t$ could also be advantageous because it lowers the concentrations of the unbound monomers that may be toxic. Indeed, past studied found better expression co-fluctuations of genes encoding members of the same protein complex than random gene pairs (Budnik et al., 2018; Sigal et al., 2006), suggesting a demand for expression co-fluctuation of members of the same protein complex. We also simulated the concentration of [AB] under a wide range of K (K=0.1, 1,10, 100, 1000, 10000, 100000), our results remain largely unchanged, and the lower bound mean [AB] is 3% higher under co-fluctuation than under no co-fluctuation. Moreover, the effect size rises substantially if CV of protein is larger. For example, when CV = 0.5, the effect rises to 20%. For eukaryotic species, CV of protein generally ranges from 0.1 to 1 (Milo et al., 2009).We also considered dimers with different stoichiometries and suboptimal mean concentrations (see Methods). In all of the combinations of the parameters, the mean concentration of the protein complex increases as the correlation in expression levels of A and B increases, albeit with a wide range of effect sizes (0.001% to 27% higher under co-fluctuation than under no co-fluctuation).

To test if genes encoding components of the same protein complex tend to be linked, we used the mouse protein complex data from CORUM and downloaded the chromosomal positions of all mouse protein-coding genes from Ensembl (Aken et al., 2016). Because genes may be linked due to their origins from tandem duplication(Ibn-Salem et al., 2016), the data were pre-processed to produce a set of duplicate-free mouse protein-coding genes (see Methods). We then randomly shuffled the genomic positions of the retained genes encoding protein complex components among all possible positions of the duplicate-free mouse protein-coding genes. The observed number of linked pairs of genes encoding components of the same protein complex is significantly greater than the random expectation (**Fig.3-4B**). For comparison, we also computed the number of linked pairs of genes encoding components of different protein complexes. This number is not significantly greater than the random expectation (**Fig.3-4C**). Thus, the enrichment in gene linkage is specifically related to coding for components of the same protein complex. Interestingly, the observed median distance between the TSSs of two linked genes encoding protein complex components is not significantly different from the random expectation, regardless of whether components of the same (**Fig.3-4D**) or different (**Fig.3-4E**) protein complexes are considered.

The phenomenon that members of the same protein complex tend to be encoded by linked genes could have arisen for one or both of the following reasons. First, selection for co-fluctuation among proteins of the same complex has driven the evolution of gene linkage.

Second, due to their co-fluctuation, products of linked genes may have been preferentially recruited to the same protein complex in evolution. Under the first hypothesis, originally unlinked genes encoding members of the same protein complex are more likely to become linked in evolution than originally unlinked genes that do not encode members of the same complex. To verify this prediction, we examined mouse genes using rat and human as outgroups (**Fig.3-4F**). We obtained pairs of genes encoding components of the same protein complex in both human and mouse. Hence, these pairs likely encode members of the same protein complex in the common ancestor of the three species. Among them, 875 pairs are unlinked in human and rat, suggesting that they were unlinked in the common ancestor of the three species. Of the 875 pairs, 25 pairs become linked in the mouse genome, significantly more than the random expectation under no requirement for gene pairs to encode members of the same complex (P = 0.005; **Fig.3-4F**; see Methods). Therefore, the first hypothesis is supported. Under this hypothesis, the result in **Fig.3-4D** may be explained by the long-range linkage effect on expression co-fluctuation, such that once two genes encoding components of the same protein complex move to the same chromosome, selection is not strong enough to drive them closer to each other. To test the second hypothesis, we need gene pairs encoding proteins that belong to the same protein complex in mouse but not in human nor rat, which require such low false negative errors in protein complex identification that no current method can meet. Hence, we leave the validation of the second hypothesis to future studies.

As mentioned, our theoretical consideration suggests that, due to their intermediate diffusion coefficient, H1 histones may be responsible for the observed chromosome-wide expression co-fluctuation. Because the local H1 concentration fluctuates more when its cellular concentration is lower, we predict that the benefit of and the coefficient of selection for linkage of genes encoding members of the same protein complex is greater in tissues with lower H1 concentrations. Given that gene expression is costly, for a given gene, it is reasonable to assume that the relative importance of its function in a tissue increases with its expression level in the tissue (Cherry, 2010; Gout et al., 2010). Hence, we predict that, the more negative the across-tissue expression correlation is between a protein complex member gene and H1 histones, the higher the likelihood that the gene is driven to be linked with other genes encoding members of the same protein complex. To verify the above prediction, we used a recently published RNA-seq dataset (Söllner et al., 2017) to measure Pearson's correlation between the mRNA concentration of a gene that encodes a protein complex member and the mean mRNA concentration of all H1 histone genes across 13 mouse tissues. Indeed, the linked protein complex genes show more negative correlations than the unlinked protein complex genes (*P* = 0.012, one-tailed Mann-Whitney *U* test; **Fig.3-4G**). The disparity is even more pronounced when we compare linked protein complex genes that become linked in the mouse lineage with unlinked protein complex genes (*P* = 0.00068, one-tailed Mann-Whitney *U* test; **Fig.3-4G**). This is likely owing to the enrichment of genes that are linked due to the linkage effect in the group of evolved linked protein complex genes ($\frac{\text{Observed}-\text{null expectation}}{\text{null expectation}} = \frac{25-13}{13} = 92\%$) when compared with the group of linked protein complex genes ($\frac{\text{Observed}-\text{null expectation}}{\text{null expectation}} = \frac{200-161}{161} = 24\%$). The above three groups of genes (evolved linked protein complex genes, linked protein complex genes, and unlinked protein complex genes) were constructed using stratified sampling so that their mean expression levels across tissues are not significantly different (see Methods). For comparison, we performed the same analysis but replaced H1 histones with TFIIB, a general transcription factor that is involved in the formation of the RNA polymerase II preinitiation complex and has a high diffusion rate (Vosnakis et al., 2017). The trends shown in **Fig.3-4G** no longer holds (unlinked vs. linked: *P* = 0.11, one-tailed Mann-Whitney *U* test; unlinked vs. evolved linked: *P* = 0.63, one-tailed Mann-Whitney *U* test). We also performed the same analysis but replaced H1 histones with core histone proteins, which are immoblized (Lever et al., 2000).

Again, the trends in **Fig.3-4G** disappeared (unlinked vs. linked: *P* = 0.48, one-tailed Mann-Whitney *U* test; unlinked vs evolved linked: *P* = 0.89, one-tailed Mann-Whitney *U* test). These results support our hypothesis about the role of H1 histones in the linkage effect of expression co-fluctuation.

3.4 Discussion

Using allele-specific single-cell RNA-seq data, we discovered chromosome-wide expression co-fluctuation of linked genes in mammalian cells. We hypothesize and provide evidence that genes on the same chromosome tend to have close 3D proximity, which results in a shared chemical environment for transcription and leads to expression co-fluctuation. While the linkage effect on expression co-fluctuation is likely an intrinsic cellular property, when the expression co-fluctuation of certain genes improves fitness, natural selection may drive the relocation of these genes to the same chromosome. Indeed, we provide evidence suggesting that the chromosomal linkage of genes encoding components of the same protein complex is beneficial owing to the resultant expression co-fluctuation that minimizes the dosage imbalance among these components and has been selected for in genome evolution.

Although many statistical results in this study are highly significant, the effect sizes appear small in several analyses, most notably the δ_e and δ_a values for linked genes. The small effect sizes are generally due to the large noise in the data, less ideal types of data used, and mismatches between the data sets co-analyzed. For instance, δ_e between linked genes estimated here (**Fig.3-2D**) is much smaller than what was previously estimated for a pair of linked florescent protein genes (Raj et al., 2006), due in a large part to the inherently large error in quantifying mRNA concentrations by single-cell RNA-seq (Marinov et al., 2014). The small size of δ_a (**Fig.3-3E**) is likely caused at least in part by the low efficiency of ATAC-seq in detecting open chromatin (see Methods). The positive correlation between *trans*-r_a and *trans*-r_e (**Fig.3-3G**) is likely an underestimate due to the use of different cell types in RNA-seq and ATAC-seq. As shown in Figs. 2E and 2F, the actual effect sizes would be much larger should better experimental methods and/or data become available. Hence, it is likely that many effects are underestimated in this study. In addition, the co-fluctuation effect detected by Raj et al. may be unusually large because in that study the chromosomal distance between the two genes was extremely small and the two genes used identical regulatory elements (Raj et al., 2006).Regardless, it is important to stress that whether an effect is large or not depends on the sensitivity of natural selection. According to the results of **Fig.3-4 B~G**, the effects appear visible to natural selection, as reflected in the preferential chromosomal linkage of genes encoding members of the same protein complex. It may seem surprising that the apparently small effect of single cell co-fluctuation can be detected by natural selection. However, based on basic population genetics (Ohta, 1992), natural selection can detect a selection adavantage as small as the inverse of the effective population size. The mouse effective population size is about 70,000 (Phifer-Rixey et al., 2012), so natural selection can detect a fitness differential that is as small as 1/70000.

Because we used RNA-seq to measure expression co-fluctuation, our results apply to the co-fluctuation of mRNA concentrations. In the case of protein complex components, it is presumably the co-fluctuation of protein concentrations rather than mRNA concentrations that is directly beneficial. Although the degree of covariation between mRNA and protein concentrations is under debate (Kustatscher et al., 2017; Liu et al., 2016), the two concentrations correlate well at the steady state (Raj et al., 2006). One key factor in this correlation is the protein half-life, because, when the protein half-life is long, mRNA and protein concentrations may not correlate well due to the delay in the effect of a change in mRNA concentration on protein concentration (Raj et al., 2006). It is interesting to note that in Raj et al.'s study (Raj et al., 2006), mRNA and protein concentrations still correlate reasonably well ($r = 0.43$) when the protein half-life is 25 hours, which is much longer than the reported mean protein half-life of 9 hours in mammalian cells (Eden et al., 2011). Corroborating this finding is the recent report (Popovic et al., 2018) that mRNA and protein concentrations correlate well across single cells in the steady state (mean *r* = 0.732). Note that, although the correlation between mRNA and protein concentrations measured at the same moment may not be high when the protein half-life is long, the current protein level can still correlate well with a past mRNA level (Gedeon and Bokes, 2012). Because our study focuses on cells at the steady state, co-fluctuation of mRNA concentrations is expected to lead to co-fluctuation of protein concentrations.

We attributed the preferential linkage of genes encoding components of the same protein complex to the benefit of expression co-fluctuation, while a similar phenomenon of linkage was previously reported in yeast and attributed to the potential benefit of co-expression of protein complex components across environments (Teichmann and Veitia, 2004), where co-expression refers to the correlation in mean expression level. In mammalian cells, our hypothesis is more plausible than the co-expression hypothesis for five reasons. First, across-environment (or among-tissue) variation in mean mRNA concentration does not translate well to the corresponding variation in mean protein concentration (Franks et al., 2017; Kustatscher et al., 2017), while mRNA concentration fluctuation explains protein concentration fluctuation quite well (Popovic et al., 2018; Raj et al., 2006). Hence, gene linkage, which enhances mRNA concentration co-fluctuation and by extension protein concentration co-fluctuation, may not improve protein co-expression across environments. Second, co-expression of linked genes appears to occur at a much smaller genomic distance than the linkage effect on co-fluctuation reported here (Hurst et al., 2004). Thus, if selection on co-expression were the cause for the non-random distribution of genes encoding members of the same protein complex, these genes should be closely linked. This, however, is not observed (**Fig.3-4D**). Hence, the previous finding that genes encoding members of (usually not the same) protein complexes tend to be clustered is best explained by the fact that certain chromosomal regions have inherently low expression noise and that these regions attract genes encoding protein complex members because stochastic expressions of these genes are especially harmful (i.e., the noise reduction hypothesis) (Batada and Hurst, 2007; Chen and Zhang, 2016). Third, the protein complex stoichiometry often differs among environments, which makes co-expression of complex components disfavored in the face of environmental changes (Ori et al., 2016; Slavov et al., 2015). Nonetheless, under a given environment, protein concentration co-fluctuation remains beneficial because of the presence of an optimal stoichiometry at each steady state. Fourth, gene linkage is not necessary for the purpose of co-expression, because the genes involved can use similar *cis*-regulatory sequences to ensure co-expression even when they are unlinked. In fact, a large fraction of co-expression of linked genes is due to tandem duplicates (Hurst et al., 2004), which have similar regulatory sequences by descent. However, even for genes with the same regulatory sequences, linkage improves expression co-fluctuation at the steady state. Finally, the co-expression hypothesis or noise reduction hypothesis cannot explain our observation of the relationship between the expression levels of H1 histones and those of linked genes encoding protein complex members across tissues (**Fig.3-4G**). Taken together, these considerations suggest that it is most likely the selection for expression co-fluctuation rather than co-expression across environments that has driven the evolution of linkage of genes encoding members of the same protein complex.

Several previous studies reported long-range coordination of gene expression (Fukuoka et al., 2004; Ghanbarian and Hurst, 2015; Kustatscher et al., 2017; Lercher and Hurst, 2006; Levesque and Raj, 2013; Liao and Zhang, 2008; Sémon and Duret, 2006; Singer et al., 2004; Spellman and Rubin, 2002), but most of them was about co-expression. As discussed, co-expression is the correlation in mean expression level across different tissues or environments and differs from expression co-fluctuation across single cells in the same environment. One study used fluorescent in situ hybridization of intronic RNA to detect nascent transcripts in individual cells (Levesque and Raj, 2013). The authors reported independent transcriptions of most linked genes with the exception of two genes about 14 million bases apart that exhibit a negative correlation in transcription. Their observations are not contradictory to ours, because they measured the nearly instantaneous rate of transcription, whereas we measured the mRNA concentration that is the accumulated result of many transcriptional bursts. As explained, having a similar biochemical environment makes the activation/inactivation cycles of linked genes coordinated to some extent, even though the stochastic transcriptional bursts in the activation period may still look independent.

Our work suggests several future directions of research regarding expression co-fluctuation and its functional implications. First, it would be interesting to know if the linkage effect on expression co-fluctuation varies across chromosomes. Although we analyzed individual chromosomes (**Fig.A2-3**), addressing this question fully requires better single-cell expression data, because the current single-cell RNA-seq data are noisy. This also makes it difficult to detect any unusual chromosomal segment in its δ_e distribution. Second, our results suggest that 3D proximity is a major cause for the linkage effect on expression co-fluctuation. In particular, diffusion of proteins with intermediate diffusion coefficients such as H1 histones is likely one mechanistic basis of the effect. However, the diffusion behaviors of most proteins involved in transcription are largely unknown. A thorough research on the diffusion behaviors of proteins inside the nucleus will help us identify other proteins that are important in the linkage effect. As mentioned, our data do not allow a clear distinction between 3D looping and 3D diffusion in causing the linkage effect on tightly linked genes. To distinguish between these two mechanisms definitively, we would need allele-specific models of mouse chromosome conformation (Naumova et al., 2013), which require more advanced algorithms and more sensitive allele-specific Hi-C methods. Third, our study highlights the importance of the impact of sub-nucleus spatial heterogeneity in gene expression. This can be studied more thoroughly via real-time imaging and spatial modeling of chemical reactions (Elf and Barkefors, 2018; Mahmutovic et al., 2012). The lack of knowledge about the details of transcription reactions prevents us from constructing an accurate quantitative model of gene expression, which can be achieved only by more accurate measurement and more advanced computational modeling. Fourth, we used protein complexes as an example to demonstrate how the linkage effect on expression co-fluctuation influences the evolution of gene order. But, to understand the broader evolutionary impact of the linkage effect, a general prediction of the fitness consequence of expression co-fluctuation is necessary. To achieve this goal, whole-cell modeling may be required (Carrera and Covert, 2015). Note that some other mechanisms such as cell cycle (Rustici et al., 2004) can also lead to gene expression co-fluctuation and so should be considered when predicting the relationship between gene expression and fitness. Fifth, because expression co-fluctuation could be beneficial or harmful, an alteration of expression co-fluctuation should be considered as a potential mechanism of disease caused by mutations that relocate genes in the genome. Sixth, our analysis focused primarily on highly expressed genes due to the limited sensitivity of single-cell RNA-seq. Because lowly expressed genes are affected more than highly expressed genes by expression noise (Raj et al., 2010), expression co-fluctuation may be more important to lowly expressed genes than highly expressed ones. More sensitive and accurate single-cell expression profiling methods are needed to study the expression co-fluctuation of lowly expressed genes. Seventh, we focused on mouse fibroblast cells because of the limited availability of allele-specific single-cell RNA-seq data. To study how expression co-fluctuation impacts the evolution of gene order, it will be important to have data from multiple cell types and species. Last but not least, as we start designing and synthesizing genomes (Baker, 2011), it will be important to consider how gene order affects expression co-fluctuation and potentially fitness. It is possible that the fitness effect associated with expression co-fluctuation is quite large when one compares an ideal gene order with a random one. It is our hope that our discovery will stimulate future researches in above areas.

3.5 Methods

3.5.1 High-throughput sequencing data

The processed allele-specific single-cell RNA-seq data were downloaded from https://github.com/RickardSandberg/Reinius_et_al_Nature_Genetics_2016?files=1 (mouse.c57.counts.rds and mouse.cast.counts.rds). The Hi-C data (Giorgetti et al., 2016) were downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72697, and we analyzed the 500kb-resolution Hi-C interaction matrix with high SNP density (iced-snpFiltered). The processed ATAC-seq data were provided by authors(Xu et al., 2017), and the data from 16 NPC cell populations were analyzed. All analyses were performed using custom programs in R or python.

3.5.2 Protein complex data and pre-processing

The mouse protein complex data were downloaded from the CORUM database (http://mips.helmholtz-muenchen.de/corum/) (Ruepp et al., 2009). The coordinates for all mouse protein-coding genes were downloaded from Ensembl BioMart (GRC38m.p5) (Aken et al., 2016). To produce duplicate-free gene pairs, we also downloaded all paralogous gene pairs from Ensembl BioMart. Note that these gene pairs can be redundant, meaning that a gene may be paralogous with multiple other genes and appear in multiple gene pairs. We then iteratively removed duplicate genes based on the following rules. First, if one gene in a pair of duplicate genes has been removed, the other gene is retained. Second, if neither gene in a duplicate pair has been removed and neither encodes a protein complex component, one of them is randomly removed. Third, if neither gene in a duplicate pair has been removed and only one of them encodes a protein complex member, we remove the other gene. Fourth, if neither gene in a duplicate pair has been removed and both genes encode protein complex components, one of them is randomly removed. Applying the above rules resulted in a set of duplicate-free genes with as many of them encoding protein complex members as possible.

3.5.3 Gibbs sampling for testing protein complex-driven evolution of gene order

We obtained all mouse genes that have one-to-one orthologs in both human and rat, and acquired from Ensembl their chromosomal locations in human, mouse, and rat. Gene pairs are formed if their products belong to the same protein complex in human as well as mouse, based on protein complex information in the CORUM database mentioned above. Among them, 875 gene pairs from 342 genes are unlinked in both human and rat, of which 25 pairs become linked in mouse. To test whether the number 25 is more than expected by chance, we compared these 342 genes with a random set of 342 genes that also form 875 unlinked gene pairs in human and rat. These unlinked pairs are highly unlikely to encode members of the same complex, so serve as a negative control. Because of the difficulty in randomly sampling 342 genes that form 875 unlinked gene pairs, we adopted Gibbs sampling (Geman and Geman, 1987), one kind of Markov-Chain Monte-Carlo sampling (Gilks, 2005). The procedure was as follows. Starting from the observed 342 genes, represented by the vector of (gene 1, gene 2, …, gene 342), we swapped gene 1 with a randomly picked gene from the mouse genome such that the 342 genes still satisfied all conditions of the original 342 genes described above. We then similarly swapped gene 2, gene 3, ..., and finally gene 342, at which point a new gene set was produced. To allow the Markov chain to reach the stationary phase, we discarded the first 1000 gene sets generated. Starting the 1001st gene set, we retained a set every 50 sets produced until 1000 sets were retained; this ensured relative independence among the 1000 retained sets. In each of these 1000 sets, we counted the number of gene pairs that are linked in mouse. The fraction of sets having the number equal to or greater than 25 was the probability reported in **Fig.3-4F**.

3.5.4 Chromatin co-accessibility among cells vs. among cell populations

Let us consider the chromatin accessibilities of two genomic regions, *A* and *B*, in a population of *N* cells (*N* = 50,000 in the data analyzed) (Xu et al., 2017). Let us denote the chromatin accessibilities for the two regions in cell *i* by random variables A_i and B_i, respectively, where i=1, 2, 3, ..., and *N*. We further denote the corresponding total accessibilities in the population as random variables *AT* and *BT*, respectively. We assume that A_i follows the distribution *X*, while B_i follows the distribution *Y*. We then have the following equations.

$$AT = \sum_{i=1}^{N} A_i \text{ and } BT = \sum_{i=1}^{N} B_i \, . \tag{1}$$

Pearson's correlation between *AT* and *BT* across cell populations all of size *N* is

$$Corr(AT, BT) = \frac{E(AT \cdot BT) - E(AT)E(BT)}{\sqrt{Var(AT)Var(BT)}} = \frac{E(\sum_{i=1}^{N}\sum_{j=1}^{N} A_i B_j) - N^2 E(X)E(Y)}{\sqrt{N^2 Var(X)Var(Y)}}$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} E(A_i B_j) - N^2 E(X)E(Y)}{N\sqrt{Var(X)Var(Y)}} \, . \tag{2}$$

Because cells are independent from one another, when $i \neq j$,

$$E(A_i B_j) = E(A_i)E(B_j). \tag{3}$$

Thus,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} E(A_i B_j) = \sum_{i=1}^{N} E(A_i B_i) + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j \neq i}}^{N} E(A_i)E(B_i)$$

$$= NE(XY) + (N^2 - N)E(X)E(Y). \tag{4}$$

Combining Eq. (2) with Eq. (4), we have

$$Corr(AT, BT) = \frac{NE(XY) - NE(X)E(Y)}{N\sqrt{Var(A) \cdot Var(B)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X) \cdot Var(Y)}} = Corr(X, Y). \tag{5}$$

Hence, if the number of cells per population is a constant and there is no measurement error, correlation of chromatin accessibilities of two loci among cells is expected to equal the correlation of total chromatin accessibilities per population of cells among cell populations.

To examine how violations of some of the above conditions affect the accuracy of Eq. (5), we conducted computer simulations. We assume that the accessibility of a genomic region in a single cell is either 1 (accessible) or 0 (inaccessible). This assumption is supported by previous single-cell ATAC-seq data (Buenrostro et al., 2015b), where the number of reads mapped to each peak in a cell is nearly binary. Now let us consider two genomic regions whose chromatin states are denoted by A and B, respectively. The probabilities of the four possible states of this system are as follows.

$$\begin{aligned} &\Pr(A=0,B=0)=p,\\ &\Pr(A=0,B=1)=q,\\ &\Pr(A=1,B=0)=r,\\ \text{and}\quad &\Pr(A=1,B=1)=s, \end{aligned} \tag{6}$$

where $p + q + r + s = 1$. Hence, we have

$$\begin{aligned} &E(A)=r+s,\\ &E(B)=q+s,\\ &E(AB)=s,\\ &Var(A)=(r+s)(p+q),\\ &Var(B)=(q+s)(p+r). \end{aligned} \tag{7}$$

With Eq. (7), we can compute $Corr(A,B)$. In other words, for any given set of p, q, r, and s, we can compute the among-cell correlation in chromatin accessibility between the two regions.

We then generated 10,000 random sets of p, q, r, s from a Dirichlet distribution. For each set of p, q, r, and s, we simulated the state of a cell by a random sampling from the four possible states. We did this for 16 cells as well as 16 cell populations each composed of

50,000 cells. We computed the total accessibility of each region in each cell population by summing up the corresponding accessibility of each cell. As expected, the among-cell correlation between the two regions in accessibility matches the true correlation (**Fig.A2-5A**). The deviation from the true correlation is due to sampling error. Based on Eq. (5), the among-cell-population correlation between the two regions in total accessibility approximates the true correlation, which is indeed observed in our simulation (**Fig.A2-5B**).

Nevertheless, accessibility of a region may be undetected due to low detection efficiencies of high-throughput methods, which makes the observed correlation between the accessibilities of two regions lower than the true correlation. To assess the impact of such low detection efficiencies on the correlation, we simulated a scenario with a 10% detection efficiency, which is common in high-throughput methods (Marinov et al., 2014). That is, for every accessible region, it is detected as accessible with a 10% chance and inaccessible with a 90% chance; every inaccessible region is detected as inaccessible with a 100% chance. Our simulation showed that the observed correlation between the accessibilities of two regions is weaker than the true correlation regardless of whether the data are from individual cells (**Fig.A2-5C**) or cell populations (**Fig.A2-5D**).

True δ_a vs observed δ_a

The framework we developed in previous section "Chromatin co-accessibility among cells vs. among cell populations" allows us to perform a simulation to get a lower bound of true δ_a. We simulated the δ_a by considering two pairs of regions simultaneously. For each pair of regions, we first randomly sampled p, q, r and s, and computed the true correlation using Eq (7). The difference between the true correlations of the two pairs of regions is the true $\delta_{a.}$ Then, for each pair of regions, the estimated correlation can be obtained by simulation, from which we can get the estimated δ_a . In our allelic specific ATAC-seq data, only 55% of the reads are allelic specific. Given that in high-throughput sequencing data, the detection efficiency is around 10%~20% when considering all reads (Hwang et al., 2018), we choose 8.25%($= 0.15 \times 0.55$) as the detection efficiency in our simulation. We repeated this procedure 10000 times. And the result is plotted in **Fig.A2-6B**

3.5.5 True δ_e vs observed δ_e

To obtain a lower bound of the true δ_e, we performed a simulation incorporating the known parameters of single-cell RNA-seq in our dataset. The simulation was performed as follows:

(1) We first decided the mean expression levels for a pair of genes, A and B. The mean expression levels were sampled from the distribution of mean expression levels of genes we analyzed. The mean expression level distribution of observed genes were obtained based on the estimation that 1 RPKM correspondent to 1 transcript per cell in the original dataset (Reinius et al., 2016). Notice that the mean expression level of each allele (A_1, A_2, B_1, B_2) will be half of the above sampled values;

(2) We than generated the expression levels across 60 cells for a pair of alleles (A_1 and B_1) from joint multivariate normal distribution. The multivariate normal distribution can be uniquely determined once the correlation coefficient between two alleles and their CV are chosen. We fixed the CV of the two alleles as 0.5, based on sm-FISH experiments for mammalian cells for genes whose expression levels are similar to the genes that we analyzed. (Battich et al., 2015). Notice the CV we used here is the mRNA CV but not the protein CV. The correlation between the two alleles was randomly sampled from range (-1, 1). We call this correlation r_1.

(3) For each allele in each cell, we used binomial sampling to determine the detected transcript level. In our data set, only 17% of the reads are allelic specific. Since the capturing efficiency is around 10%~20% for full-length single cell RNA-seq data (Hwang et al., 2018), we used 2.55%($= 0.15 \times 0.17$) as the sampling probability;

(4) We than computed the observed correlation of A_1 and B_1 across cells after binomial sampling,;

(5) We repeated step (2)-(4). We call the newly sampled correlation r_2. The true δ_e would be $r_1 - r_2$, and the observed δ_e is the difference between the observed correlations;

(6) Steps (1)-(5) were repeated 10000 times, with all true δ_e and observed δ_e recorded. And the result is plotted in **Fig. A2-6A**.

Our simulation showed that the observed δ_e will be much weaker than the true δ_e (**Fig.A2-6A**). Notice this lower bound is conservative: further signal loss due to technical noise down-stream of the reverse-transcription (transcript capturing) is not modeled.

3.5.6 Simulation of protein complex concentrations

Let the concentration of protein complex AB be [AB]. To study the average [AB] across cells in a population, we first simulated the concentrations of subunit A and subunit B in each cell. We assumed that the total concentrations of A and B, denoted by $[A]_t$ and $[B]_t$ respectively, are both normally distributed with mean = 1 and *CV* = 0.2. We used *CV* = 0.2 because this is the median expression noise measured by *CV* for enzymes in yeast(Wang and Zhang, 2011), the only eukaryote with genome-wide protein expression noise data (Newman et al., 2006). Thus, the joint distribution of $[A]_t$ and $[B]_t$ is multivariate normal, which can be specified if the correlation (*r*) between $[A]_t$ and $[B]_t$ is known. With a given *r*, we simulated $[A]_t$ and $[B]_t$ for 10,000 cells by sampling from the joint distribution. We set the concentration to 0 if the simulated value is negative. We computed [AB] in each cell by solving the following set of equations.

$$[A]_t = [A] + [AB],\ [B]_t = [B] + [AB], \text{ and } [AB] = K[A][B], \qquad (8)$$

where we used $K = 10^5$ based on the empirical values of association constants of protein complexes (Milo et al., 2009). We then took the average [AB] among all cells to acquire the mean complex concentration. We also performed our simulation with a wide range of K values (K=0.1, 1,10, 100, 1000, 10000, 100000), our results remain largely unchanged.

The above simulation can be further extended to simulate the concentration of protein complex with different stoichiometry. In general, for protein complex $A_M B_N$:

$$[A]_t = [A] + M[A_M B_N],\ [B]_t = [B] + N[A_M B_N], \text{ and } [AB] = K[A]^M B^N \qquad (9)$$

Besides, by altering the mean expression level of A and B, we can futher simulate the effect when the relative concentration between A and B is suboptimal. Based on the general model, we simulated the concentration of protein complex across a wide range of parameters (K=0.1, 1,10, 100, 1000, 10000, 100000; (M, N)=(1, 1), (1,2), (1, 3), (2,2), (2, 3), (3,3); mean(A)=M, 2M, 0.5M whereas mean(B) keeps at N; CV=0.2 or 0.5). In all of the combinations of the parameters, the mean concentration of the protein complex increases as the correlation in expression levels of A and B increases, albeit with a wide range of effect size (0.001% to 27% higher under co-fluctuation than under no co-fluctuation).

3.5.7 Analysis of the relationship in expression level between protein complex genes and linker histone genes across tissues

This analysis used the RNA-seq data from 13 mouse tissues (Söllner et al., 2017) as well as the protein complex data aforementioned. We divided all protein complex genes into three groups: unlinked genes, linked genes, and evolved linked genes. The first two groups are from duplicate-free protein complex gene pairs. A gene is assigned to the "linked" group if it is linked with at least one gene that encodes a member of the same protein complex. We found that the gene expression levels tend to be higher for the "linked" group than the "unlinked" group. To allow a fair comparison between these two groups, we computed the mean expression level of each gene across tissues and performed a stratified sampling as follows. We lumped all genes from the two groups and divided them into 20 bins based on their expression levels. For each bin, we counted the numbers of linked and unlinked genes respectively, and randomly down-sampled the larger group to the size of the smaller group. After the downsampling, the expression levels of the two groups of genes are comparable (*P* = 0.9, two-tailed Mann-Whitney *U* test). The third gene group contains genes that are linked in mouse but not in human nor in rat (i.e., "evolved linked"). We did not require them to be duplicate-free, but they were ancestrally unlinked so could not have resulted from tandem duplication. The expression levels of the third group of genes are not significantly different from those of the first two groups after the stratified sampling (*P* = 0.68).

After obtaining the three groups of genes, we examined the among-tissue correlation between the expression level of each of these genes and the total expression level of all 11 H1 histone genes in mouse (Medrzycki et al., 2012). For control, we performed the same analysis but replaced H1 histones with TFIIB, a rapidly diffused transcription factor. In another control, we replaced H1 histones with immobilized core histones (H2A, H2B, H3, and H4). H2A, H2B, H3, and H4 genes are obtained from Mouse Genome Informatics (http://www.informatics.jax.org/) (Bult et al., 2008):

http://www.informatics.jax.org/vocab/pirsf/PIRSF002048

http://www.informatics.jax.org/vocab/pirsf/PIRSF002050

http://www.informatics.jax.org/vocab/pirsf/PIRSF002051

http://www.informatics.jax.org/vocab/pirsf/PIRSF002052

3.6 Data and software availability

All statistical analyses were performed using custom R and python scripts that are available upon request.

3.7 References

Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., and Clapham, P. (2016). Ensembl 2017. Nucleic Acids Res *45*, D635-D642.

Bahar, R., Hartmann, C.H., Rodriguez, K.A., Denny, A.D., Busuttil, R.A., Dolle, M.E., Calder, R.B., Chisholm, G.B., Pollock, B.H., Klein, C.A., *et al.* (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature *441*, 1011-1014.

Baker, M. (2011). Synthetic genomes: The next step for the synthetic genome. Nature *473*, 403-408.

Batada, N.N., and Hurst, L.D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet *39*, 945-949.

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of transcript variability in single mammalian cells. Cell *163*, 1596-1610.

Becskei, A., Kaufmann, B.B., and van Oudenaarden, A. (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat Genet *37*, 937.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi–C: a comprehensive technique to capture the conformation of genomes. Methods *58*, 268-276.

Bernas, T., Brutkowski, W., Zarębski, M., and Dobrucki, J. (2014). Spatial heterogeneity of dynamics of H1 linker histone. Eur Biophys J *43*, 287-300.

Blake, W.J., Kærn, M., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633.

Brown, C.R., Mao, C., Falkovskaia, E., Jurica, M.S., and Boeger, H. (2013). Linking stochastic fluctuations in chromatin structure and gene expression. PLoS Biol *11*, e1001621.

Budnik, B., Levy, E., Harmange, G., and Slavov, N. (2018). Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. bioRxiv, 102681.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015a). ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol *109*, 21.29. 21-21.29. 29.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486-490.

Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., Blake, J.A., and Group, M.G.D. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res *36*, D724-D728.

Carrera, J., and Covert, M.W. (2015). Why build whole-cell models? Trends Cell Biol *25*, 719-722.

Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature *453*, 544.

Chen, X., and Zhang, J. (2016). The genomic landscape of position effects on protein expression level and noise in yeast. Cell Syst *2*, 347-354.

Cherry, J.L. (2010). Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol *2*, 757-769.

Dar, R.D., Razooky, B.S., Singh, A., Trimeloni, T.V., McCollum, J.M., Cox, C.D., Simpson, M.L., and Weinberger, L.S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. Proceedings of the National Academy of Sciences.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet *14*, 390.

Dekker, J., and Mirny, L. (2016). The 3D genome as moderator of chromosomal communication. Cell *164*, 1110-1121.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Eden, E., Geva-Zatorsky, N., Issaeva, I., Cohen, A., Dekel, E., Danon, T., Cohen, L., Mayo, A., and Alon, U. (2011). Proteome half-life dynamics in living human cells. Science *331*, 764-768.

Elf, J., and Barkefors, I. (2018). Single-molecule kinetics in living cells. Annu Rev Biochem.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Franks, A., Airoldi, E., and Slavov, N. (2017). Post-transcriptional regulation across human tissues. PLoS Comput Biol *13*, e1005535.

Fukuoka, Y., Inaoka, H., and Kohane, I.S. (2004). Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics *5*, 4.

Fyodorov, D.V., Zhou, B.-R., Skoultchi, A.I., and Bai, Y. (2018). Emerging roles of linker histones in regulating chromatin structure and function. Nature Reviews Molecular Cell Biology *19*, 192.

Gedeon, T., and Bokes, P. (2012). Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations. Biophys J *103*, 377-385.

Geman, S., and Geman, D. (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In Readings in Computer Vision (Elsevier), pp. 564-584.

Ghanbarian, A.T., and Hurst, L.D. (2015). Neighboring genes show correlated evolution in gene expression. Mol Biol Evol *32*, 1748-1766.

Gilks, W.R. (2005). Markov chain monte carlo. Encyclopedia of Biostatistics.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., *et al.* (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575-579.

Gout, J.F., Kahn, D., and Duret, L. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet *6*, e1000944.

Hager, G.L., McNally, J.G., and Misteli, T. (2009). Transcription dynamics. Mol Cell *35*, 741-753.

Hahn, S., and Kim, D. (2013). Physical origin of the contact frequency in chromosome conformation capture data. Biophys J *105*, 1786-1795.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., and Rozenblatt-Rosen, O. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol *17*, 77.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. Development *136*, 3853-3862.

Hurst, L.D., Pál, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet *5*, 299-310.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med *50*, 96.

Ibn-Salem, J., Muro, E.M., and Andrade-Navarro, M.A. (2016). Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res *45*, 81-91.

Kemkemer, R., Schrank, S., Vogel, W., Gruler, H., and Kaufmann, D. (2002). Increased noise as an effect of haploinsufficiency of the tumor-suppressor gene neurofibromatosis type 1 in vitro. Proc Natl Acad Sci U S A *99*, 13783-13788.

Kustatscher, G., Grabowski, P., and Rappsilber, J. (2017). Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol Syst Biol *13*, 937.

Lehner, B. (2008). Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol Syst Biol *4*, 170.

Lercher, M.J., and Hurst, L.D. (2006). Co-expressed yeast genes cluster over a long range but are not regularly spaced. J Mol Biol *359*, 825-831.

Lever, M.A., Th'ng, J.P., Sun, X., and Hendzel, M.J. (2000). Rapid exchange of histone H1. 1 on chromatin in living human cells. Nature *408*, 873.

Levesque, M.J., and Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nature methods *10*, 246.

Liao, B.-Y., and Zhang, J. (2008). Coexpression of linked genes in Mammalian genomes is generally disadvantageous. Mol Biol Evol *25*, 1555-1565.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell *165*, 535-550.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., and Martersteck, E.M. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202-1214.

Mahmutovic, A., Fange, D., Berg, O.G., and Elf, J. (2012). Lost in presumption: stochastic reactions in spatial models. Nature methods *9*, 1163.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res *24*, 496-510.

Medrzycki, M., Zhang, Y., Cao, K., and Fan, Y. (2012). Expression analysis of mammalian linker-histone subtypes. Journal of visualized experiments: JoVE.

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2009). BioNumbers—the database of key numbers in molecular and cell biology. Nucleic Acids Res *38*, D750-D753.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. Science *342*, 948-953.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst *23*, 263-286.

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol *17*, 47.

Phifer-Rixey, M., Bonhomme, F., Boursot, P., Churchill, G.A., Piálek, J., Tucker, P.K., and Nachman, M.W. (2012). Adaptive evolution and effective population size in wild house mice. Mol Biol Evol *29*, 2949-2955.

Phillips, R., Theriot, J., Kondev, J., and Garcia, H. (2012). Physical biology of the cell (Garland Science).

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc *9*, 171.

Popovic, D., Koch, B., Kueblbeck, M., Ellenberg, J., and Pelkmans, L. (2018). Multivariate Control of Transcript to Protein Variability in Single Mammalian Cells. Cell systems.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol *4*, e309.

Raj, A., Rifkin, S.A., Andersen, E., and Van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913-918.

Raj, A., Van Den Bogaard, P., Rifkin, S.A., Van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nature methods *5*, 877.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226.

Raser, J.M., and O'shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.

Reinius, B., Mold, J.E., Ramskold, D., Deng, Q., Johnsson, P., Michaelsson, J., Frisen, J., and Sandberg, R. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet *48*, 1430-1435.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2009). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res *38*, D497-D501.

Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bähler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. Nat Genet *36*, 809.

Sanchez, A., Choubey, S., and Kondev, J. (2013). Regulation of noise in gene expression. Annu Rev Biophys *42*, 469-491.

Sémon, M., and Duret, L. (2006). Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol *23*, 1715-1723.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. Nature *444*, 643.

Singer, G.A., Lloyd, A.T., Huminiecki, L.B., and Wolfe, K.H. (2004). Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol *22*, 767-775.

Slavov, N., Semrau, S., Airoldi, E., Budnik, B., and van Oudenaarden, A. (2015). Differential stoichiometry among core ribosomal proteins. Cell reports *13*, 865-873.

Söllner, J.F., Leparc, G., Hildebrandt, T., Klein, H., Thomas, L., Stupka, E., and Simon, E. (2017). An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Scientific data *4*, 170185.

Spellman, P.T., and Rubin, G.M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol *1*, 5.

Stewart-Ornstein, J., Nelson, C., DeRisi, J., Weissman, J.S., and El-Samad, H. (2013). Msn2 coordinates a stoichiometric gene expression program. Curr Biol *23*, 2336-2345.

Stewart-Ornstein, J., Weissman, J.S., and El-Samad, H. (2012). Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol Cell *45*, 483-493.

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. Science *332*, 472-474.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533-538.

Teichmann, S.A., and Veitia, R.A. (2004). Genes encoding subunits of stable complexes are clustered on the yeast chromosomes. Genetics *167*, 2121-2125.

Turing, A.M. (1952). The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences *237*, 37-72.

Veening, J.-W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol *62*, 193-210.

Veitia, R.A. (2010). A generalized model of gene dosage and dominant negative effects in macromolecular complexes. The FASEB Journal *24*, 994-1002.

Vosnakis, N., Koch, M., Scheer, E., Kessler, P., Mély, Y., Didier, P., and Tora, L. (2017). Coactivators and general transcription factors have two distinct dynamic populations dependent on transcription. The EMBO journal, e201696035.

Wang, Z., and Zhang, J. (2011). Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc Natl Acad Sci U S A *108*, E67-76.

Xu, J., Carter, A.C., Gendrel, A.-V., Attia, M., Loftus, J., Greenleaf, W.J., Tibshirani, R., Heard, E., and Chang, H.Y. (2017). Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells. Nat Genet *49*, 377-386.

Yan, C., Wu, S., Pocetti, C., and Bai, L. (2016). Regulation of cell-to-cell variability in divergent gene expression. Nat Commun *7*, 11099.

Zhang, Z., Qian, W., and Zhang, J. (2009). Positive selection for elevated gene expression noise in yeast. Mol Syst Biol *5*, 299.

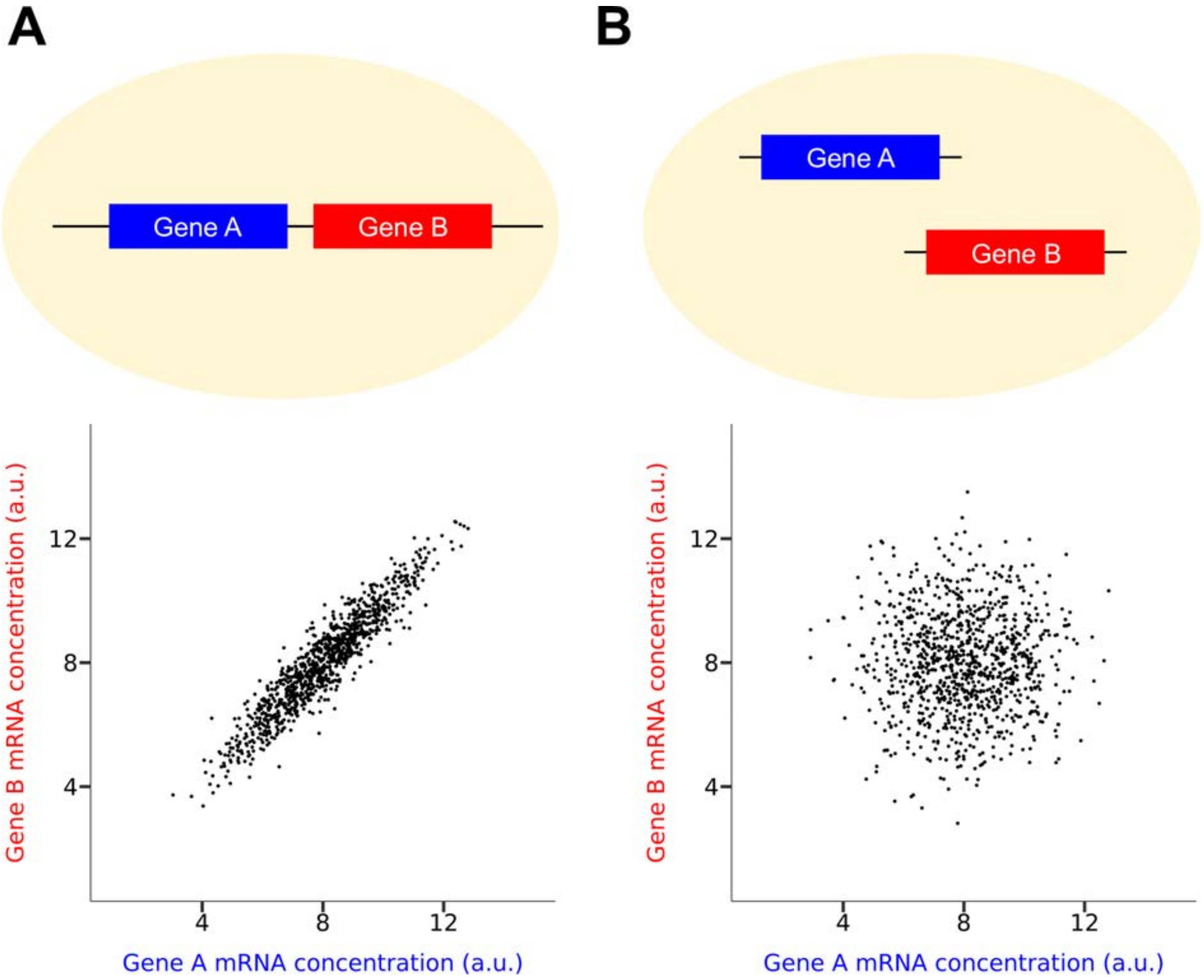
A
B
Gene A
Gene B
Gene A
Gene B
Gene B mRNA concentration (a.u.)
12
8
4
4
8
12
Gene A mRNA concentration (a.u.)
Gene B mRNA concentration (a.u.)
12
8
4
4
8
12
Gene A mRNA concentration (a.u.)

Fig.3-1 The hypothesized linkage effect on gene expression co-fluctuation. The cellular mRNA concentrations of two genes should be better correlated among isogenic cells in a population under a constant environment **(A)** when the two genes are chromosomally linked than **(B)** when they are unlinked. In the dot plot, each dot represents a cell.

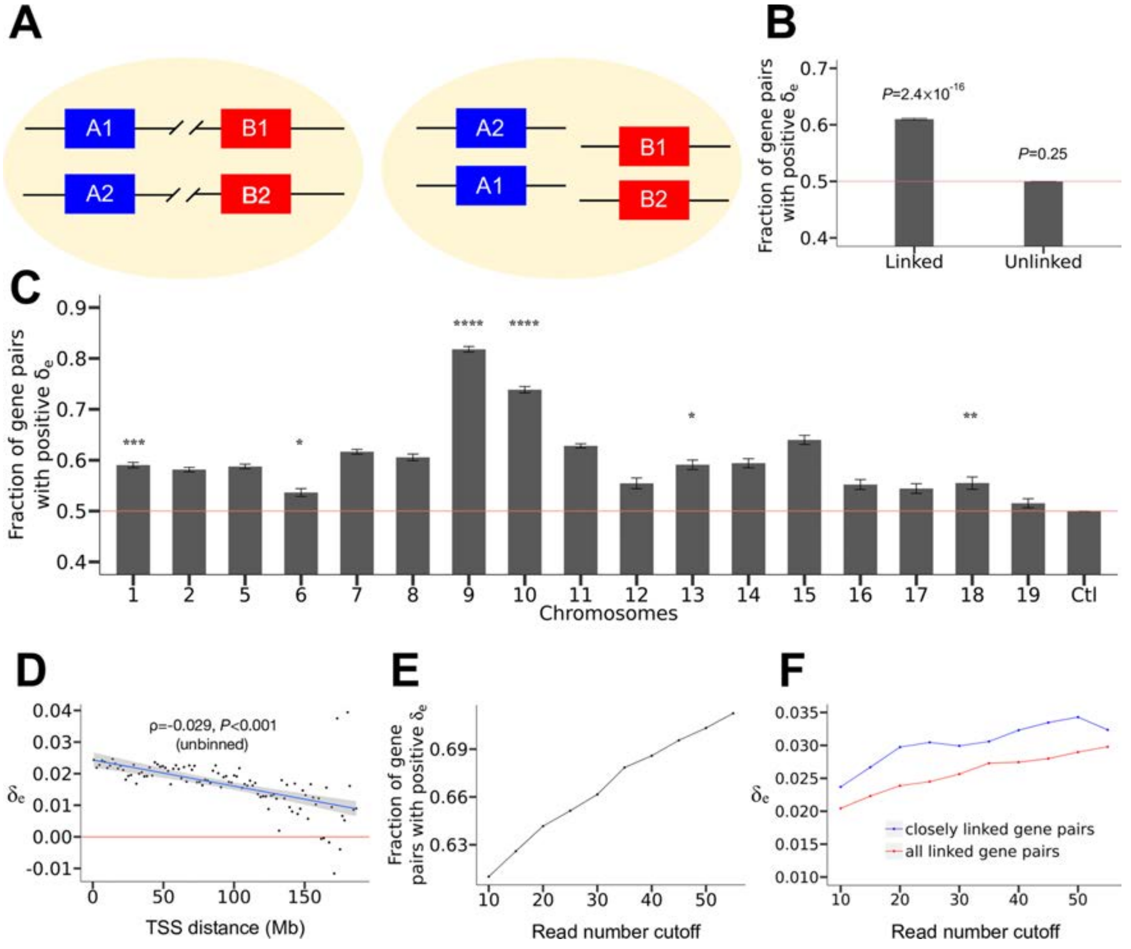

Fig.3-2 Chromosome-wide linkage effects on gene expression co-fluctuation in mouse fibroblast cells. **(A)** The logic of the method for testing the linkage effect. When gene *A* and gene *B* are linked, the correlations between the mRNA concentrations of the alleles of *A* and *B* that are physically linked (*cis*-correlations) should exceed the corresponding correlations of the alleles that are physically unlinked (*trans*-correlations). That is, δ_e = (sum of *cis*-correlations − sum of *trans*-correlations)/2 should be positive. This relationship should disappear if gene *A* and gene *B* are unlinked. **(B)** Fraction of gene pairs with positive δ_e. The red line represents the null expectation under no linkage effect. The confidence bands based on raw number of pairs are presented. *P*-values from binomial tests on independent gene pairs are presented. **(C)** Fraction of gene pairs with positive δ_e in each chromosome. The confidence bands based on raw number of pairs are presented. Binomial *P*-values are indicated as follows. NS, not significant; *, $0.01 < P < 0.05$; **, $0.001 < P < 0.01$; ***, $0.0001 < P < 0.001$; ****, $P < 0.0001$. The red line represents the null expectation under no linkage effect. The control (Ctl) shows the fraction of unlinked gene pairs with positive δ_e. **(D)** Median δ_e in a bin decreases with the median genomic distance of linked genes in the bin. All bins have the same genomic distance interval. TSS, transcription start site. The blue line shows the linear regression of the binned data, and the confidence band is presented. Spearman's ρ from unbinned data and associated *P*-value determined by a shuffling test are presented. **(E)** Fraction of linked gene pairs showing positive δ_e increases with the minimal number of reads per allele required. **(F)** Median δ_e for all linked gene pairs (red) and median δ_e in the left-most bin of panel D (blue) increase with the minimal read number per allele required.

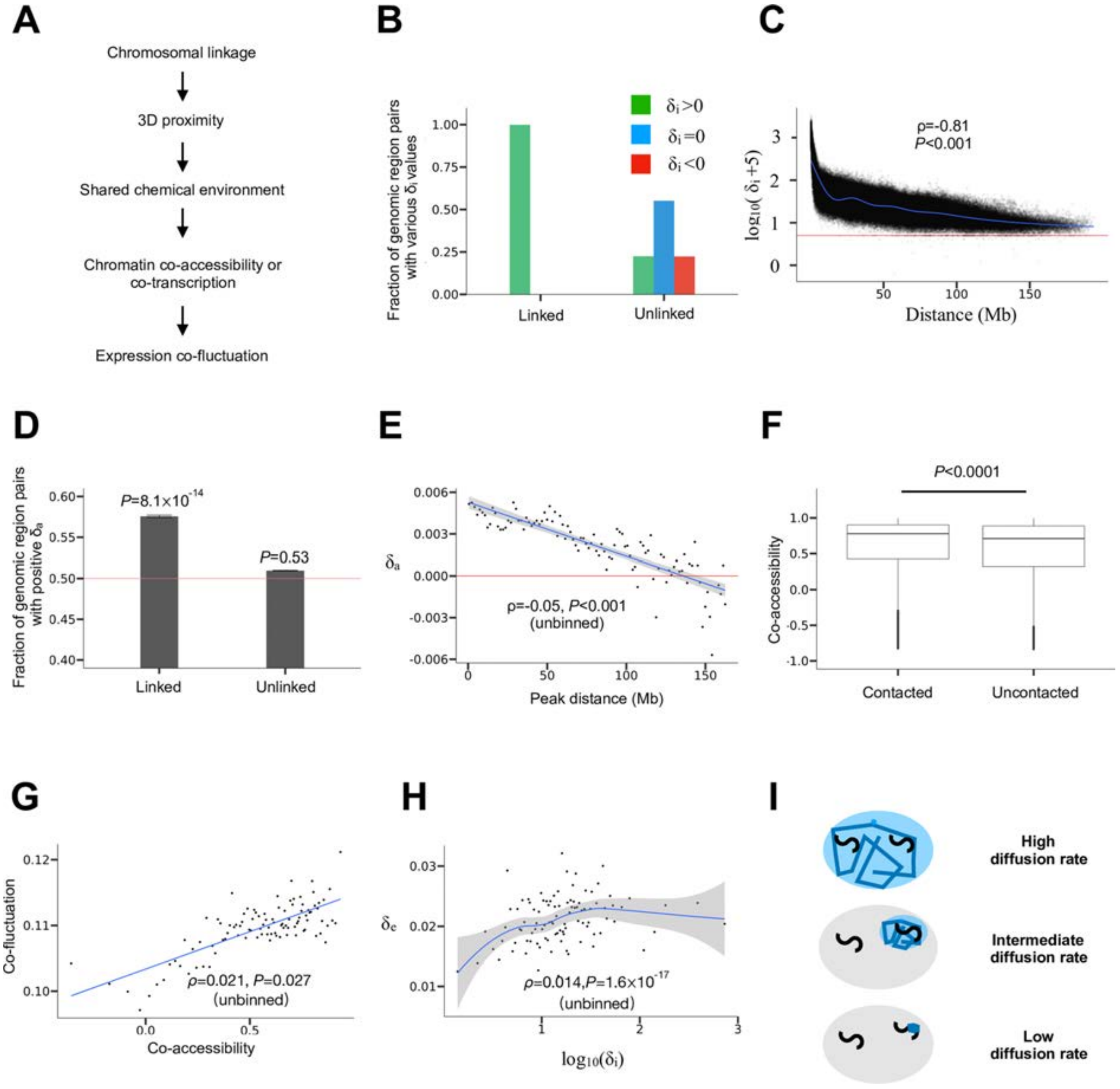
A
Chromosomal linkage
3D proximity
Shared chemical environment
Chromatin co-accessibility or co-transcription
Expression co-fluctuation
B
Fraction of genomic region pairs with various δi values
1.00
0.75
0.50
0.25
0.00
🟩 δi >0
🟦 δi =0
🟥 δi <0
Linked
Unlinked
C
log10(δi +5)
ρ=-0.81
P<0.001
0
1
2
3
50
100
150
Distance (Mb)
D
Fraction of genomic region pairs with positive δa
0.60
0.55
0.50
0.45
0.40
P=8.1×10^-14
P=0.53
Linked
Unlinked
E
δa
0.006
0.003
0.000
-0.003
-0.006
ρ=-0.05, P<0.001
(unbinned)
0
50
100
150
Peak distance (Mb)
F
P<0.0001
Co-accessibility
1.0
0.5
0.0
-0.5
-1.0
Contacted
Uncontacted
G
Co-fluctuation
0.12
0.11
0.10
ρ=0.021, P=0.027
(unbinned)
0.0
0.5
Co-accessibility
H
δe
0.03
0.02
0.01
ρ=0.014,P=1.6×10^-17
(unbinned)
1
2
3
log10(δi)
I
High diffusion rate
Intermediate diffusion rate
Low diffusion rate

Fig.3-3 Mechanistic basis of the linkage effect on expression co-fluctuation. **(A)** A model on how chromosomal linkage causes expression co-fluctuation. **(B)** Fractions of linked or unlinked genomic region pairs with positive, 0, and negative δ_i values, respectively. δ_i = (sum of *cis*-interactions − sum of *trans*-interactions)/2, where chromatin interactions are based on Hi-C data. All fractions are shown, but the blue and red bars for linked regions are too low to be visible. **(C)** δ_i decreases with the genomic distance between the linked regions considered. Each dot represents one pair of linked genomic regions. Shown here is $\log_{10}(\delta_i + 5)$ because δ_i is occasionally negative and it decreases with genomic distance very quickly. The horizontal red line indicates $\delta_i = 0$. The blue line is a cubic spline regression of δ_i on the genomic distance. Spearman's ρ from unbinned data and associated *P*-value determined by a shuffling test are presented. **(D)** Fraction of linked or unlinked pairs of ATAC peaks with positive δ_a. δ_a = (sum of *cis*-correlations in accessibility − sum of *trans*-correlations in accessibility)/2. The confidence bands calcuated from the raw number of pairs are presented. *P*-values from binomial tests on independent peak pairs are presented. The red line shows the fraction of 0.5. **(E)** δ_a decreases with the distance between linked ATAC peaks. Each dot represents a bin. All bins have the same distance interval. The red line shows $\delta_a = 0$. The blue line shows the linear regression of the binned data. For better viewing, one bin (X=156, Y= -0.02) is not shown; the extreme δ_a of the bin is probably due to the small sample size of the bin ($n = 13$). Spearman's ρ computed from unbinned data and associated *P*-value determined from a shuffling test are presented. **(F)** Co-accessibility (*trans*-r_a) is greater for 3D contacted (*trans*-$F > 0$) than uncontacted (*trans*-$F = 0$) non-allelic genomic regions located on homologous chromosomes. The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). *P*-value is determined by a Mantel test. **(G)** Expression co-fluctuation (*trans*-r_e) improves with the co-accessibility (*trans*-r_a) of non-allelic ATAC peaks located on homologous chromosomes. Each dot represents a bin. All bins have the same distance interval. The blue line shows the linear regression of the binned data. The confidence band is presented. Spearman's ρ computed from unbinned data and associated *P*-value determined by a Mantel test are presented. **(H)** Expression co-fluctuation is positively correlated with ln (Hi-C contact frequency). The blue line is a linear regression of expression co-fluctuation and ln (Hi-C contact frequency). The confidence band is presented. **(I)** Diffusion rates for molecules responsible for the chromosome-wide linkage effect should be neither too high nor too low. If the diffusion is too fast, the concentration of the molecule will be similar across the nucleus (top); if the diffusion is too slow, the concentration cannot even be similar for loci loosely linked on the same chromosome (bottom). Only when the diffusion rate is intermediate, the local chemical environment could be homogeneous for genes on the same chromosome but heterogeneous for genes on different chromosomes (middle). The large oval represents the nucleus and each black "S" curve represents a chromosome. Blue zig-zags show molecular diffusions, while the blue area depicts a chemically homogenous environment.

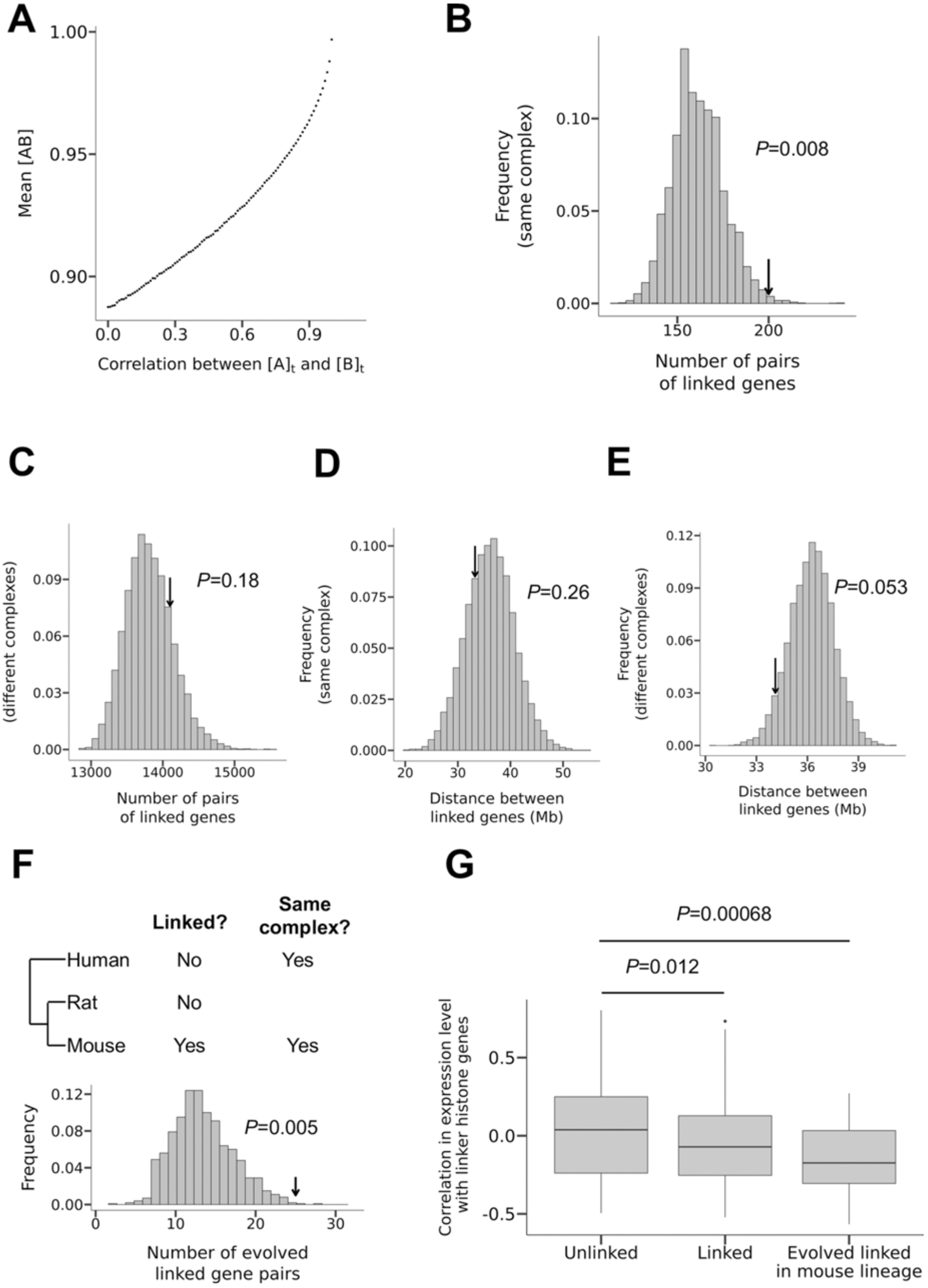
A
1.00
0.95
0.90
Mean [AB]
0.0
0.3
0.6
0.9
Correlation between $[A]_t$ and $[B]_t$
B
Frequency (same complex)
0.10
0.05
0.00
150
200
P=0.008
Number of pairs of linked genes
C
Frequency (different complexes)
0.09
0.06
0.03
0.00
13000
14000
15000
P=0.18
Number of pairs of linked genes
D
Frequency (same complex)
0.100
0.075
0.050
0.025
0.000
20
30
40
50
P=0.26
Distance between linked genes (Mb)
E
Frequency (different complexes)
0.12
0.09
0.06
0.03
0.00
30
33
36
39
P=0.053
Distance between linked genes (Mb)
F
Linked?
Same complex?
Human
No
Yes
Rat
No
Mouse
Yes
Yes
Frequency
0.12
0.08
0.04
0.00
0
10
20
30
P=0.005
Number of evolved linked gene pairs
G
P=0.00068
P=0.012
Correlation in expression level with linker histone genes
0.5
0.0
-0.5
Unlinked
Linked
Evolved linked in mouse lineage

Fig.3-4 Genes encoding components of the same protein complex tend to be chromosomally linked. **(A)** Mean concentration of the protein complex AB ([AB]) in 10,000 cells increases with the co-fluctuation of the concentrations of its two components measured by the correlation of the total concentration of protein A ($[A]_t$) and that of B ($[B]_t$). **(B-C)** The frequency distribution of the number of pairs of linked genes encoding components of the same protein complex **(B)** and components of different protein complexes **(C)** in 10,000 randomly shuffled genomes. Arrows indicate the observed values. **(D-E)** The frequency distribution of the median distance between two linked genes that encode components of the same protein complex **(D)** and components of different protein complexes **(E)** in 10,000 randomly shuffled genomes. Arrows indicate the observed values. **(F)** Test of the hypothesis of protein complex-driven evolution of gene linkage, which asserts that the probability for an originally unlinked pair of genes to become linked is higher if they encode members of the same protein complex. Of 875 pairs of genes that are unlinked in both human and rat and encode members of the same protein complex in both human and mouse, 25 become linked in mouse, as indicated by the arrow. The frequency distribution of the corresponding expected number is shown by the distribution. **(G)** Protein complex genes that are linked with at least one gene encoding a member of the same complex tend to be highly expressed in tissues with low abundances of linker histones. Y-axis shows the correlation in expression level between protein complex genes and the linker histone genes across tissues.

Chapter 4: Preferred Synonymous Codons Are Translated More Accurately: Proteomic Evidence, Among-Species Variation, and Mechanistic Basis

I love the right words. I think economy and precision of language are important.

-Chelsea Clinton

4.1 Abstract

A commonly stated cause of the widespread phenomenon of unequal uses of synonymous codons is their differential translational accuracies. However, this long-standing translational accuracy hypothesis (TAH) of codon usage bias has had no direct evidence beyond anecdotes. Analyzing proteomic data from *Escherichia coli*, we observe higher translational accuracies of more frequently used synonymous codons, offering direct, global evidence for the TAH. The experimentally measured codon-specific translational accuracies validate a sequence-based proxy; this proxy provides support for the TAH from the vast majority of over 1000 taxa surveyed in all domains of life. We find that the relative translational accuracies of synonymous codons vary substantially among taxa and are strongly correlated with the amounts of cognate tRNAs relative to those of near-cognate tRNAs. These and other observations suggest a model in which selections for translational efficiency and accuracy drive codon usage bias and its coevolution with the tRNA pool.

4.2 Introduction

Eighteen of the 20 amino acids are each encoded by more than one codon, but the synonymous codons are usually unequally used in a genome(Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). Among the synonymous codons of an amino acid, those used more often than the average are referred to as preferred codons while the rest unpreferred. This phenomenon of codon usage bias (CUB), initially discovered over four decades ago from the first few determined gene sequences(Air et al., 1976; Efstratiadis et al., 1977; Fiers et al., 1975; Ikemura, 1981), is a result of the joint forces of mutation, genetic drift, and natural selection, but the specific selective agents have not been fully deciphered(Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). One long-standing hypothesis known as the translational accuracy hypothesis (TAH) asserts that different synonymous codons are translated with different accuracies and that CUB results at least in part from natural selection for translational accuracy(Akashi, 1994). Indeed, the importance of accurate protein translation cannot be overstated, because mistranslation may lead to the loss of normal protein functions and gain of cellular toxicity(Drummond and Wilke, 2009) and cause severe diseases including cancer and neurodegenerative diseases(Chen et al., 2011). In fact, several cellular mechanisms are known to ensure the overall fidelity of protein synthesis. For example, conformational changes of the ribosome decoding center can be more efficiently induced by cognate codon-anticodon interactions than near-cognate codon-anticodon interaction(Ibba and Söll, 1999), allowing discrimination against incorrect decoding. Additionally, the accuracy of many steps in translation, such as tRNA aminoacylation(Ibba and Söll, 1999) and codon-anticodon matching, is enhanced by the energy-consuming kinetic proofreading(Hopfield, 1974). Notwithstanding, even if synonymous codons differ in translational accuracy, relatively accurate synonymous codons may not be preferentially used. This is because synonymous codons differ in other properties such as the translational elongation speed(Hussmann et al., 2015; Weinberg et al., 2016); selection related to these other features(Qian et al., 2012) could triumph selection for translational accuracy.

Several groups have attempted to test the TAH of CUB. In particular, Akashi(Akashi, 1994) developed an indirect test based on the idea that the benefit of using relatively accurate codons should be greater at evolutionarily conserved amino acid sites than unconserved sites of the same protein; consequently, a higher usage of preferred codons at conserved sites than at unconserved sites supports the TAH. While Akashi's test is positive for several model organisms investigated(Akashi, 1994; Drummond and Wilke, 2008; Stoletzki and Eyre-Walker, 2007), this test does not directly compare translational accuracies of synonymous codons so cannot completely exclude alternative explanations(Akashi, 1994; Shah and Gilchrist, 2010). In an early study, Precup and Parker used site-directed mutagenesis followed by peptide sequencing to show that AAU, an unpreferred codon of Asn, is misread as Lys 4-9 times more often than is AAC, a preferred codon of Asn, at a particular position of the coat protein gene of the bacteriophage MS2 under Asn starvation(Precup and Parker, 1987). Similarly, Kramer and Farabaugh observed that AAU has a significantly higher rate of mistranslation to Lys than AAC at a particular position of a reporter gene in *Escherichia coli*(Kramer and Farabaugh, 2007). Nonetheless, Kramer and Farabaugh also observed that the unpreferred Arg codons of CGA and CGG and the preferred Arg codons of CGU and CGC exhibited similar rates of mistranslation to Lys(Kramer and Farabaugh, 2007). While the above experiments directly tested the TAH, they were each based on the investigation of one amino acid site of one protein, so its genome-wide generality is unknown. As such, a direct and global test of the TAH is needed.

Capitalizing on a proteome-wide probe of mistranslation in *E. coli*(Mordret et al., 2019), we here provide direct evidence that preferred codons are generally translated more accurately than unpreferred codons. We then use the *E. coli* data to validate a sequence-based proxy for relative translational accuracies of synonymous codons. Using this proxy, we show that the TAH of CUB is supported in the vast majority of over 1000 diverse taxa surveyed, but that the relative translational accuracies of synonymous codons vary substantially among taxa. We find that the relative translational accuracy of a synonymous codon is strongly correlated with its cognate tRNA abundance relative to near-cognate tRNA abundance. These and other results suggest a model in which selections for translational efficiency and accuracy drive the CUB and its coevolution with the tRNA pool.

4.3 Results

4.3.1 Preferred codons are more accurately decoded

A direct test of the TAH of CUB requires comparing the mistranslation rate among synonymous codons. Using mass spectrometry, Mordret *et al.* quantified mistranslations at individual sites of the *E. coli* proteome(Mordret et al., 2019). After removing sites and codons where mistranslation rates cannot be quantified due to technical reasons (see Methods), we grouped mistranslation events according to the identities of their original codons. We then computed the absolute mistranslation rate of a codon as the ratio of the total intensity of mistranslated peptides to that of all peptides mapped to the codon. Finally, we computed the relative mistranslation rate (*RMR*) of a codon by dividing its absolute mistranslation rate by the mean absolute mistranslation rate of all codons coding for the same amino acid. *RMR* >1 means that the codon has a higher mistranslation rate than the average among all codons for the same amino acid, whereas *RMR* <1 means the opposite. Codon usage was assessed by the relative synonymous codon usage (*RSCU*). The *RSCU* of a codon equals its frequency in the genome relative to the average frequency of all codons for the same amino acid(Sharp et al., 1986). A codon with *RSCU* >1 is preferred while a codon with *RSCU* <1 is unprefered.

We were able to estimate the *RMR* for 27 codons of nine amino acids (**Fig. 4-1a**). Except for Gly, the most preferred synonymous codon of an amino acid shows *RMR* <1, providing a significant support for the TAH (P = 0.020, one-tailed binomial test). Similarly, except for Gly and Val, the least prevalent synonymous codon of an amino acid shows *RMR* >1 (P = 0.090, one-tailed binomial test). Because both *RSCU* and *RMR* of a codon are relative to the mean of all codons for the same amino acid, they can be compared among codons of different amino acids. Indeed, a strong negative correlation was observed between *RSCU* and *RMR* among the 27 codons (Pearson's r = -0.56, P < 0.001, permutation test; Spearman's ρ = -0.49, P = 0.005, permutation test; **Fig. 4-1b**). Together, these findings from the proteomic data of *E. coli* demonstrate that preferred codons tend to have lower mistranslation rates, supporting the TAH of CUB.

4.3.2 Relative translational accuracies of synonymous codons vary across taxa

How do certain synonymous codons achieve higher translational accuracies than others? There are two general scenarios. In the first scenario, referred hereinafter as the constant accuracy hypothesis, the translational accuracy is intrinsically higher for a synonymous codon than another because of their different chemical nature(Hershberg and Petrov, 2009). Consequently, the relative translational accuracies of synonymous codons should be more or less the same in different species. For instance, given that AAA (Lys) is more accurate than AAG (Lys) in *E. coli* (**Fig. 4-1a**), we expect the same trend in the vast majority if not all species. Alternatively, relative translational accuracies of synonymous codons may be greatly influenced by species-specific factors such as the tRNA pool. Under this scenario, referred to as the variable accuracy hypothesis hereinafter, the relative accuracies of synonymous codons vary among species. That is, AAA is more accurate than AAG in many species but the opposite is true in many other species.

Measuring the relative translational accuracies of synonymous codons in a large number of species will allow differentiating between the above two hypotheses, which will in turn help understand the mechanism underlying the translational accuracy differences among synonymous codons. Because codon-specific, proteome-based translational accuracies have not been measured beyond *E. coli*, we resort to a sequence-based proxy referred to as the odds ratio (*OR*) that originated from Akashi's test(Akashi, 1994). Specifically, the *OR* of synonymous codon X that encodes amino acid Y in a gene is the number of times that X is used at invariant Y sites relative to the number of times that X is not used at invariant Y sites, divided by the number of times that X is used at variant Y sites relative to the number of times that X is not used at variant Y sites (**Fig. 4-2a**). Here, invariant and variant Y sites refer to Y sites in the focal species whose counterparts in the ortholog from a related species have Y and non-Y, respectively. The *OR* values computed from individual genes can be combined to yield a single *OR* using the Mantel-Haenszel procedure (see Methods). While *OR* was originally developed for preferred codons, it can be computed for any codon of the 18 amino acids that have multiple synonymous codons(Qian et al., 2012). Based on Akashi's test, *OR* has been used as a proxy for the relative translational accuracy of a codon(Qian et al., 2012). To verify the relationship between *OR* and relative translational accuracy, we computed *OR* values by aligning *E. coli* genes with their *Salmonella enterica* orthologs. Indeed, for the 27 codons with *RMR* estimates, *OR* and *RMR* are strongly negatively correlated (r = -0.63, P = 0.001; ρ = -0.43, P = 0.01; **Fig. 4-2b**), confirming that the *OR* of a codon is a valid proxy for its relative translational accuracy.

To examine whether the relative translational accuracies of synonymous codons vary among species, we took advantage of a recently built phylogenetic tree of 10,575 microbial taxa(Zhu et al., 2019). Because most taxa (9,906) in the tree are from the domain Bacteria, we first focused our analysis on Bacteria. We picked all 1,197 pairs of sister bacterial taxa from the tree and aligned their orthologous genes (see Methods). We randomly assigned one taxon in each pair as the focal taxon and computed *OR* for each codon as described above. We found a positive correlation between *RSCU* and *OR* across codons in 95% of the taxa examined (**Fig. 4-2c**), demonstrating an overwhelming support for the TAH of CUB in Bacteria.

We computed ln(*OR*) to make its distribution relatively symmetric to aid visualization, and examined as an example ln(*OR*) for codon CAT (His) in each of the focal taxa arranged according to the bacterial tree (one taxon per order is presented in **Fig. 4-2d**). We found ln(*OR*) to vary greatly from negative values to positive values, with a high density near 0 (**Fig. 4-2e**). Furthermore, the extreme values of ln(*OR*) (bright red and bright green in **Fig. 4-2d**) are scattered across the tree rather than concentrated in a few clades, suggesting that the relative translational accuracy of CAT has changed substantially and frequently in evolution. The across-taxon variation of *OR* indicates that CAT is the relatively inaccurate one of the two synonymous codons of His in many taxa (red in **Fig. 4-2d**) but the relatively accurate one in many other taxa (green), supporting the variable accuracy hypothesis. From **Fig. 4-2e**, which shows the 18 amino acids each with multiple codons, it is clear that the pattern observed for CAT applies to all codons. Furthermore, every codon has *OR* >1 in at least 8.9% of the taxa examined (**Fig. A3-1a**). These results thus support the variable accuracy hypothesis for all synonymous codons. The above observations of *OR* variation among taxa are not primarily caused by sampling error, because a similar pattern was detected when we analyzed a subset of taxa for each amino acid where the number of occurrences of each synonymous codon considered in *OR* estimation is at least 1000 per taxon (**Fig. A3-1b**). They are not mainly caused by genetic drift either, because a similar pattern was found when we analyzed a subset of taxa with strong signals of selection for translational accuracy (correlation between *RSCU* and *OR* exceeding 0.5) (**Fig. A3-1c**). It is worth pointing out that, despite the general support for the variable accuracy hypothesis, for a minority of codons such as ATA (Ile), AGA (Arg), and AGG (Arg), the distribution of ln(*OR*) is strongly skewed toward negative values (**Fig. 4-2e**), suggesting that their relative translational accuracies are somewhat constrained although not invariable in evolution.

To investigate if the above observations from Bacteria are generalizable to the other two domains of life, we first expanded our analysis to Archaea represented in the large phylogeny mentioned(Zhu et al., 2019). We found that the correlation between *RSCU* and *OR* is positive in 90% of taxa examined and that ln(*OR*) varies greatly across taxa for each codon (**Fig. A3-2**), further supporting the TAH and the variable accuracy hypothesis. For Eukaryota, we analyzed five commonly used model organisms: human, mouse, worm, fly, and budding yeast (see Methods). In each of these species, the correlation between *RSCU* and *OR* is significantly positive (**Table A3-1**), supporting the TAH. Except for the two mammals, which are closely related, the *OR*s estimated from one species are not well correlated with those estimated from another species (**Fig. A3-3**). Furthermore, the correlation in *OR* generally declines with the divergence time between the two species (**Fig. A3-3**), consistent with the variable accuracy hypothesis. Taken together, our results show that the TAH is generally supported in all domains of life but the relative translational accuracies of synonymous codons vary across taxa.

4.3.3 Mechanistic basis of among-codon and across-taxon variations of translational accuracies

The empirical support for the variable accuracy hypothesis strongly suggests that the determinants of the *RMR*s of synonymous codons vary among species. In the aforementioned Kramer-Farabaugh study(Kramer and Farabaugh, 2007), the authors found that artificially increasing the expression level of the cognate tRNA for Arg codons AGA and AGG reduces their mistranslations to Lys, so proposed that the competition between cognate and near-cognate tRNAs determines the mistranslation rate of a codon. Here, the cognate tRNA is the tRNA whose anticodon pairs with the codon correctly (allowing wobble pairing), whereas the near-cognate tRNA corresponds to a different amino acid and has an anticodon that mismatches the codon at one position. Consistent with the above proposal, Mordret *et al.*(Mordret et al., 2019) inferred that most of the mistranslation events in *E. coli* arose from mispairing between codons and near-cognate tRNAs. They further noted that, for certain types of mistranslation, there is a negative correlation across growth phases between the mistranslation rate and the ratio ($R_{\mathrm{c/nc}}$) in abundance between cognate and near-cognate tRNAs, although the correlation was rarely statistically significant(Mordret et al., 2019). Based on these past observations, we hypothesize that the relative translational accuracy of a synonymous codon increases with its relative $R_{\mathrm{c/nc}}$, or $RR_{\mathrm{c/nc}}$, which is $R_{\mathrm{c/nc}}$ divided by the mean $R_{\mathrm{c/nc}}$ of all codons coding for the same amino acid (see Methods). We further hypothesize that, because the tRNA pool varies substantially among species(Chan and Lowe, 2009), the among-species variation of relative translational accuracies arises from the among-species variation in $RR_{\mathrm{c/nc}}$.

To test the above hypotheses, we computed $RR_{\mathrm{c/nc}}$ for each codon using published tRNA expression levels in *E. coli*(Mordret et al., 2019). Indeed, we observed a significant negative correlation between $RR_{\mathrm{c/nc}}$ and *RMR* (r = -0.47, P = 0.009; ρ = -0.53, P = 0.005; **Fig. 4-3a**) and a significant positive correlation between $RR_{\mathrm{c/nc}}$ and *OR* (r = 0.49, P = 0.07; ρ = 0.74, P = 0.00001; **Fig. 4-3b**) across codons, supporting the hypothesis that the relative ratio of cognate to near-cognate tRNA abundances is a major determinant of a codon's relative translational accuracy in *E. coli*. Note that the relative cognate tRNA abundance alone is not significantly correlated with *RMR* (r = -0.22, P = 0.2; ρ = -0.04, P = 0.4; **Fig. A3-4a**), supporting the role of competition between cognate and near-cognate tRNAs in determining *RMR*. As previously reported(Qian et al., 2012), the relative cognate tRNA level is highly correlated with *RSCU* ($r = 0.61$, $P = 0.03$; $\rho = 0.48$, $P = 0.02$; **Fig. A3-4b**), which is likely a result of selection for high translational efficiency (i.e., more codons translated per unit time per cell) because balanced codon usage relative to cognate tRNA concentrations maximizes translational efficiency(Qian et al., 2012).

We next investigated whether the above finding in *E. coli* applies to other bacterial taxa surveyed in Fig. 4-2. Because tRNA expression levels are unknown for the vast majority of these taxa, we used the gene copy number of each tRNA species as a proxy for the total expression level of the tRNA species(Tuller et al., 2010). Indeed, *E. coli* $RR_{\mathrm{c/nc}}$ computed from tRNA gene copy numbers is highly correlated with that computed from tRNA expression levels ($r = 0.77$, $P = 1.64\times10^{-6}$; $\rho = 0.90$, $P = 1.36\times10^{-10}$). Furthermore, *E. coli* $RR_{\mathrm{c/nc}}$ computed from tRNA gene copy numbers is significantly correlated with *RMR* (**Fig. 4-3c**), confirming the validity of using this proxy. We obtained the tRNA gene annotations for 1094 of the 1197 focal bacterial taxa examined in Fig. 4-2. However, in many of these taxa, there is little tRNA gene redundancy or variation in cognate tRNA gene copy number among synonymous codons despite considerable CUB; in these taxa, the tRNA gene copy number is unlikely a good proxy for tRNA abunadnce(Wei et al., 2019). Because the tRNA gene copy number is a good proxy for tRNA abundance in *E. coli*, which has 85 tRNA genes, we decided to filter out taxa with fewer than 81 tRNA genes to strike a balance between the noise level and number of taxa in our analysis. This filtering left us with 59 taxa, in each of which we correlated the *OR* of a codon with its $RR_{\mathrm{c/nc}}$ computed from tRNA gene copy numbers. The vast majority (92%) of the taxa show a positive correlation (**Fig. 4-3d**), supporting the generality of our hypothesis on the role of $RR_{\mathrm{c/nc}}$ in determining the relative translational accuracy of a codon in Bacteria.

To investigate whether the above finding is generalizable to other domains of life, we analyzed tRNA genes in Archaea taxa and Eukaryotic model organisms. Unfortunately, no

Archaea taxa examined have more than 80 tRNA genes. For each of the five eukaryotes (human, mouse, fly, worm, and yeast), the correlation between *OR* and $RR_{\mathrm{c/nc}}$ computed from tRNA gene copy numbers is significantly positive for linear or rank correlation (**Table A3-2**). Together, our findings strongly support that, in the diverse taxa surveyed, the ratio of cognate tRNA abundance to near-cognate tRNA abundance is generally a major determinant of the relative translational accuracy of a codon. Hence, the variation of the tRNA pool among species can explain the across-species variation of the relative translational accuracies of synonymous codons.

4.4 Discussion

Analyzing published proteomic data from *E. coli*, we provided direct, global evidence that preferred synonymous codons are generally decoded more accurately than unpreferred codons. We found that relative translational accuracies of synonymous codons vary substantially among species, supporting the variable accuracy hypothesis. We obtained strong evidence that the ratio of cognate tRNA abundance to near-cognate tRNA abundance is a major determinant of a codon’s relative translational accuracy. Hence, the variable accuracies observed are mechanistically explained by the variation of the tRNA pool across species. These findings, together with the previous report on the selection for translational efficiency(Qian et al., 2012), suggest a model in which the tRNA pool and codon usage coevolve to improve both translational efficiency and accuracy (**Fig. 4-4a**). Specifically, mutation and drift can alter both codon frequencies and tRNA concentrations. The cellular translational efficiency is maximized when (transcriptomic) codon frequencies equal relative cognate tRNA concentrations(Qian et al., 2012), whereas the translational accuracy of a codon is maximized when the ratio of its cognate tRNA concentration to near-cognate tRNA concentration is maximized. Under this model, selections for translational efficiency and translational accuracy are related but not perfectly aligned, which could introduce tradeoffs between translational efficiency and accuracy(Yang et al., 2014). Indeed, our simulation of a simple genetic system with two amino acids, each encoded by two synonymous codons (**Fig. 4-4b**), found that imposing a selection for translational accuracy can lower translational efficiency (**Fig. 4-4c**).

Interestingly, our results imply that, even in the absence of selection for translational accuracy, the positive correlation between synonymous codon frequency and cognate tRNA concentration resulting from selection for translational efficiency(Qian et al., 2012) will likely render the cognate tRNA concentration relative to near-cognate tRNA concentration higher for more frequently used synonymous codons. Consequently, the positive correlation between the relative codon frequency and relative translational accuracy may arise in the absence of selection for translational accuracy. In fact, in *E. coli*, for 16 of the 18 amino acids with multiple synonymous codons, the codon with the highest cognate tRNA concentration has the highest $RR_{\mathrm{c/nc}}$. Upon shuffling the expression levels among tRNA species, we found that, for over one half of the 18 amino acids, the codon with the highest cognate tRNA concentration has the highest $RR_{\mathrm{c/nc}}$. This was true in each of 1000 shufflings. Nevertheless, in only 6 of these 1000 shufflings did all 18 amino acids exhibit the above feature. Thus, a high but non-perfect concordance between translational efficiency and accuracy is expected. Therefore, strictly speaking, the correlation in Fig. 4-1b by itself does not prove selection for translational accuracy. However, this correlation, in conjunction with the correlation between *RSCU* and *OR*, demonstrates that evolutionarily conserved sites tend to use preferred synonymous codons, which tend to be relatively accurately translated, hence proving the role of selection for translational accuracy in causing CUB, or the TAH. How codon usage and the tRNA pool evolve under the joint forces of selections for translational efficiency and accuracy in addition to mutation and drift is quite complex. For instance, because any tRNA is simultaneously a cognate tRNA for one or more codons and a near-cognate tRNA for some other codons, increasing the translational accuracy of a particular codon might be at the expense of the translational accuracy of another codon. Indeed, a previous study showed that artificially increasing the cognate tRNA expression levels for arginine codons can result in proteotoxic stress(Yona et al., 2013). This subtle tradeoff could cause non-independent uses of codons of different amino acids. This was indeed observed in the simulation aforementioned (**Fig. 4-4d)**. Future modeling work with realistic parameters might shed more light on this issue. In addition to impacting translational efficiency and accuracy, synonymous mutations also affect mRNA folding(Park et al., 2013), mRNA stability(Presnyak et al., 2015), mRNA concentration(Chen et al., 2017; Presnyak et al., 2015; Zhou et al., 2016), pre-mRNA splicing(Chamary et al., 2006), and co-translational protein folding(Buhr et al., 2016; Walsh et al., 2020), so additional selections may shape CUB and its evolution.

Our study has several caveats. First, in our calculation of a codon's mistranslation rate, we lumped all mistranslations of the codon regardless of the erroneous amino acid it is translated to. Because different mistranslations of the same codon likely have differential fitness costs and because selection for translational accuracy likely minimizes the total fitness reduction caused by mistranslation instead of the mistranslation rate *per se*, properly weighting different mistranslations in *RMR* calculation will likely strengthen its correlation with *RSCU*. Second, when calculating the ratio of cognate tRNA concentration to near-cognate tRNA concentration, we did not consider the difference in interaction strength between different codons and anticodons(Reis et al., 2004). Future research that takes into account this interaction under physiological conditions may significantly improve the signal in the correlation analysis of Fig. 4-3. Third, our analysis in Fig. 4-3d was limited to taxa with >80 tRNA genes. Future research using tRNA expression levels(Wei et al., 2019) when they become available can confirm if the same pattern holds for taxa with fewer tRNA genes.

Finally, due to data limitation, we did not consider tRNA expression variations across environments, cell cycle stages, or tissues(Gingold et al., 2014). In the future, it would be interesting to study how such variations simultaneously impact translational efficiency and accuracy.

Our results might help design organisms with expanded code tables(Ros et al., 2020). Expanding the code table is realized by introducing unnatural tRNAs that are charged with non-canonical amino acids. The introduction of these tRNAs often leads to fitness defects due to mistranslation of normal codons(Chin, 2017). Our research suggests that one way to alleviate the proteotoxic stress is to identify potential near-cognate codons that could be mistranslated by the unnatural tRNA and adjust the natural tRNA pool to minimize the impact.

4.5 Methods

4.5.1 Estimating relative mistranslation rates of synonymous codons from *E. coli* proteomic data

The proteomic data analyzed came from Table A3-1 in Modret *et al.*(Mordret et al., 2019). The authors separately measured mistranslation events from high-solubility and low-solubility proteins using mass spectrometry, and both groups of events were considered in our analysis. We focused on the data from the wild-type strain BW25113 in the MOPS complete medium because (i) this dataset is the largest among datasets from all strain-medium combinations and (ii) no artificial perturbation such as mutation, drug, or amino acid depletion was applied(Mordret et al., 2019). We first removed sites that cannot be traced to a unique original codon. We also filtered out sites showing an intensity of “NaN” for the unmodified (aka base) peptide or mistranslated (aka dependent) peptide. Because different synonymous codons tend to generate different mistranslations by mispairing with different near-cognate tRNAs, if these different mistranslations have different detection probabilities, the comparison between synonymous codons would be unfair. Unfortunately, some mistranslations produce mass shifts indistinguishable from post-translational modifications so cannot be reliably identified through mass spectrometry(Wei et al., 2019), which would produce exactly this situation in some cases. Therefore, we removed amino acids with undetectable mistranslations except for Leu and Ile. We kept these two amino acids because the only undetectable mistranslations for them are Leu to Ile and Ile to Leu, both can be considered benign due to the high physicochemical similarity between Leu and Ile(Henikoff and Henikoff, 1992). Considering the structure of the genetic code table, we found that the underestimation of the mistranslation rate due to the negligence of mistranslations between Leu and Ile is severer for unpreferred than preferred codons, suggesting that the actual strength of evidence for higher mistranslation rates of unpreferred than preferred synonymous codons is stronger than what is shown in Fig. 4-1. We then computed each codon's absolute mistranslation rate by dividing the total intensity of mistranslated (i.e., dependent) peptides by that of all (i.e., dependent + base) peptides mapped to the codon. We divided each codon's absolute mistranslation rate by the mean absolute mistranslation rate of all codons coding for the same amino acid to obtain the codon’s relative mistranslation rate (*RMR*). We removed amino acids without data for all of its synonymous codons because calculating *RMR* requires having data for all synonymous codons of an amino acid. In total, we computed *RMR* for 27 codons of 9 amino acids.

4.5.2 Relative synonymous codon usage (*RSCU*), odds ratio (*OR*), and relative ratio of cognate tRNA concentration to near-cognate tRNA concentration (*RR*$_{\mathbf{c/nc}}$) for *E. coli*

Peptide and cDNA sequences of *E. coli* (genome assembly ASM584v2) and *S. enterica* (genome assembly ASM78381v1) were downloaded from Ensembl Bacteria(Howe et al., 2021). We computed *RSCU* of codon *j* of amino acid *i* from all coding sequences of *E. coli* by $RSCU_{i,j} = \frac{n_i x_{i,j}}{\sum_{j=1}^{n_i} x_{i,j}}$, where n_i is the number of synonymous codons of amino acid *i* and

$x_{i,j}$ is the number of codon *j* of amino acid *i* in all coding sequences(Sharp et al., 1986). Conventionally, *RSCU* is computed from highly expressed genes(Sharp et al., 1986). However, due to the lack of gene expression information from most of the species analyzed, we computed *RSCU* from all genes. This should not qualitatively affect our analysis, because *RSCU* computed from highly expressed genes (e.g., the top 20% of genes) is nearly perfectly correlated with that computed from all genes (e.g., in *E. coli*, $r = 0.96$, $P < 2.2\times10^{-16}$).

To calculate the *OR* of each codon, we first identified one-to-one orthologous proteins between *E. coli* and *S. enterica* using OrthoFinder(Emms and Kelly, 2019). Next, we aligned these one-to-one orthologs using MUSCLE(Edgar, 2004), separating all amino acid sites into conserved and non-conserved sites. For a focal codon in gene *i*, we tabulated a_i, number of times the focal codon is observed at conserved amino acid sites; b_i, number of times the focal codon is observed at unconserved sites; c_i, total number of times the focal codon's synonymous codons are observed at conserved sites; and d_i, total number of times the focal codon's synonymous codons are observed at unconserved sites. Here, the focal codon's synonymous codons do not include itself. *OR* for gene *i* equals $(a_i d_i)/(b_i c_i)$. Using the Mantel-Haenszel procedure, we combined the odds ratios of the focal codon from individual genes into one odds ratio(Akashi, 1994) by $OR = \frac{\sum_i \frac{a_i d_i}{(a_i+b_i+c_i+d_i)}}{\sum_i \frac{b_i c_i}{(a_i+b_i+c_i+d_i)}}$.

To compute $RR_{\mathrm{c/nc}}$ of a codon, we tabulated the cognate tRNAs and near-cognate tRNAs of the codon. Cognate tRNAs are all tRNAs that can pair with the focal codon allowing wobble pairing at the 3rd codon position, while near-cognate tRNAs are tRNAs coded for a different amino acid but can pair with the focal codon with one base-pair mismatch (allowing wobble pairing at the 3rd codon position). We then weighted each tRNA by their average relative expression levels across three growth stages in the MOPS complete media (GEO number: GSE128812). Finally, we normalized the ratio for each codon by the average ratio of all codons coding for the same amino acid.

4.5.3 *RSCU, OR*, and *RR*$_{\text{c/nc}}$ for other species

RSCU, *OR*, and *RR*$_{\text{c/nc}}$ were calculated for non-*E. coli* taxa as for *E. coli*, with the differences noted below. For the non-*E. coli* prokaryotic taxa, we downloaded the phylogenetic tree of 10,575 taxa from the Web of Life(Zhu et al., 2019) (https://biocore.github.io/wol/) and identified sister taxa from the tree. Briefly, each pair of sister taxa are two taxa that are the single closest relative to each other in the tree. For each pair of sister taxa, we downloaded from the same web site their protein-coding DNA sequences, protein sequences, and tRNA gene copy number data. For eukaryotic model organisms, we downloaded protein-coding DNA sequences and protein sequences of human (*Homo sapiens*), mouse (*Mus musculus*), fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), and budding yeast (*Saccharomyces cerevisiae*) from the NCBI refseq database(O'Leary et al., 2016). We further downloaded the protein sequences of *Macaca mulatta* (as a relative of *H. sapiens*), *Rattus norvegicus* (as a relative of *M. musculus*), *Drosophila erecta* (as a relative of *D. melanogaster*), *Caenorhabditis briggsae* (as a relative of *C. elegans*), and *Saccharomyces paradoxus* (as a relative of *S. cerevisiae*) from the NCBI refseq database. The tRNA gene annotations in the five model organisms were downloaded from GtRNAdb(Chan and Lowe, 2009). *RR*$_{\text{c/nc}}$ was computed using tRNA gene copy numbers instead of tRNA expression levels.

4.5.4 Statistical analysis

Many of the quantities estimated in our work, such as *RMR*, *RR*$_{\text{c/nc}}$, *RSCU*, and *OR*, are not independent among synonymous codons. To deal with this non-independence in statistical tests, we applied permutation tests. Specifically, in **Fig. 4-1b**, we generated 1000 permuted samples by shuffling the absolute mistranslation rates among all codons and then re-estimated *RMR* values. We then computed the correlation between *RMR* and *RSCU* in each permuted sample while holding the *RSCU* value of each codon unchanged. *P* equals the fraction of permuted samples with the correlation coefficient more negative than that observed in the original sample. Similarly, when testing the correlation between *RMR* and *OR* (**Fig. 4-2b**), we shuffled the absolute mistranslation rate among all codons and recomputed *RMR* while holding the *OR* for each codon unchanged. When testing the correlation between *RMR* (or *OR*) and $RR_{\text{c/nc}}$ (**Fig. 4-3**), we shuffled the absolute mistranslation rates among codons and the expression levels (or gene copy numbers) among tRNAs. Finally, when testing the correlation between *RMR* (or *RSCU*) and relative cognate tRNA concentration (**Fig. A3-4**), we shuffled the absolute mistranslation rate among codons and the expression level among tRNAs.

To estimate the standard error (SE) of the *RMR* of each codon, we constructed 1000 bootstrap samples by resampling the sites in the original data with replacement. Similarly, we estimated the SE of the *OR* of each codon by constructing 1000 bootstrapped *E. coli* genomes via resampling its genes that have one-to-one orthologs in *S. enterica*.

4.5.5 Simulation

To assess the impact of selections for translational accuracy and efficiency on codon usage, we built a toy model with two amino acids: aa_0 and aa_1. Amino acid aa_0 is encoded by synonymous codons 00 and 01 while aa_1 is encoded by synonymous codons 10 and 11 (**Fig. 4-4b**). Codon-anticodon pairing follows the rule that 0 pairs with 1 and vice versa. The cognate tRNA of a codon has an anticodon that pairs perfectly with the codon, while the near-cognate tRNA has an anticodon that pairs with the codon with exactly one mismatch and carries the other amino acid.

We considered a unicellular organism with one gene consisting of n codons. We assumed that the mRNA level of the gene does not change in the evolution simulated and that ribosomes are in shortage. We defined the organismal fitness as follows.

$$Absolute\ fitness = Function - Cost,\ \text{where}$$

$$Function = TE \times \sum_{i=1}^{n} f_i \text{ and } Cost = \sum_{i=1}^{n} c_i.$$

Here, f_i and c_i are the function and cost of codon i, respectively. We set $f_i = F_i$ if codon i encodes the pre-specified optimal amino acid at the codon; otherwise, $f_i = 0$. For each i, F_i is a random variable sampled from an exponential distribution with the mean equal to 1 (Eyre-Walker and Keightley, 2007). Following a previous study(Qian et al., 2012), we set the expected codon selection time per amino acid aa$_0$ as $t_0 = p_1^2/q_1+p_2^2/q_2$, where p_1 and $p_2 = 1-p_1$ are the fractions of amino acid aa$_0$ encoded by codon 00 and 01, respectively, and q_1 and q_2 $=1-q_1$ are the fractions of corresponding cognate tRNAs among all tRNAs of aa$_0$, respectively. We similarly set the expected codon selection time per amino acid aa$_1$ and computed the total codon selection time of all codons. Translational efficiency TE, which is the number of codons translated per unit time, is the inverse of the total codon selection time. We set $c_i = TE \times C_i$ if codon i does not encode the pre-specified optimal amino acid at the codon; otherwise, $c_i = C_i \times TE \times \frac{1}{RR_{c/nc}}$. When there is no selection for translational accuracy, $C_i = 0$; otherwise, C_i for codon i is a random variable sampled from an exponential distribution with mean equal to 1. Note that C_i and F_i are independent from each other. $RR_{c/nc}$ is computed as described in Results; the inverse of $RR_{c/nc}$ measures the mistranslation rate.

We started the simulation with a coding sequence of 200 nucleotides, coding for 100 amino acids. Each site had a 50% chance to be 0 or 1. For simplicity, we assumed that the initial amino acid sequence is optimal such that the evolution in our simulation is largely about codon usage. For each of the four different tRNAs (with anticodons of 00, 01, 10, and

11, respectively), we sampled the initial copy number from 1, 2, and 3 with equal probabilities.

Next, we simulated the coevolution between the tRNA pool and codon usage following a strong selection, weak mutation regime. We first generate a mutation. With a probability of 0.02, it alters the copy number of a tRNA. In this case, we randomly pick a tRNA species to change its copy number by +1 or -1 with equal probabilities unless the copy number is 1, in which case it is +1. With a probability of 0.98, the mutation is a random point mutation at a randomly picked site of the coding sequence. The fitness of the mutant is then computed following the above fitness definition. The mutation is fixed with a probability of $\frac{1-r^{-1}}{1-r^{-N}}$, where r is the ratio of the absolute fitness of the mutant to that of the wild-type and N is the population size(Moran, 1958). The above mutation-selection process was repeated 100,000 rounds in each simulation to reach an equilibrium. For each N, we simulated 200 times with and 200 times without selection for translational accuracy.

4.6 References

Air, G.M., Blackburn, E.H., Coulson, A.R., Galibert, F., Sanger, F., Sedat, J.W., and Ziff, E.B. (1976). Gene F of bacteriophage phiX174. Correlation of nucleotide sequences from the DNA and amino acid sequences from the gene product. J Mol Biol *107*, 445-458.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics *136*, 927-935.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M.V., and Komar, A.A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Mol Cell *61*, 341-351.

Chamary, J.V., Parmley, J.L., and Hurst, L.D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet *7*, 98-108.

Chan, P.P., and Lowe, T.M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res *37*, D93-D97.

Chen, B., Retzlaff, M., Roos, T., and Frydman, J. (2011). Cellular strategies of protein quality control. Cold Spring Harb Perspect Biol *3*, a004374.

Chen, S., Li, K., Cao, W., Wang, J., Zhao, T., Huan, Q., Yang, Y.F., Wu, S., and Qian, W. (2017). Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. Mol Biol Evol *34*, 2944-2958.

Chin, J.W. (2017). Expanding and reprogramming the genetic code. Nature *550*, 53-60.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341-352.

Drummond, D.A., and Wilke, C.O. (2009). The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet *10*, 715-724.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Efstratiadis, A., Kafatos, F.C., and Maniatis, T. (1977). The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. Cell *10*, 571-585.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol *20*, 1-14.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat Rev Genet *8*, 610-618.

Fiers, W., Contreras, R., Duerinck, F., Haegmean, G., Merregaert, J., Jou, W.M., Raeymakers, A., Volckaert, G., Ysebaert, M., Van de Kerckhove, J., *et al.* (1975). A-protein gene of bacteriophage MS2. Nature *256*, 273-278.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sorensen, K.D., *et al.* (2014). A dual program for translation regulation in cellular proliferation and differentiation. Cell *158*, 1281-1292.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915-10919.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annu Rev Genet *42*, 287-299.

Hershberg, R., and Petrov, D.A. (2009). General rules for optimal codon choice. PLoS Genet *5*, e1000556.

Hopfield, J.J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proceedings of the National Academy of Sciences *71*, 4135-4139.

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., and Bhai, J. (2021). Ensembl 2021. Nucleic Acids Res *49*, D884-D891.

Hussmann, J.A., Patchett, S., Johnson, A., Sawyer, S., and Press, W.H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLoS Genet *11*, e1005732.

Ibba, M., and Söll, D. (1999). Quality control mechanisms during translation. Science *286*, 1893-1897.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol *151*, 389-409.

Kramer, E.B., and Farabaugh, P.J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA *13*, 87-96.

Moran, P.A.P. (1958). Random processes in genetics. Paper presented at: Mathematical proceedings of the cambridge philosophical society (Cambridge University Press).

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G.D., Cox, J., Geiger, T., Lindner, A.B., and Pilpel, Y. (2019). Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. Mol Cell *75*, 427-441. e425.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., and Ako-Adjei, D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-D745.

Park, C., Chen, X., Yang, J.R., and Zhang, J. (2013). Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proc Natl Acad Sci U

S A *110*, E678-686.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet *12*, 32-42.

Precup, J., and Parker, J. (1987). Missense misreading of asparagine codons as a function of codon identity and context. J Biol Chem *262*, 11351-11355.

Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., *et al.* (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111-1124.

Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet *8*, e1002603.

Reis, M.d., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res *32*, 5036-5044.

Ros, E., Torres, A.G., and de Pouplana, L.R. (2020). Learning from Nature to Expand the Genetic Code. Trends Biotechnol.

Shah, P., and Gilchrist, M.A. (2010). Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. PLoS Genet *6*, e1001128.

Sharp, P.M., Tuohy, T.M., and Mosurski, K.R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res *14*, 5125-5143.

Stoletzki, N., and Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol *24*, 374-381.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344-354.

Walsh, I.M., Bowman, M.A., Soto Santarriaga, I.F., Rodriguez, A., and Clark, P.L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. Proc Natl Acad Sci U S A *117*, 3528-3534.

Wei, Y., Silke, J.R., and Xia, X. (2019). An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Sci Rep *9*, 1-11.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Rep *14*, 1787-1799.

Yang, J.R., Chen, X., and Zhang, J. (2014). Codon-by-codon modulation of translational speed and accuracy via mRNA folding. PLoS Biol *12*, e1001910.

Yona, A.H., Bloom-Ackermann, Z., Frumkin, I., Hanson-Smith, V., Charpak-Amikam, Y., Feng, Q., Boeke, J.D., Dahan, O., and Pilpel, Y. (2013). tRNA genes rapidly change in evolution to meet novel translational demands. Elife *2*, e01339.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.H., Fu, J., Chen, S., and Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A *113*, E6117-E6125.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J.G., Belda-Ferre, P., Al-Ghalith, G.A., Kopylova, E., and McDonald, D. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature communications *10*, 1-14.

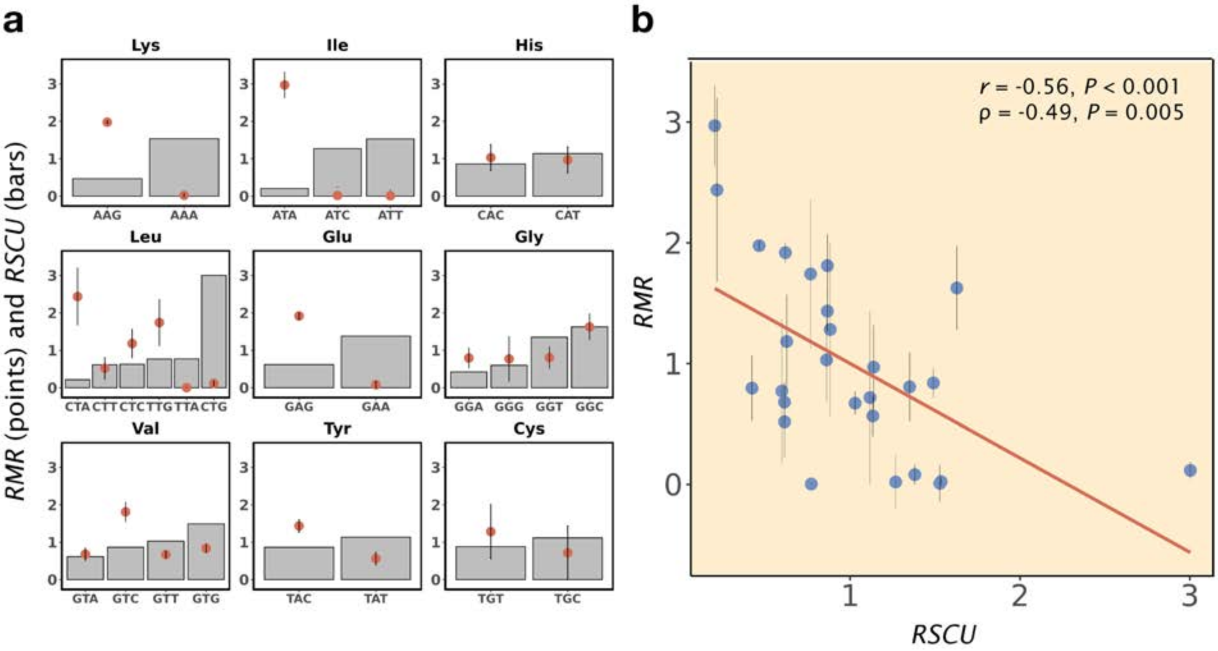

Fig. 4-1 More frequently used synonymous codons tend to be decoded more accurately in *E. coli*. **a**, Comparison of relative synonymous codon usage (*RSCU*, bars) and relative mistranslation rate (*RMR*, dots) among synonymous codons for nine amino acids with proteome-based *RMR* estimates. **b**, A significant negative correlation between *RSCU* and *RMR* across the 27 codons in panel a. The red line is the linear regression. In both panels, error bars represent one standard error estimated by the bootstrap method. The standard error of *RSCU* estimated by the bootstrap method is negligible due to the large number of each codon in the genome, so is not shown. *P*-values are based on permutation tests.

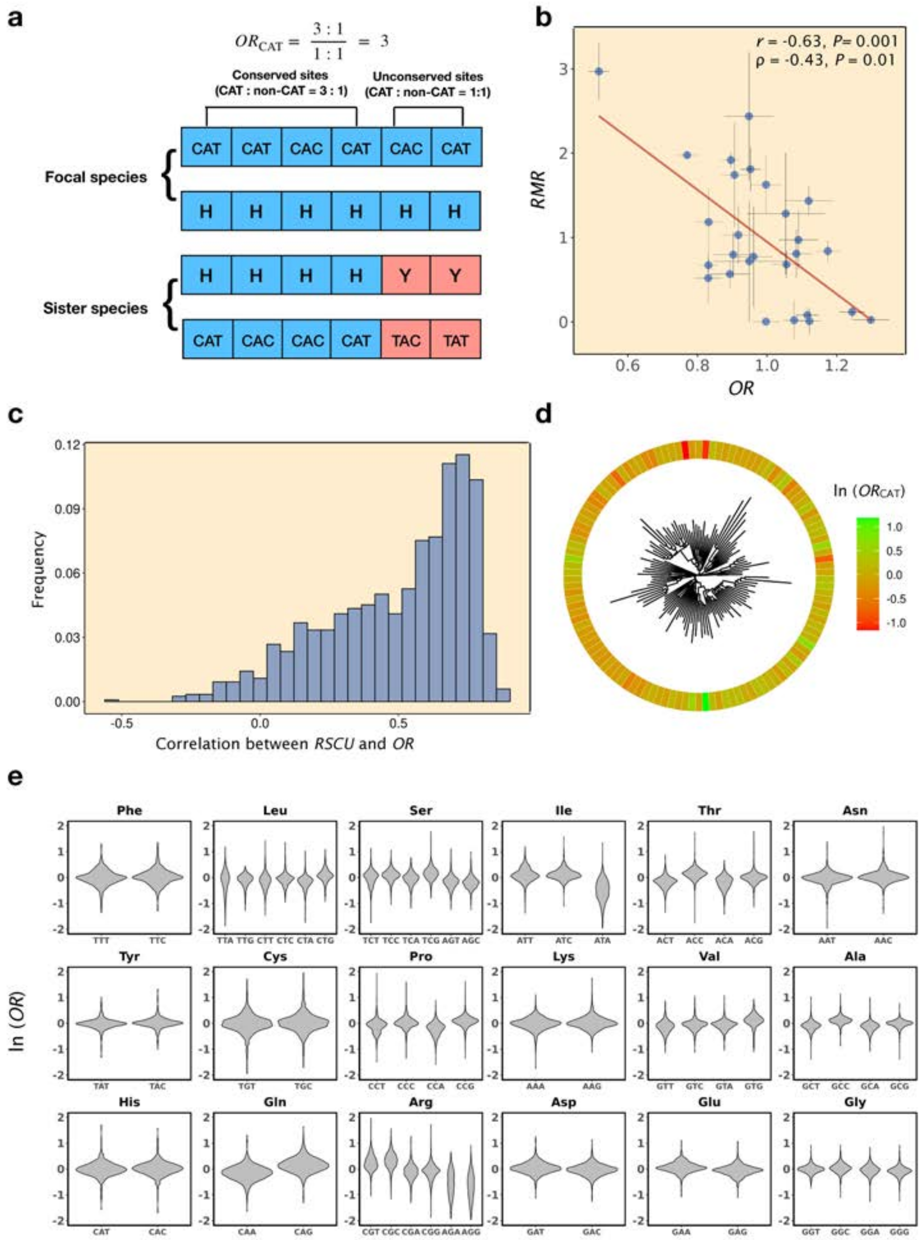

Fig 4-2 Variation of relative translational accuracies of synonymous codons across taxa. **a**, Diagram explaining the calculation of odds ratio (*OR*) of the codon CAT that serves as a proxy for its relative translational accuracy. Showing here is a hypothetical alignment of orthologous proteins (and the underlying coding sequences) between the focal species and a related species. **b**, *OR* is negatively correlated with *RMR* across codons in *E. coli*. *P*-values are based on permutation tests. The red line is the linear regression. **c**, Frequency distribution of Pearson's correlation between *RSCU* and *OR* in 1197 bacterial taxa. Ninety-five percent of these taxa show positive correlations. **d**, ln(*OR*) of codon CAT for each of 118 bacterial taxa, one per order, arranged according to their phylogeny shown in the middle. **e**, Violin plots showing frequency distributions of ln(*OR*) of individual codons of the 18 amino acids that have multiple synonymous codons across 1197 bacterial taxa.

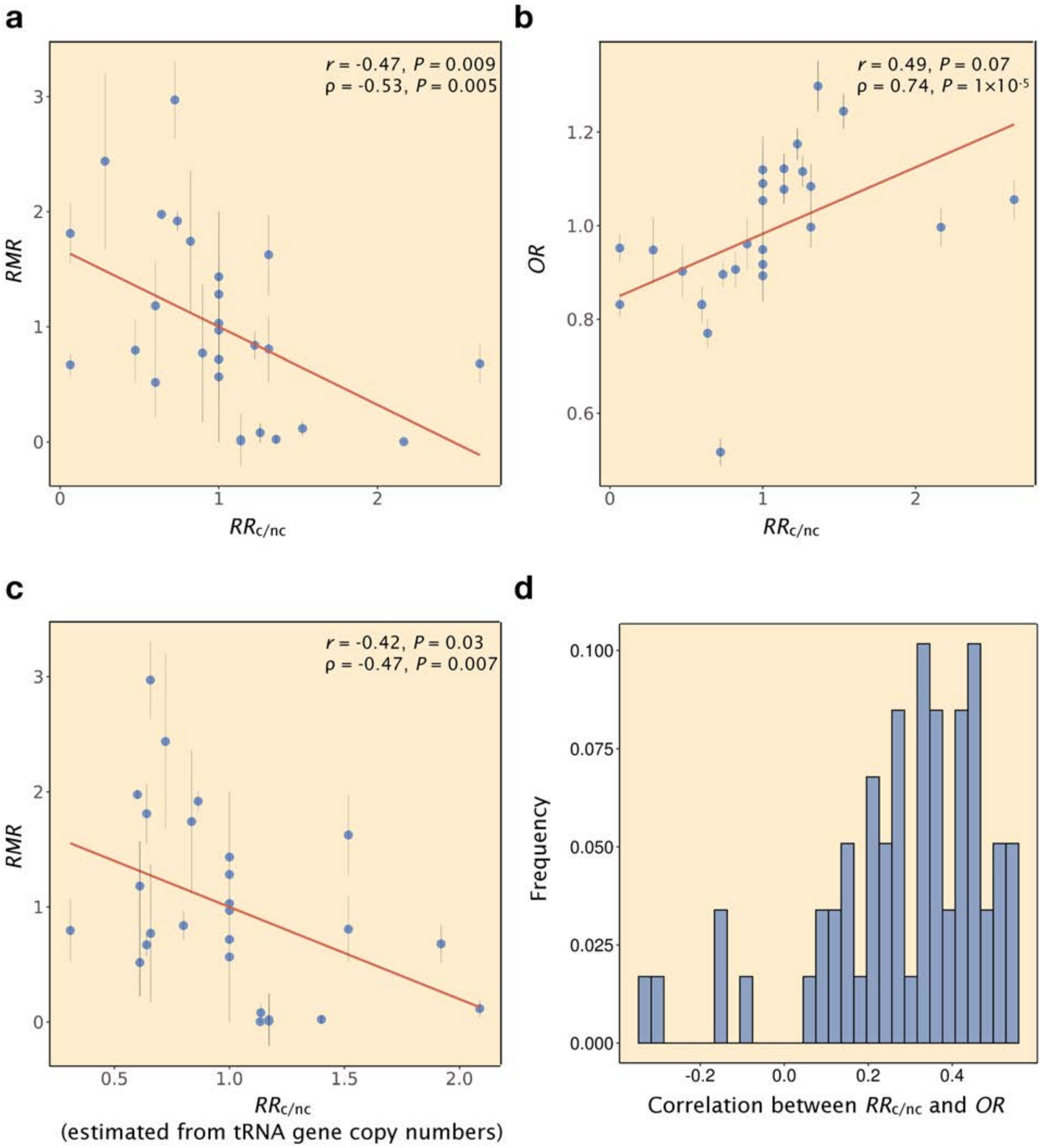

Fig. 4-3 The relative ratio of cognate tRNA concentration to near-cognate tRNA concentration ($RR_{\mathrm{c/nc}}$) is a major determinant of a codon's relative translational accuracy. **a**, *RMR* is negatively correlated with $RR_{\mathrm{c/nc}}$ across codons in *E. coli*. **b**, *OR* is positively correlated with $RR_{\mathrm{c/nc}}$ across codons in *E. coli*. **c**, *RMR* is negatively correlated with $RR_{\mathrm{c/nc}}$ computed using tRNA gene copy numbers instead of tRNA concentrations in *E. coli*. **d**, Frequency distribution of Pearson's correlation between *OR* and $RR_{\mathrm{c/nc}}$ computed using tRNA gene copy numbers in bacterial taxa with >80 tRNA genes. All *P*-values are based on permutation tests. In a-c, the red line is the linear regression.

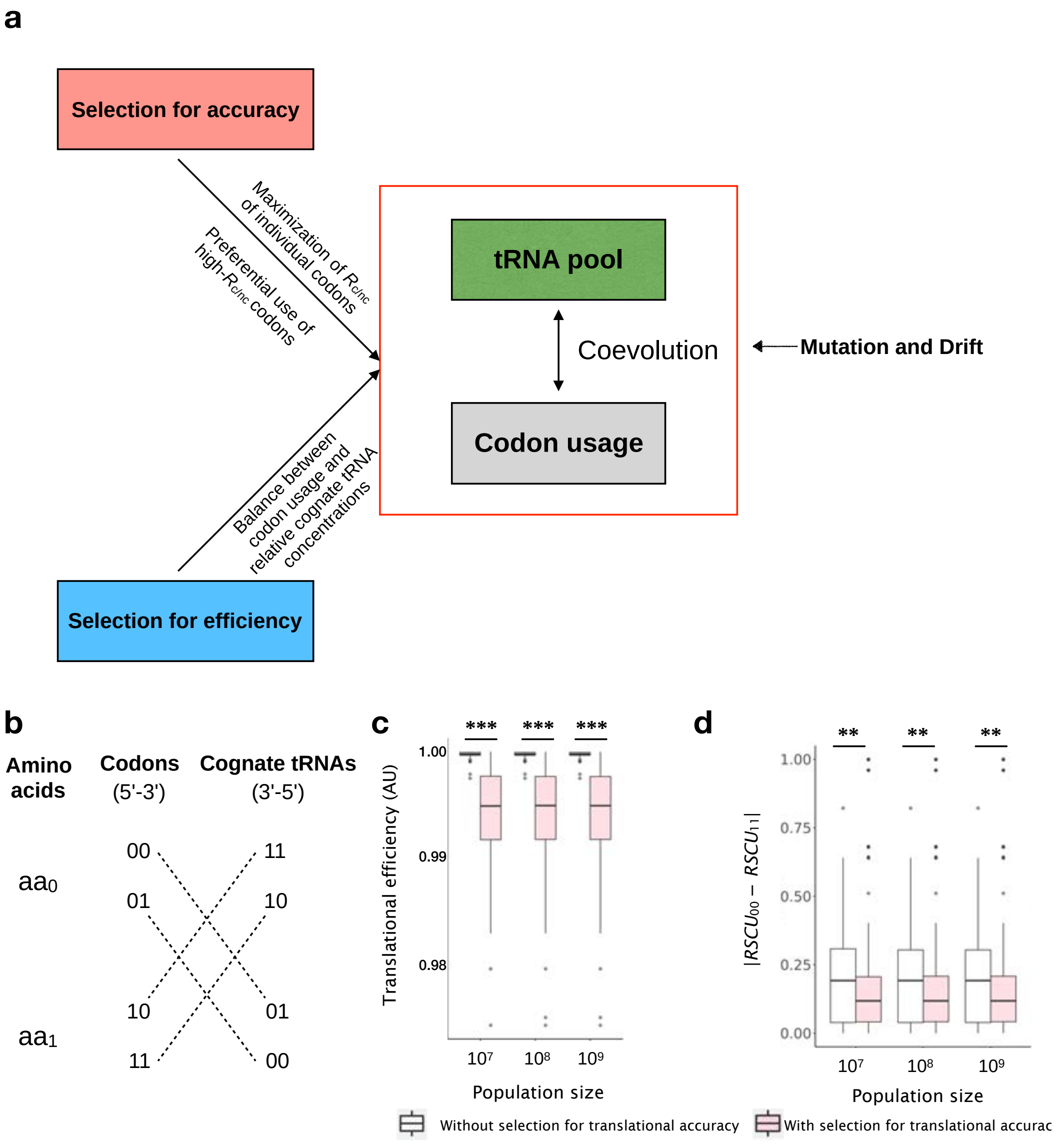

Fig. 4-4 Selections for translational efficiency and accuracy shape the tRNA pool and codon usage. **a**, A model for the coevolution of the tRNA pool and codon usage driven by selections for translational efficiency and accuracy. **b.** A toy model with two amino acids, each encoded by two synonymous codons. A dotted line connects a codon with its near-cognate tRNA. **c**, Translational efficiency is significantly lower in the presence of selection for translational accuracy than in the absence of this selection. **d,** The absolute difference between the *RSCU* of 00 ($RSCU_{00}$) and that of 11 ($RSCU_{11}$) is smaller under the selection for translational accuracy than without this selection. With the selection, codon usage for aa_0 and that for aa_1 become coupled, because selection disfavors the cognate tRNA of the common codon of aa_0 to become the near-cognate tRNA of the common codon of aa_1, and vice versa. In c and d, each box plot shows the distribution from 200 replicates. The lower and higher edges of a box represent the first (qu_1) and third (qu_3) quartiles, respectively; the horizontal line inside the box indicates the median (md); the whiskers extend to the most extreme values inside inner fences, md ± 1.5(qu_3-qu_1); and the dots are outliers. *, $0.01 \leq P < 0.05$, Wilcoxon rank-sum tests; **, $0.001 \leq P < 0.01$; ***, $P < 0.001$.

Chapter 5: Conclusions

Sometimes we may learn more from a man's errors than from his virtues.

-Henry Wadsworth Longfellow

Summary

In my dissertation, I studied the genome-wide patterns of expression errors, with specific focuses on (i) stochastic gene expression and (ii) protein mistranslation. Three central tenets emerged from my analysis. First, errors are pervasive in biological systems. All kinds of errors I studied happen on a genome-wide scale: every gene has noise (Raser and O'shea, 2005), each pair of genes co-fluctuates to some extent (Stewart-Ornstein et al., 2012), and every codon has a non-negligible rate of mistranslation (Milo and Phillips, 2015). Second, these errors have consequences and are usually harmful. This tenet is best illustrated by the evidence for natural selection in minimizing the harms for each kind of error I studied. To be concrete, in Chapter 2, I showed that noise level varies across genes associated with different functions and is anticorrelated with the expected level of harm of each functional group. In Chapter 3, I showed that natural selection optimized genome order to alleviate the deleterious effects of dosage imbalance among cells for protein complex genes. Last but not least, in Chapter 4, I provided direct evidence for the hypothesis that preferred codons are translated more accurately, which is a hypothesis that relies on the second tenet. Third, natural selection for minimizing errors is likely a difficult optimization problem, and a perfect solution is unlikely to arise. This is most apparent in Chapter 4, where I showed that lowering the error rate of one codon might be at the expense of increasing the error rate of other codons. Furthermore, pervasive trade-offs apparently exist between translation efficiency and translation accuracy. Similar trade-offs are expected to exist in other kinds of errors, including gene expression noise (Hausser et al., 2019) I studied. Besides uncovering the three main tenets, I also explored diverse biological mechanisms of expression errors in all three main Chapters of my dissertation. Together, these results demonstrate the universality of expression errors, show the importance of considering genome-wide diversity from the perspective of errors(Warnecke and Hurst, 2011), and pave the way for future studies in other kinds of biological errors.

Limitations

My dissertation has several limitations that are worth discussing. First, all my analyses are bioinformatic analyses of publicly available data that are not specifically designed for our hypotheses. This means most of the conclusions are inherently correlational. Although I carefully controlled for the known confounding factors in my analysis, it is not definite proof of the causal relationship because there could always be some unknown confounding factors. To provide definite proof of our conclusions, carefully designed experiments are needed (Hernán and Robins, 2010; Pearl, 2009).

Second, in my analysis, most conclusions tested are about qualitative trends instead of precise quantitative predictions. To obtain and test more precise quantitative predictions, rigorous models of error evolution, in combination with realistic population parameters, have to be developed and tested in the future (Levins, 1966).

Last but not least, my research considers molecular errors as purely deleterious events. However, as mentioned earlier in the dissertation, molecular errors are also the main source for adaptation, despite beneficial errors being relatively rare (Darwin, 1909). In the future, it would be of interest to develop a study framework that could integrate both the deleterious and beneficial aspects of errors. Nevertheless, I believe that the first step to study the beneficial effects of errors should be to document the harmful effects of errors because defining the norm enables us to detect the exceptions (Fisher, 1956).

Future work

Besides documenting the cause and consequences of molecular errors at a genome-wide level, my work also suggests several future research directions. First of all, my dissertation only focuses on two types of expression errors. Because of the increasing availability of omics data (Karczewski and Snyder, 2018), it is possible to study other kinds of molecular errors, such as mutations, RNA modifications, and post-translational modifications in a similar fashion. Extending our analysis to these errors could further confirm or refute the tenets I uncovered and will certainly bring new insights. Second, as mentioned in the "limitations" section, knowing the norm (deleterious errors) allows us to detect the exceptions (adaptive errors), which are the main interest of both medical scientists and evolutionary biologists. Third, as mentioned previously, it is important to develop rigorous and realistic mathematical models in order to understand the evolution of errors. There is already some promising progress along this line, such as the “drift-barrier” model proposed by Michael Lynch (Sung et al., 2012) and the interesting model of global-local error proposed by Joanna Masel and her colleagues (Rajon and Masel, 2011). My research on the mechanisms and consequences of errors could facilitate the further development of these works. Finally, knowing the general properties of bugs sheds light on the design of better programs. My study on expression errors thus has implications for synthetic biology, whose goal is to design biological systems that could outperform natural organisms in specific tasks related to human welfare (Khalil and Collins, 2010). It is my hope that my work would stimulate further research in the above directions.

References

Darwin, C. (1909). The origin of species (PF Collier & son New York).

Fisher, R.A. (1956). Mathematics of a lady tasting tea. The world of mathematics *3*, 1514-1521.

Hausser, J., Mayo, A., Keren, L., and Alon, U. (2019). Central dogma rates and the trade-off between precision and economy in gene expression. Nature communications *10*, 1-15.

Hernán, M.A., and Robins, J.M. (2010). Causal inference (CRC Boca Raton, FL;).

Karczewski, K.J., and Snyder, M.P. (2018). Integrative omics for health and disease. Nat Rev Genet *19*, 299.

Khalil, A.S., and Collins, J.J. (2010). Synthetic biology: applications come of age. Nat Rev Genet *11*, 367-379.

Levins, R. (1966). The strategy of model building in population biology. Am Sci *54*, 421-431.

Milo, R., and Phillips, R. (2015). Cell biology by the numbers (Garland Science).

Pearl, J. (2009). Causality (Cambridge university press).

Rajon, E., and Masel, J. (2011). Evolution of molecular error rates and the consequences for evolvability. Proceedings of the National Academy of Sciences *108*, 1082-1087.

Raser, J.M., and O'shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.

Stewart-Ornstein, J., Weissman, J.S., and El-Samad, H. (2012). Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol Cell *45*, 483-493.

Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences *109*, 18488-18492.

Warnecke, T., and Hurst, L.D. (2011). Error prevention and mitigation as forces in the evolution of genes and genomes. Nat Rev Genet *12*, 875-881.

Appendix

Table A1-1 Significantly enriched GO terms among genes with extreme intrinsic and/or extrinsic expression noise in the non-clonal cells. The three most significant terms are presented if more than three terms are significantly enriched.

GO terms	Corrected *P*-values
High extrinsic noise	
Secreted	5.7×10^{-20}
Extracellular region	2.1×10^{-17}
Disulfide bond	8.4×10^{-17}
Low extrinsic noise	
Splice variant	1.7×10^{-5}
Alternative splicing	7.8×10^{-5}
Regulation of transcription, DNA-templated	1.2×10^{-2}
High intrinsic noise	
Extracellular space	1.3×10^{-13}
Extracellular region	3.0×10^{-13}
Secreted	7.1×10^{-13}
Low intrinsic noise	
Phosphoprotein	2.5×10^{-6}
IPR016024:Armadillo-type fold	8.0×10^{-5}
UbI conjugateon	1.3×10^{-4}
High extrinsic noise and high intrinsic noise	
Extracellular region	8.6×10^{-25}
Extracellular space	1.5×10^{-24}
Secreted	4.3×10^{-23}
High extrinsic noise and low intrinsic noise	
Cell division	3.6×10^{-7}
Mitosis	4.2×10^{-7}
Mitotic nuclear division	3.9×10^{-6}
Low extrinsic noise and low intrinsic noise	
poly (A) RNA binding	2.5×10^{-5}
Phosphoprotein	4.8×10^{-4}
Nucleus	6.9×10^{-4}

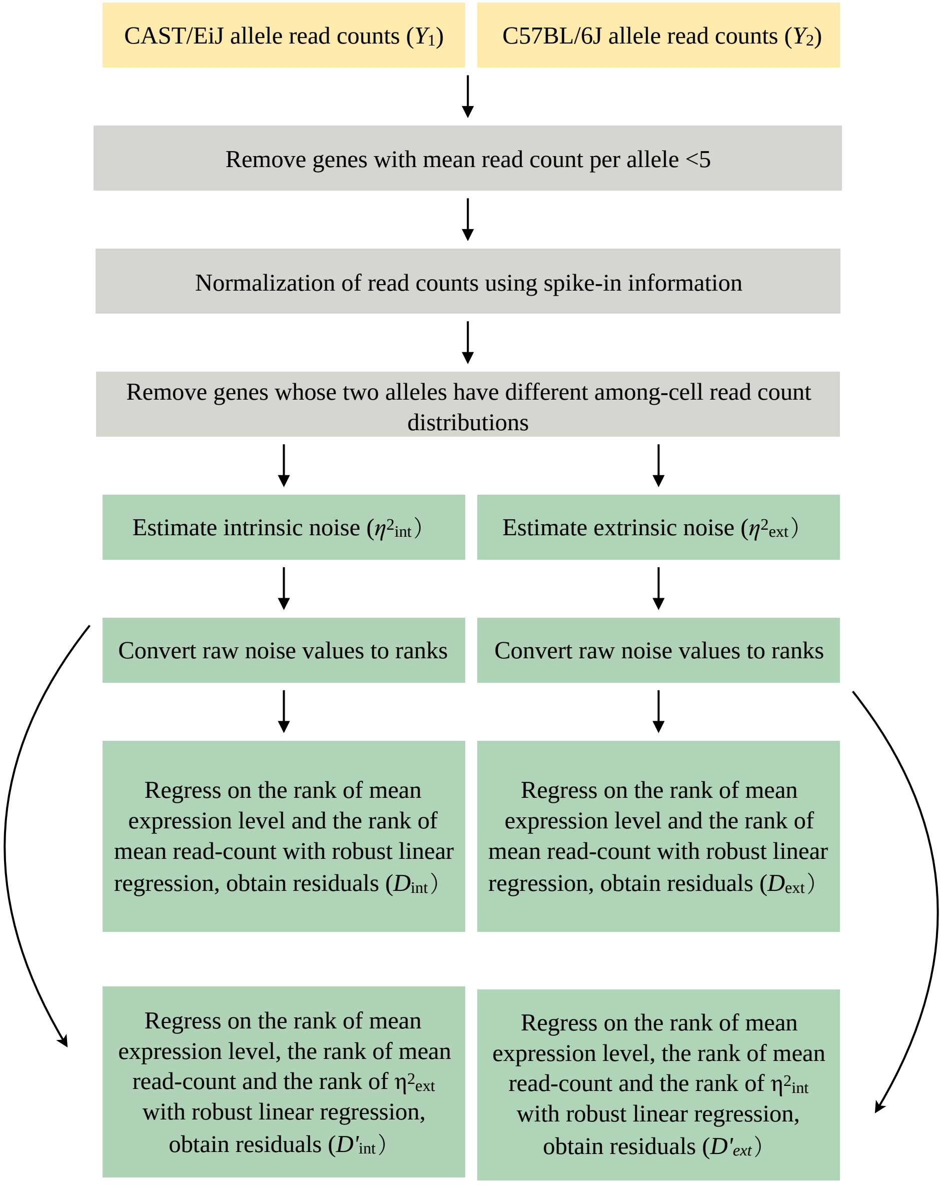
CAST/EiJ allele read counts (Y_1)
C57BL/6J allele read counts (Y_2)
Remove genes with mean read count per allele <5
Normalization of read counts using spike-in information
Remove genes whose two alleles have different among-cell read count distributions
Estimate intrinsic noise (η^2_{int})
Estimate extrinsic noise (η^2_{ext})
Convert raw noise values to ranks
Convert raw noise values to ranks
Regress on the rank of mean expression level and the rank of mean read-count with robust linear regression, obtain residuals (D_{int})
Regress on the rank of mean expression level and the rank of mean read-count with robust linear regression, obtain residuals (D_{ext})
Regress on the rank of mean expression level, the rank of mean read-count and the rank of η^2_{ext} with robust linear regression, obtain residuals (D'_{int})
Regress on the rank of mean expression level, the rank of mean read-count and the rank of η^2_{int} with robust linear regression, obtain residuals (D'_{ext})

Fig. A1-1 Procedure for clone 7 scRNA-seq data processing and noise estimation.

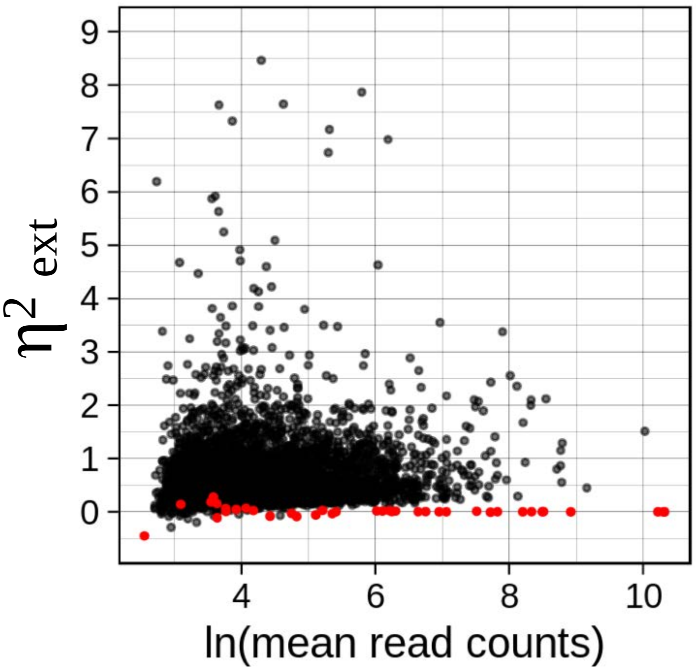

Fig. A1-2 The extrinsic noise of genes (black dots) are above technical extrinsic noises (red dots) estimated from spike-in molecules in clone 7 cells.

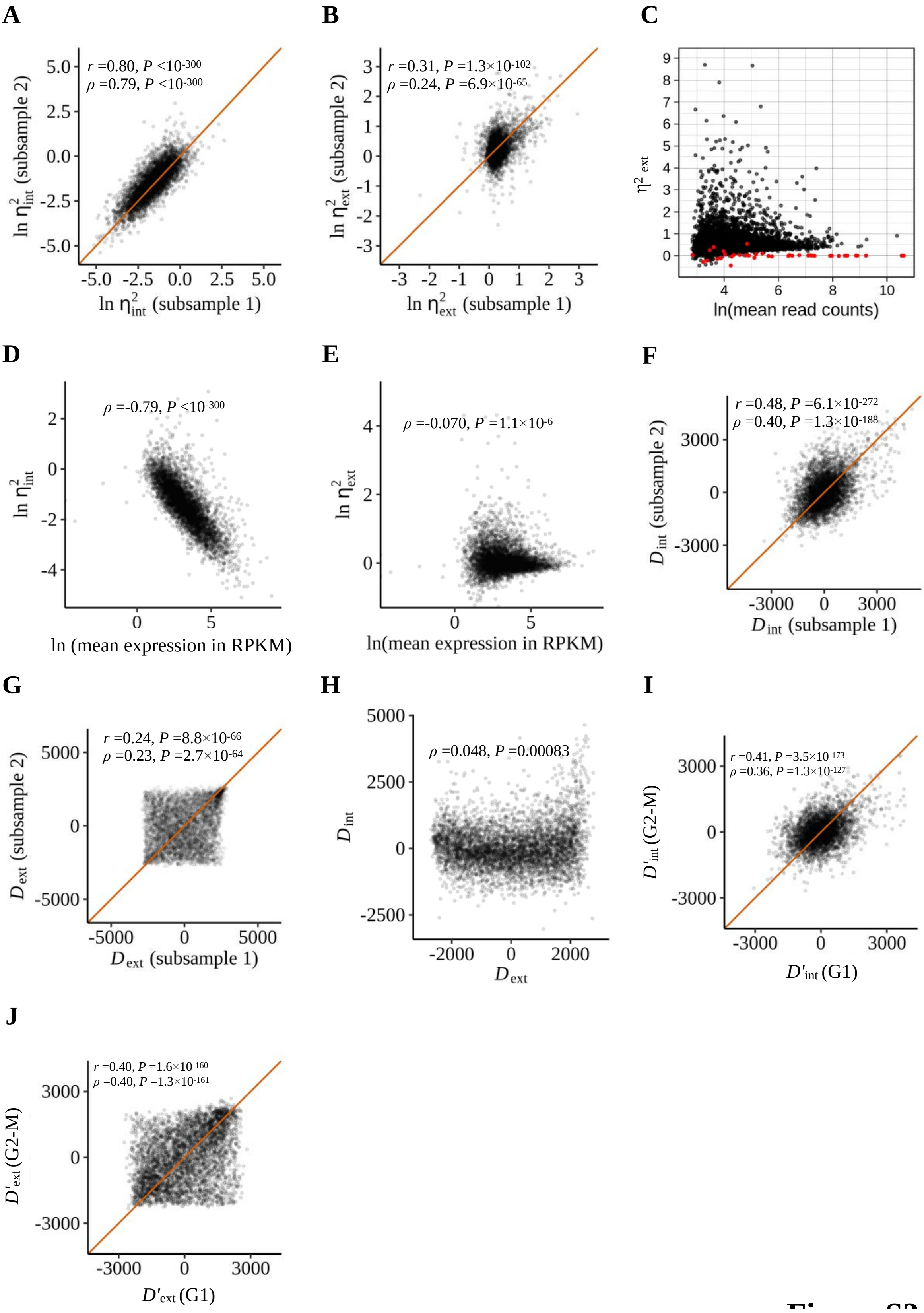
A
r =0.80, P <10-300
ρ =0.79, P <10-300
ln η2int (subsample 2)
ln η2int (subsample 1)
B
r =0.31, P =1.3×10-102
ρ =0.24, P =6.9×10-65
ln η2ext (subsample 2)
ln η2ext (subsample 1)
C
η2 ext
ln(mean read counts)
D
ρ =-0.79, P <10-300
ln η2int
ln (mean expression in RPKM)
E
ρ =-0.070, P =1.1×10-6
ln η2ext
ln(mean expression in RPKM)
F
r =0.48, P =6.1×10-272
ρ =0.40, P =1.3×10-188
Dint (subsample 2)
Dint (subsample 1)
G
r =0.24, P =8.8×10-66
ρ =0.23, P =2.7×10-64
Dext (subsample 2)
Dext (subsample 1)
H
ρ =0.048, P =0.00083
Dint
Dext
I
r =0.41, P =3.5×10-173
ρ =0.36, P =1.3×10-127
D′int (G2-M)
D′int (G1)
J
r =0.40, P =1.6×10-160
ρ =0.40, P =1.3×10-161
D′ext (G2-M)
D′ext (G1)

Figure S3

Fig. A1-3 Decomposition of gene expression noise into intrinsic and extrinsic noises in non-clonal cells. **(A)** Intrinsic noises (η^2_{int}) estimated from two sub-samples of the non-clonal cells are highly correlated with each other. Ln-transformed η^2_{int} is shown. Each dot is a gene. The orange line shows the diagonal. **(B)** Extrinsic noises (η^2_{ext}) estimated from two sub-samples of the non-clonal cells are moderately correlated with each other. Ln-transformed η^2_{ext} is shown. Each dot is a gene. The orange line shows the diagonal. **(C)** The extrinsic noise of genes (black dots) are above technical extrinsic noises (red dots) estimated from spike-in molecules.**(D)** The intrinsic expression noise of a gene is strongly negatively correlated with the mean expression level of the gene. Expression level is measured by Reads Per Kilobase of transcript per Million mapped reads (RPKM). **(E)** The extrinsic expression noise of a gene is weakly negatively correlated with the mean expression level of the gene. Because the extrinsic noise could be negative (see Materials and Methods), we added a small value (0.1 - the minimum of computed extrinsic noise) to all η^2_{ext} values before taking the natural log. **(F)** Intrinsic noise estimates adjusted for mean expression level and technical noise (D_{int}) are significantly correlated between two sub-samples of non-clonal cells. The orange line shows the diagonal. **(G)** Extrinsic noise estimates adjusted for mean expression level and technical noise (D_{ext}) are significantly correlated between two sub-samples of non-clonal cells. The orange line shows the diagonal. **(H)** D_{int} and D_{ext} are positively correlated. The blue line displays the linear regression of D_{int} on D_{ext}.

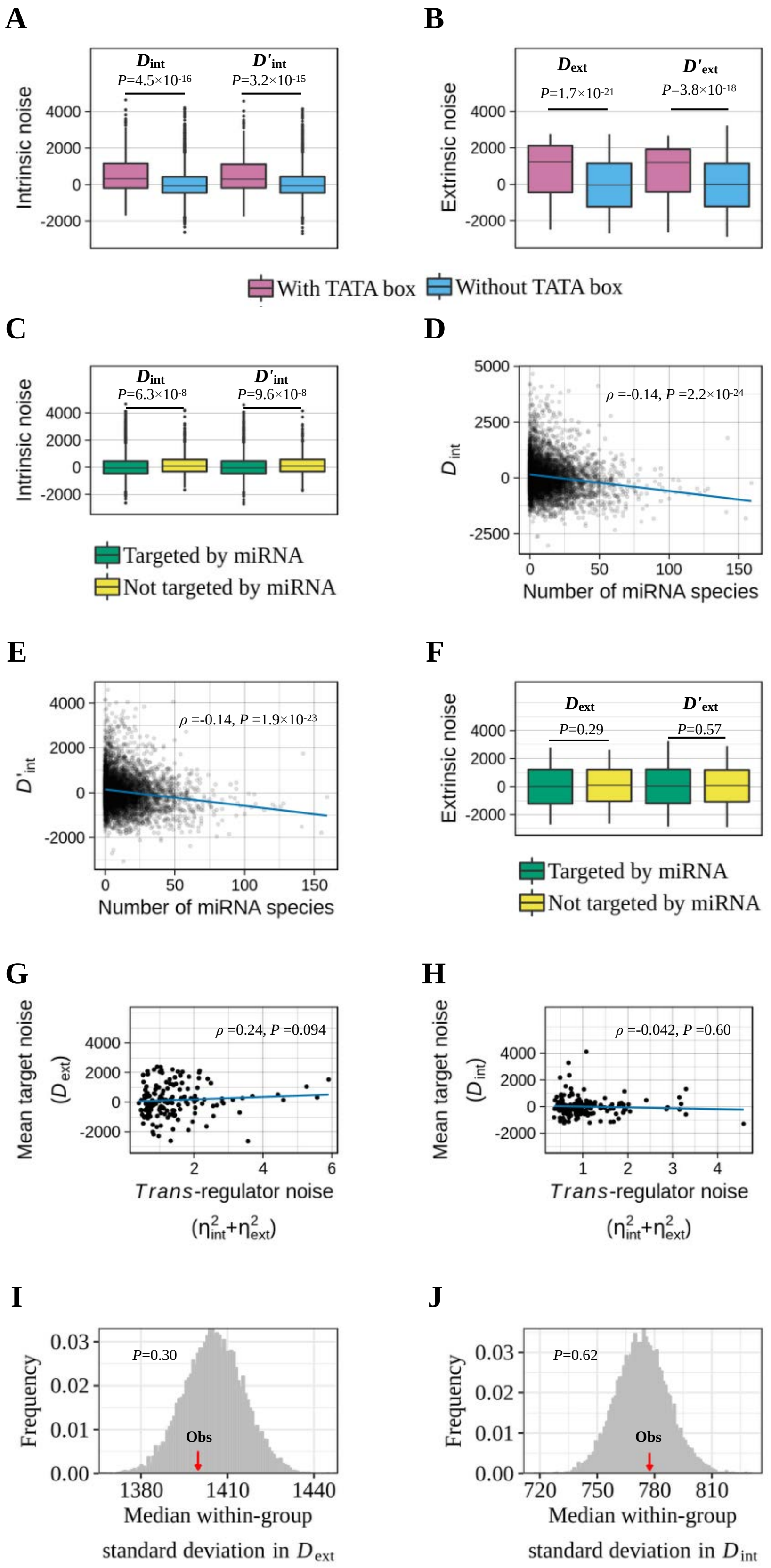
A
Intrinsic noise
Dint P=4.5×10-16
D'int P=3.2×10-15
B
Extrinsic noise
Dext P=1.7×10-21
D'ext P=3.8×10-18
With TATA box
Without TATA box
C
Intrinsic noise
Dint P=6.3×10-8
D'int P=9.6×10-8
Targeted by miRNA
Not targeted by miRNA
D
Dint
ρ =-0.14, P =2.2×10-24
Number of miRNA species
E
D'int
ρ =-0.14, P =1.9×10-23
Number of miRNA species
F
Extrinsic noise
Dext P=0.29
D'ext P=0.57
Targeted by miRNA
Not targeted by miRNA
G
Mean target noise (Dext)
ρ =0.24, P =0.094
Trans-regulator noise (η²int+η²ext)
H
Mean target noise (Dint)
ρ =-0.042, P =0.60
Trans-regulator noise (η²int+η²ext)
I
Frequency
P=0.30
Obs
Median within-group standard deviation in Dext
J
Frequency
P=0.62
Obs
Median within-group standard deviation in Dint

Fig. A1-4 Factors influencing intrinsic and/or extrinsic gene expression noise in non-clonal cells. **(A)** Genes with a TATA-box in the promoter (pink) have significantly higher intrinsic noise (D_{int}) than genes without a TATA-box (blue). The same is true when intrinsic noise is measured by D'_{int}, which is uncorrelated with extrinsic noise. The lower and upper edges of a box represent the first (qu_1) and third (qu3) quartiles, respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Genes with a TATA-box in the promoter (pink) have significantly higher extrinsic noise (D_{ext}) than genes without a TATA-box (blue). The same is true when extrinsic noise is measured by D'_{ext}, which is uncorrelated with intrinsic noise.**(C)** Genes targeted by miRNA (green) have significantly lower intrinsic noise (D_{int} and D'_{int}) than genes not targeted by miRNA (yellow). **(D)** Genes targeted by more miRNA species have lower D_{int}. The blue line displays the linear regression of D_{int} of a target gene on the number of miRNA species targeting it. **(E)** Genes targeted by more miRNA species have lower D'_{int}. The blue line displays the linear regression of D'_{int} of a target gene on the number of miRNA species targeting it. **(F)** Genes targeted by miRNA (green) have similar levels of extrinsic noise (D_{ext} and D'_{ext}) as genes not targeted by miRNA (yellow). **(G)** The mean extrinsic noise (D_{ext}) of genes targeted by the same *trans*-regulator is not significantly correlated with the total noise $(\eta_{\text{int}}^2 + \eta_{\text{ext}}^2)$ of the *trans*-regulators.**(H)** The mean intrinsic noise (D_{int}) of genes targeted by the same *trans*-regulator is not significantly correlated with the total noise $(\eta_{\text{int}}^2 + \eta_{\text{ext}}^2)$ of the *trans*-regulator. **(I)** The observed median standard deviation of D_{ext} among genes regulated by the same *trans*-regulator (red arrow) is not significantly different from the random expectation (histograms). **(J)** The observed median standard deviation of D_{int} among genes regulated by the same *trans*-regulator is not significantly different from the random expectation (histograms).

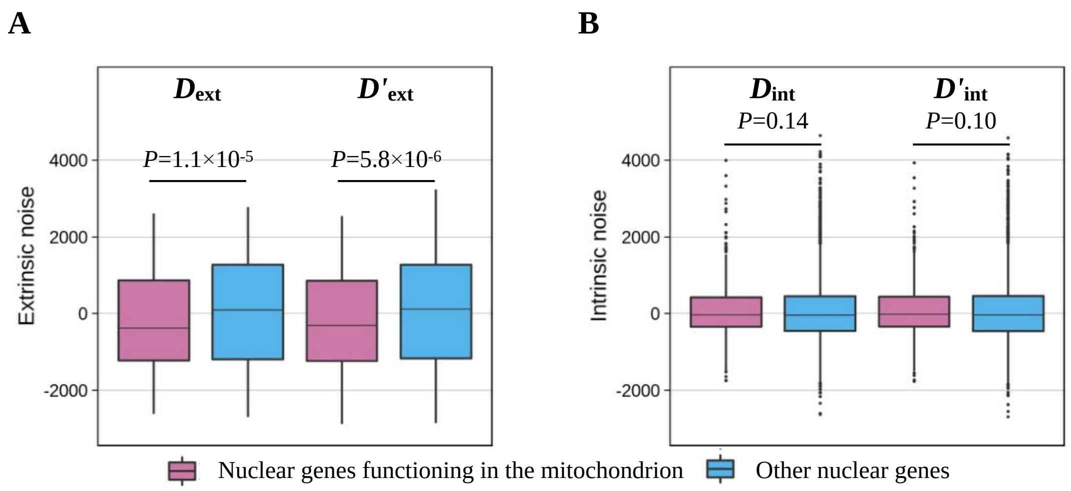

Fig. A1-5 Nuclear genes functioning in the mitochondrion have lower extrinsic noise but not lower intrinsic noise when compared with other genes in non-clonal cells. **(A)** Nuclear genes functioning in the mitochondrion (pink) have significantly lower extrinsic noise (D_{ext} and D'_{ext}) than other genes (blue). The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Nuclear genes functioning in the mitochondrion (pink) do not have significantly lower intrinsic noise D_{int} and even have significantly higher D'_{int} than other genes (blue).

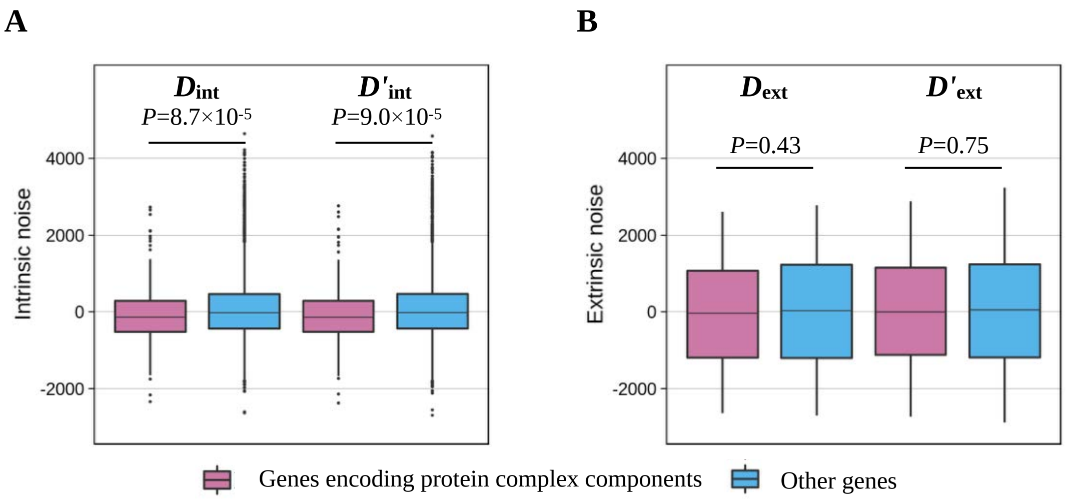

Fig. A1-6 Genes encoding protein complex components have lower intrinsic noise but not lower extrinsic noise than other genes in non-clonal cells. **(A)** Genes encoding protein complex components (pink) have significantly lower intrinsic noise (D_{int} and D'_{int}) than other genes (blue). The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Genes encoding protein complex components (pink) do not have significantly lower D_{ext} or D'_{ext} than other genes (blue).

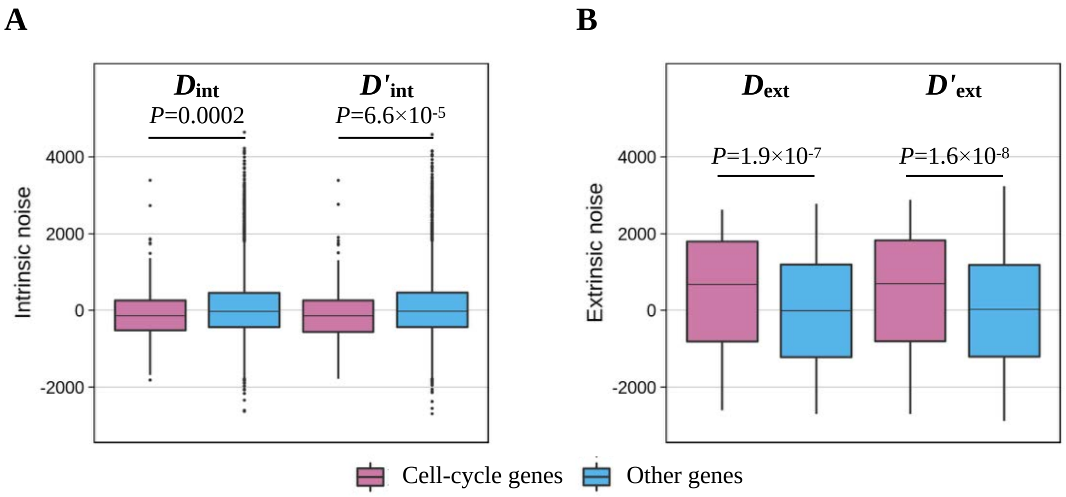

Fig. A1-7 Cell cycle genes have lower intrinsic noise but higher extrinsic noise than other genes in non-clonal cells. **(A)** Cell cycle genes (pink) have significantly lower intrinsic noise (D_{int} and D'_{int}) when compared with other genes (blue). The lower and upper edges of a box represent the first (qu_1) and third quartiles (qu_3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu_3-qu_1), and the dots represent values outside the inner fences (outliers). **(B)** Cell cycle genes (pink) have significantly higher extrinsic noise (D_{ext} and D'_{ext}) when compared with other genes.

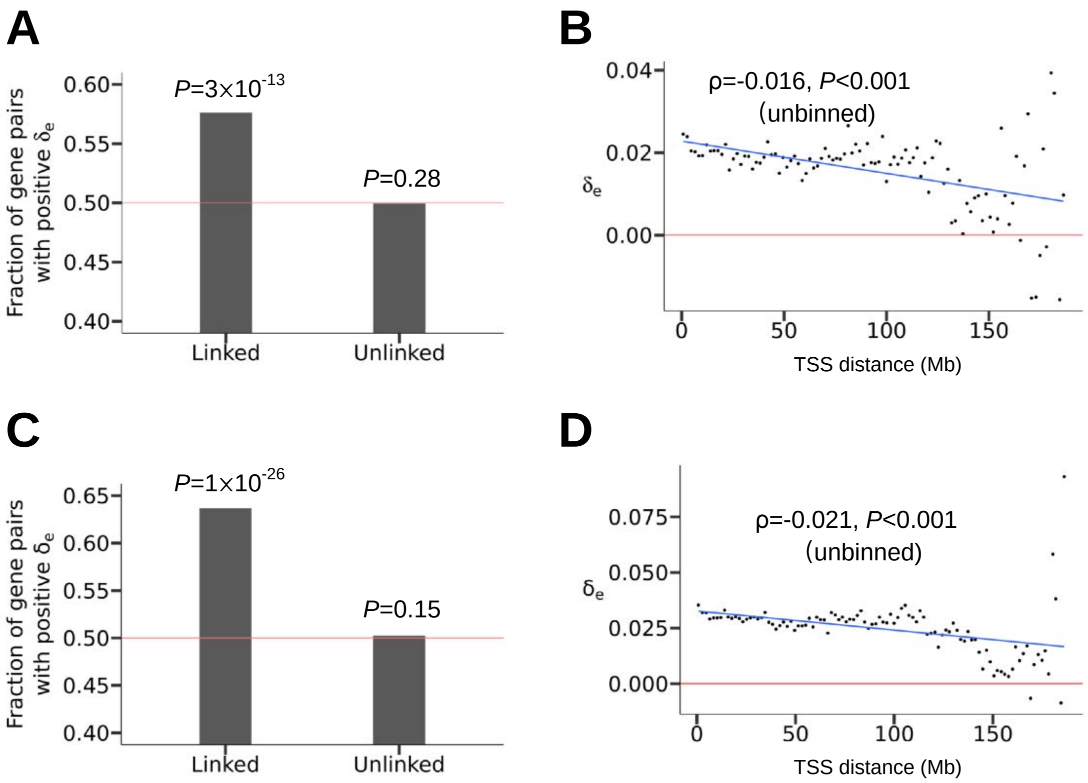

Fig. A2-1 The linkage effect on expression co-fluctuation in clone 6 cells and non-clonal cells. **(A)** Fraction of gene pairs with positive δ_e in clone 6. The red line represents the null expectation under no linkage effect. P-values from binomial tests on independent gene pairs are presented. **(B)** In clone 6, median δ_e in a bin decreases as the median genomic distance between linked genes in the bin rises. All bins have the same distance interval. TSS, transcription start site. The red line shows $\delta_e = 0$. The blue line shows the linear regression of binned data. Spearman's ρ from unbinned data and associated P-value determined by a shuffling test are presented. **(C)** Fraction of gene pairs with positive δ_e in non-clonal mouse fibroblast cells. The red line represents the null expectation under no linkage effect. P-values from binomial tests on independent gene pairs are presented. **(D)** In non-clonal cells, median δ_e in a bin decreases as the median genomic distance between linked genes in the bin rises. All bins have the same distance interval. TSS, transcription start site. The red line shows $\delta_e = 0$. The blue line shows the linear regression of binned data. Spearman's ρ from unbinned data and associated P-value determined by a shuffling test are presented.

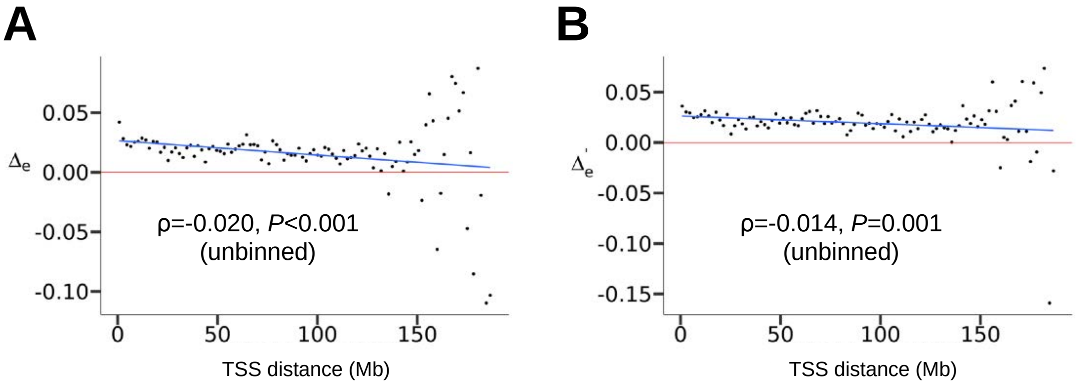

Fig. A2-2 The linkage effect on expression co-fluctuation in clone 7 cells analyzed using total reads of two alleles per locus. **(A)** Median Δ_e in a bin decreases with the median genomic distance between linked genes in the bin. Δ_e for a linked gene pair is the correlation in RNA-seq read number between the two genes minus the median correlation for pairs of unlinked genes. All bins have the same distance interval. TSS, transcription start site. The red line shows $\Delta_e = 0$. The blue line shows the linear regression of binned data. Spearman's ρ of unbinned data and associated P-value determined by a shuffling test ae presented. **(B)** Median Δ_e' in a bin decreases with the corresponding median genomic distance between linked genes in the bin. Δ_e' for a linked gene pair is the correlation in expression level measured by RPKM (Reads Per Kilobase per Million mapped reads) between the two genes minus the corresponding median correlation for pairs of unlinked genes. The blue line shows the linear regression of binned data. Spearman's ρ from unbinned data and associated P-value determined by a shuffling test are presented.

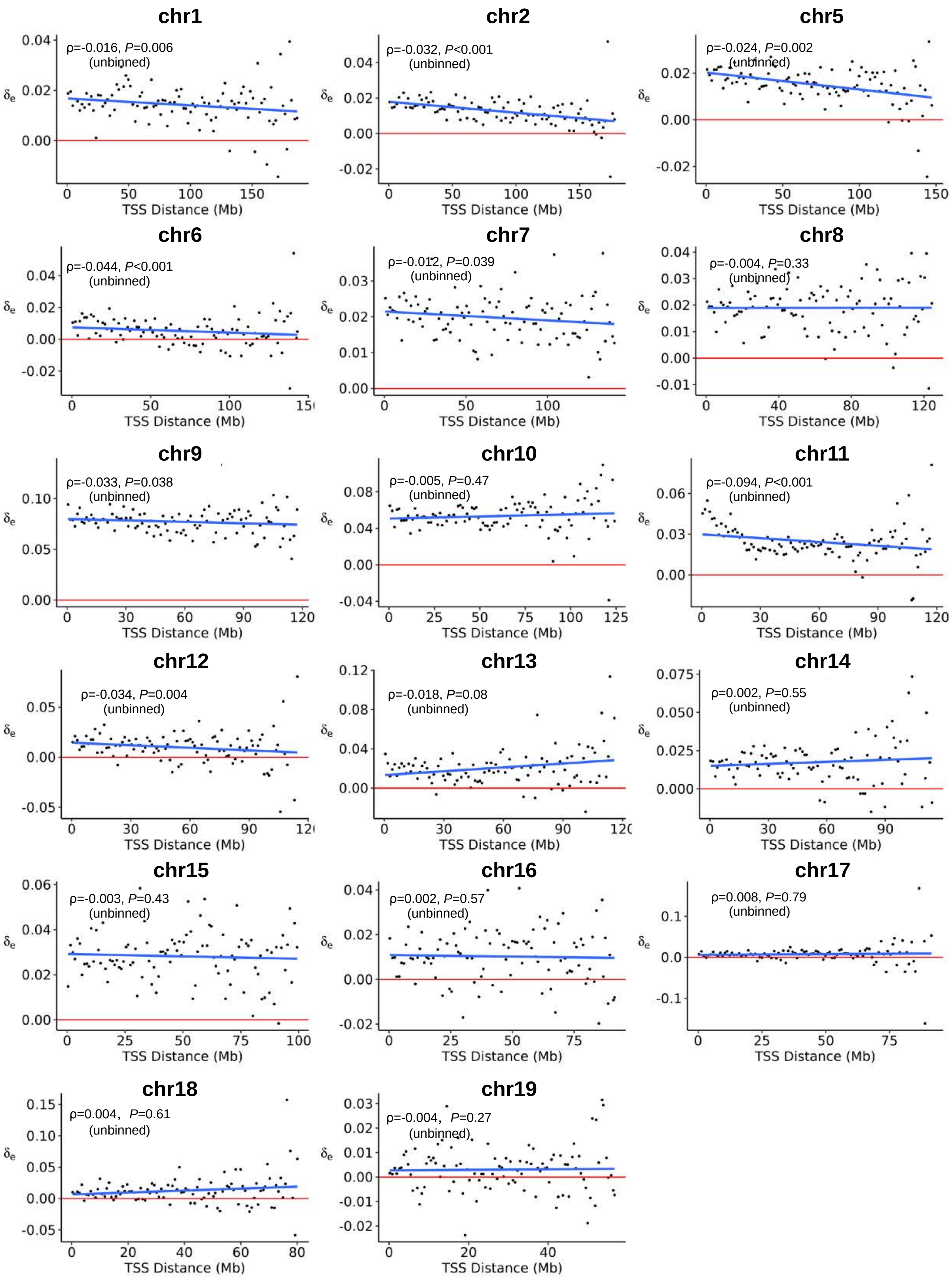
chr1
ρ=-0.016, P=0.006 (unbinned)
chr2
ρ=-0.032, P<0.001 (unbinned)
chr5
ρ=-0.024, P=0.002 (unbinned)
chr6
ρ=-0.044, P<0.001 (unbinned)
chr7
ρ=-0.012, P=0.039 (unbinned)
chr8
ρ=-0.004, P=0.33 (unbinned)
chr9
ρ=-0.033, P=0.038 (unbinned)
chr10
ρ=-0.005, P=0.47 (unbinned)
chr11
ρ=-0.094, P<0.001 (unbinned)
chr12
ρ=-0.034, P=0.004 (unbinned)
chr13
ρ=-0.018, P=0.08 (unbinned)
chr14
ρ=0.002, P=0.55 (unbinned)
chr15
ρ=-0.003, P=0.43 (unbinned)
chr16
ρ=0.002, P=0.57 (unbinned)
chr17
ρ=0.008, P=0.79 (unbinned)
chr18
ρ=0.004, P=0.61 (unbinned)
chr19
ρ=-0.004, P=0.27 (unbinned)
δ_e
TSS Distance (Mb)

Fig. A2-3 δ_e decreases with distance between genes on each mouse chromosome. Blue lines show linear regressions for binned data. All bins have the same distance intervals, while different chromosomes contain different numbers of bins depending on the chromosome length. Spearman's correlations from unbinned data and associated nominal P-values determined by shuffling tests are presented. Upon multiple testing correction, the correlations remain significant for chromosomes 1, 2, 5, 6, 11, and 12.

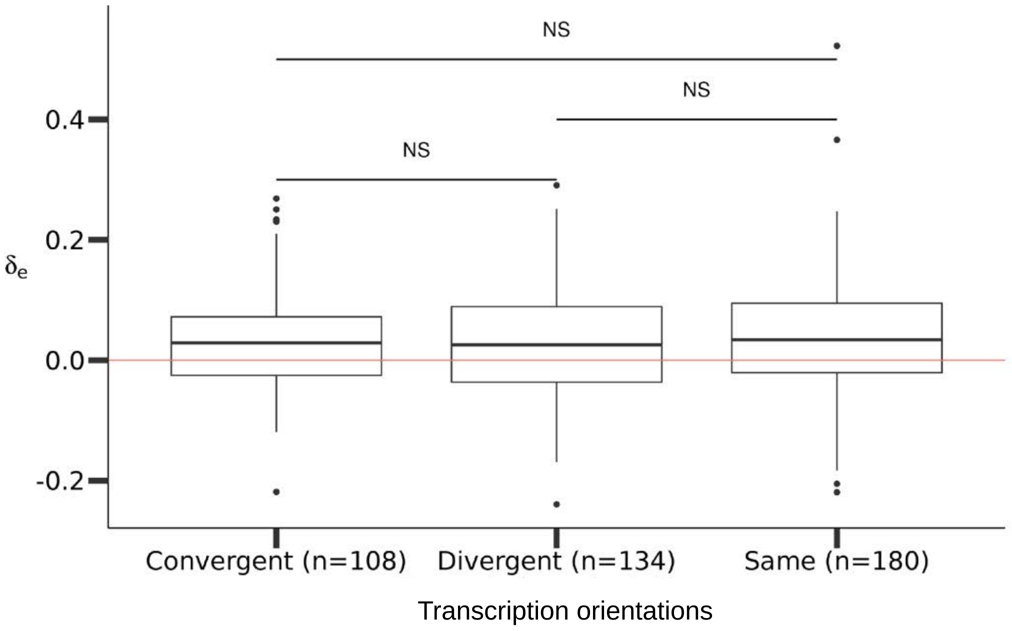

Fig. A2-4 δ_e for pairs of neighboring genes with different orientation types. The lower and upper edges of a box represent the first (qu1) and third quartiles (qu3), respectively, the horizontal line inside the box indicates the median (md), the whiskers extend to the most extreme values inside inner fences, md±1.5(qu3-qu1), and the dots represent values outside the inner fences (outliers). The nearest pairs were identified using the coordinates downloaded from Ensembl. After requiring a minimal read number of 10 for each allele, we separate neighboring gene pairs into three categories according to the orientations of their transcription directions. NS, P > 0.05, Wilcoxon rank-sum test.

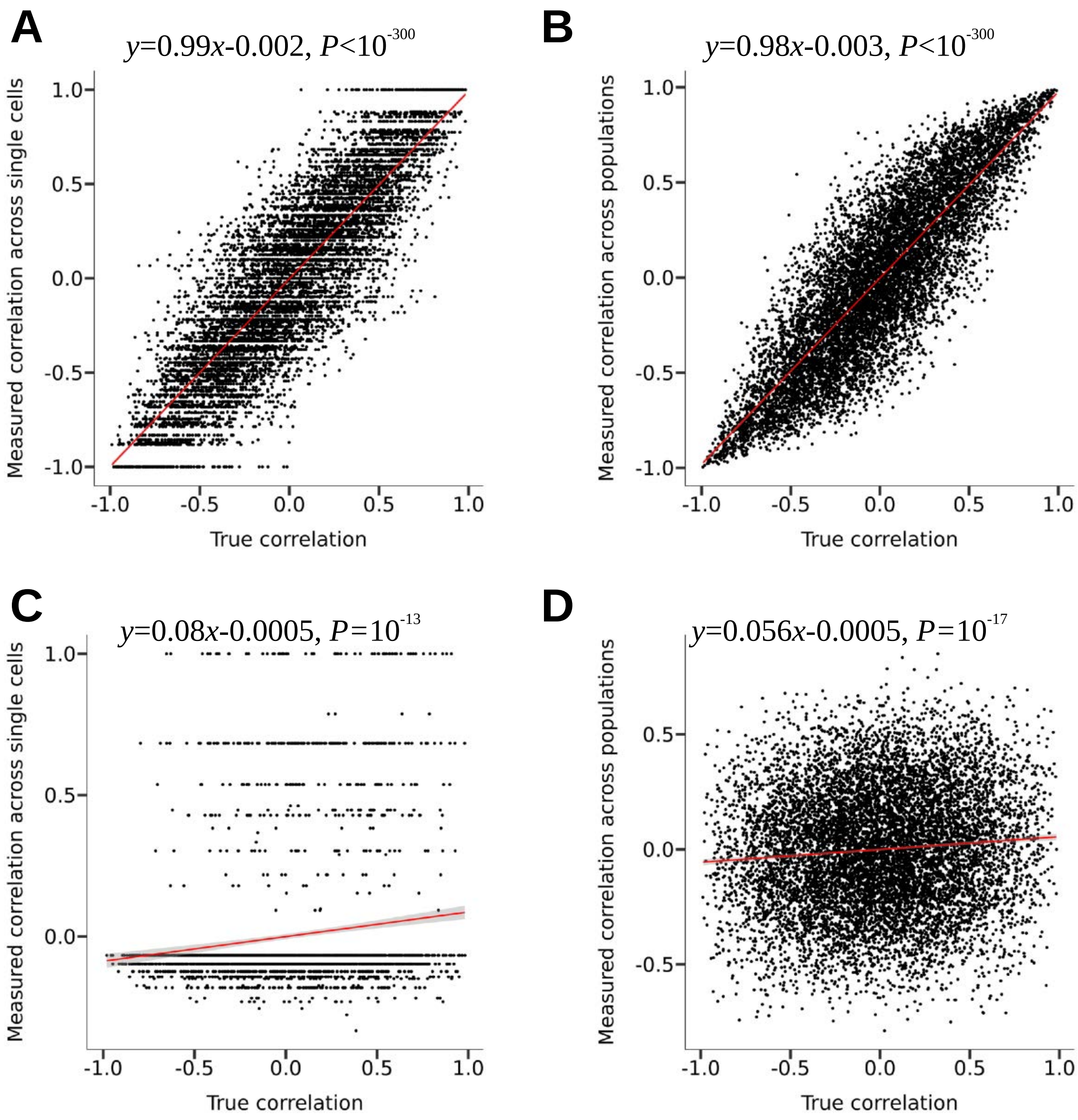

Fig. A2-5 Demonstration of chromatin co-accessibility between two ATAC peaks quantified using single-cells vs. using cell populations via simulation **(A)** The correlations quantified using single-cell-based measurements are close to their corresponding true correlations when the capturing efficiency is 100%. **(B)** The correlations quantified using cell-population-based measurements are close to the true correlations when the capturing efficiency is 100%. **(C)** The correlations quantified using single-cell-based measurements tend to be weaker than their corresponding true correlations when the capturing efficiency is 10%. **(D)** The correlations quantified using cell population-based measurements tend to be weaker than the true correlations when the capturing efficiency is 10%.

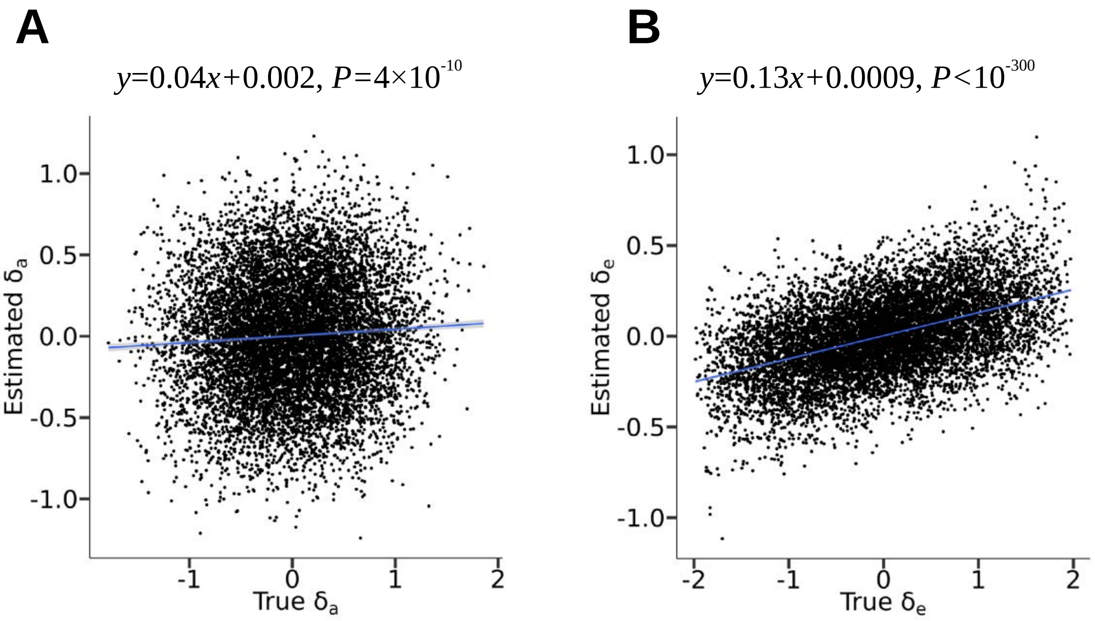

Fig. A2-6 Simulation shows low capturing efficiency will lead to underestimation of δ_e and δ_a. **(A)** The magnitude of δ_e estimated from allelic specific single cell RNA-seq is much smaller than the true δ_e. **(B)** The magnitude of δ_a estimated from allelic specific ATAC-seq is much smaller than the true δ_a.

Table A3-1 Pearson (r) and Spearman (ρ) correlations Between *OR* and *RSCU* in eukaryotic model organisms.

Species	***r***	***P***	***ρ***	***P***
Fly	0.84	<0.001	0.83	<0.001
Human	0.80	<0.001	0.74	<0.001
Mouse	0.74	<0.001	0.68	<0.001
Worm	0.42	<0.001	0.39	0.002
Yeast	0.35	0.01	0.39	0.003

Table A3-2 Pearson (*r*) and Spearman (*ρ*) correlations between *OR* and $RR_{\mathrm{c/nc}}$ in eukaryotic model organisms.

Species	***r***	***P***	***ρ***	***P***
Fly	0.21	0.109	0.25	0.046
Human	0.39	0.007	0.29	0.055
Mouse	0.29	0.081	0.41	0.018
Worm	0.54	<0.001	0.57	<0.001
Yeast	0.42	0.045	0.37	0.053

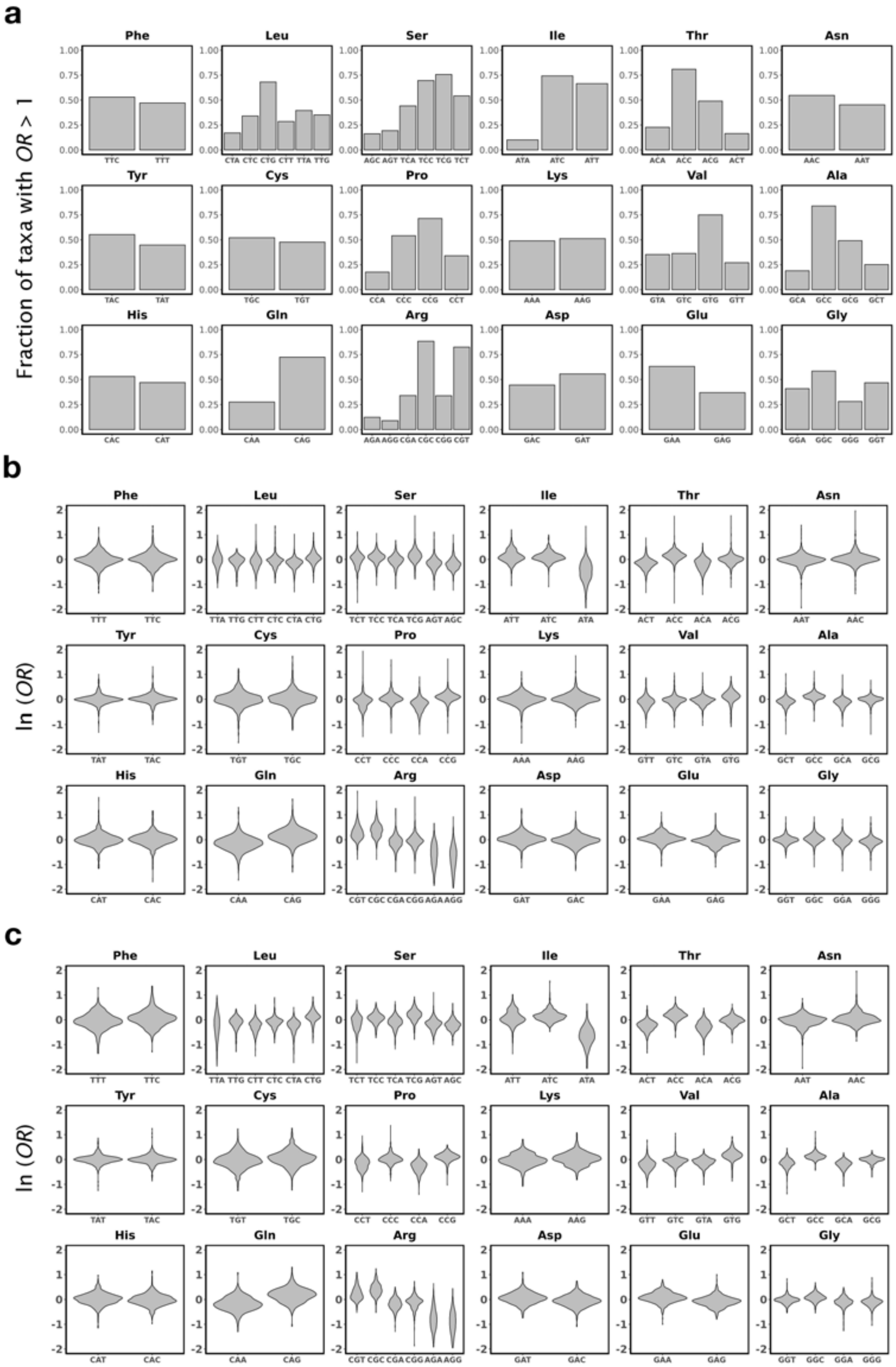

Fig. A3-1 Observed patterns of variation of relative translational accuracies of synonymous codons across bacterial taxa are robust**. a**, Fraction of 1197 taxa with *OR* > 1 for each codon. **b**, Violin plots showing frequency distributions of ln(*OR*) of individual codons across a subset of taxa in which the number of occurrences of each synonymous codon considered in *OR* estimation is at least 1000. **c**, Violin plots showing frequency distributions of ln(*OR*) of individual codons across a subset of taxa with strong signals of selection for translational accuracy (i.e., Pearson's correlation between *RSCU* and *OR* exceeds 0.5).

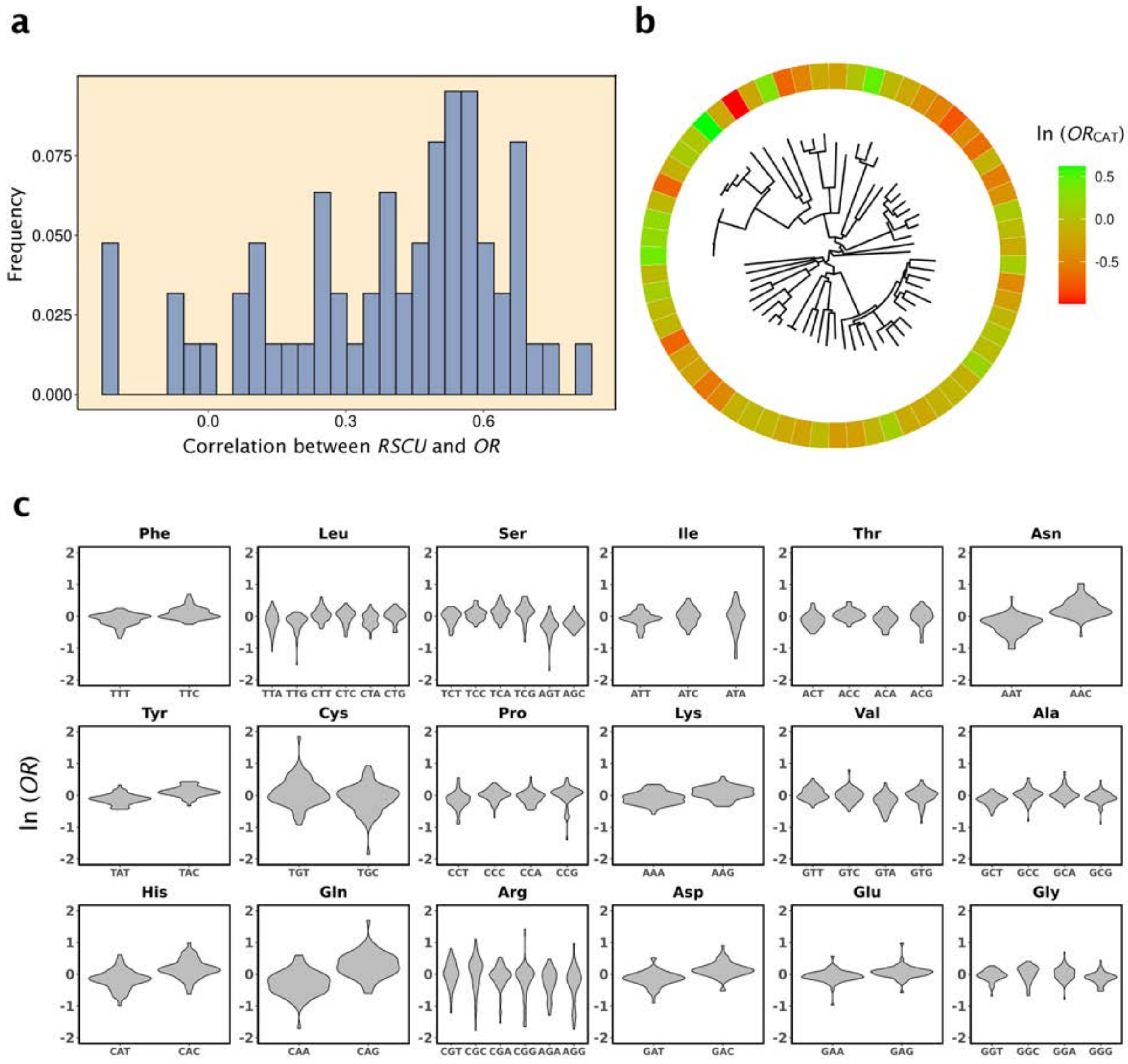

Fig. A3-2 Variation of relative translational accuracies of synonymous codons among Archaea taxa. **a**, Frequency distribution of Pearson's correlation between *RSCU* and *OR* in 63 taxa. Ninety percent of taxa show positive correlations. **b**, ln(*OR*) of codon CAT for each of the taxa arranged according to their phylogeny shown in the middle. **c**, Violin plots showing frequency distributions of ln(*OR*) of individual codons across taxa. ln(*OR*) appears less variable here than in Bacteria because of the much fewer Archaea taxa examined.

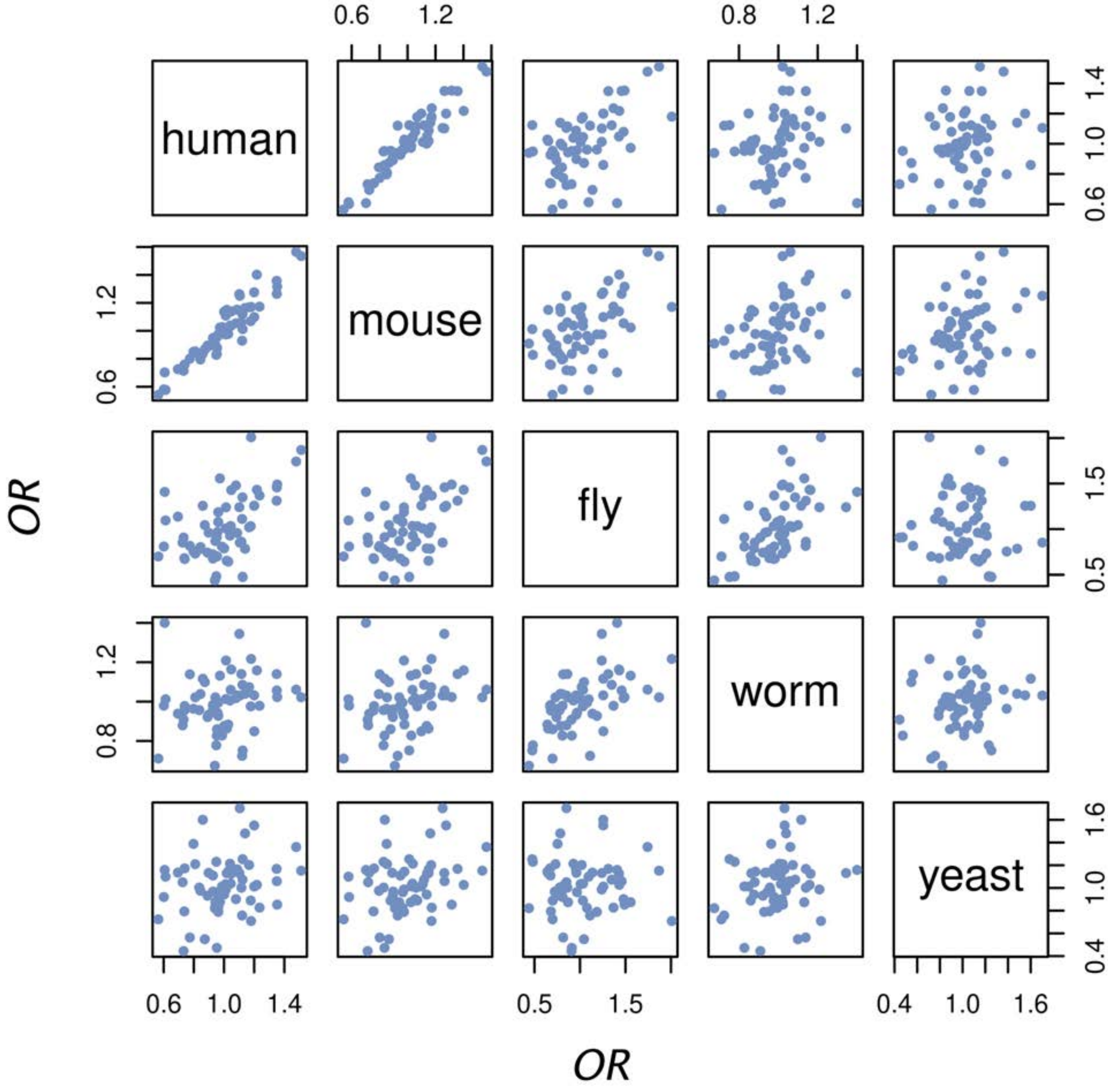

Fig. A3-3. Correlation in odds ratio (*OR*) across codons between eukaryotic model organisms. Each dot is a codon.

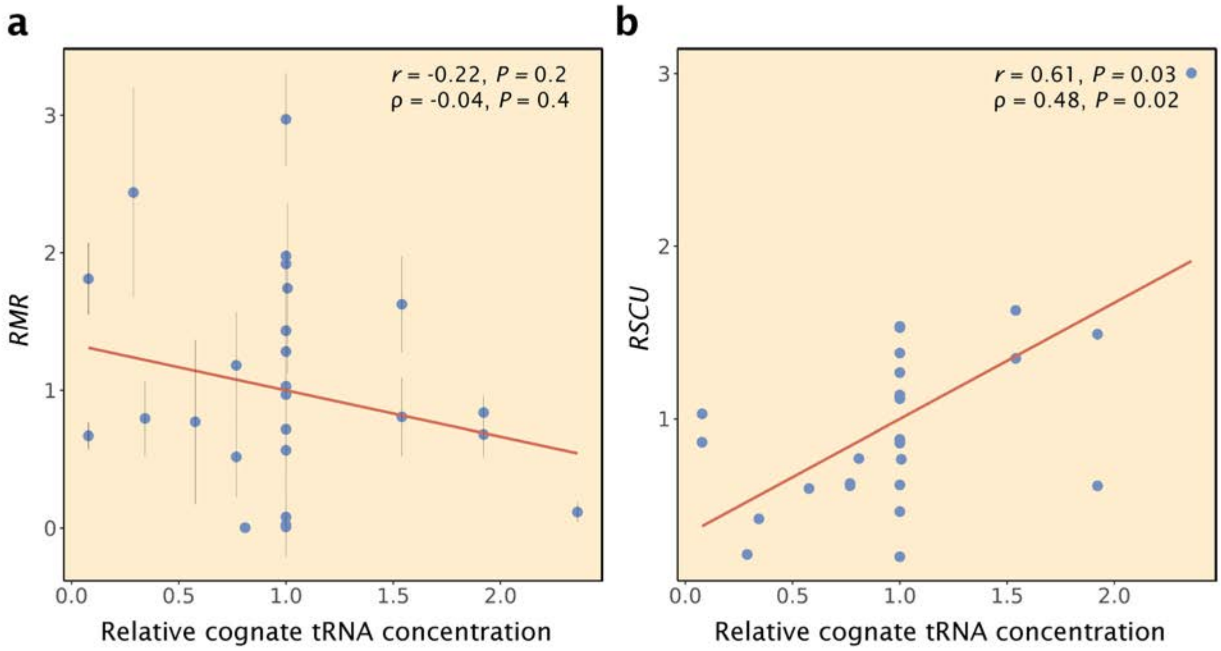

Fig. A3-4. Relationship between the cognate tRNA concentration and *RMR* or *RSCU* in *E. coli*. **a**, *RMR* of a codon is not significantly correlated with its relative cognate tRNA concentration, which is its cognate tRNA concentration divided by the mean cognate tRNA concentration of all codons coding for the same amino acid. **b**, *RSCU* of a codon is significantly positively correlated with its relative cognate tRNA concentration. *P*-values are based on permutation tests. Only the 27 codons with *RMR* estimates are analyzed in each panel to allow a direct comparison.