Feb. 6, 2019

Chromosome-wide co-fluctuation of stochastic gene expression in mammalian cells

Mengyi Sun and Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

*Corresponding author
Jianzhi Zhang
Department of Ecology and Evolutionary Biology
University of Michigan
4018 Biological Science Building
1105 North University Avenue
Ann Arbor, MI 48109
Phone: 734-763-0527
Fax: 734-763-0544
Email: jianzhi@umich.edu

Running title: Expression co-fluctuation of linked genes

Keywords: chromatin accessibility, chromosome conformation, evolution, gene expression noise, protein complex

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.

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 [1-3]. 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 [4-8]. 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 [9, 10], whereas multicellular organisms can make use of expression noise to initiate developmental processes [11-13].

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* [14] and unicellular eukaryote *Saccharomyces cerevisiae* [15]. 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[16], and the expression co-fluctuation of these genes is facilitated by their sharing of transcriptional regulators [17].

Here we explore yet another mechanism for expression co-fluctuation. We hypothesize that, due to the sharing of chromatin dynamics [18], a key contributor to gene expression noise [18-20], 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. 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 a pioneering study 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 [21]. 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 [22]. 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.

RESULTS

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. 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, $\delta_e = [r_e(A_1, B_1) + r_e(A_2, B_2) - r_e(A_1, B_2) - r_e(A_2, B_1)]/2 > 0$. 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) [22]. Single-cell RNA-seq profiles the transcriptomes of individual cells, allowing quantifying stochastic gene expression variations among isogenic cells in the same environment [23-25]. 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. 2B), significantly exceeding the null expectation of 0.5 ($P < 2.4\times10^{-16}$, binomial test). Because a gene can appear in multiple gene pairs, in the above binomial test, we considered 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. 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 (nominal $P < 0.05$; Fig. 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. 2B). The fraction of gene pairs with $\delta_e > 0$ appears to vary among chromosomes (Fig. 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 28 cells with RNA-seq data. Similar results were obtained (Fig. S1A and S1B). 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. S1C and

Fig. S1D). 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. 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. 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. 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. S1B).

Notably, a previous experiment in mammalian cells [21] 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 [21] are much more precise than those

estimated using allele-specific single-cell RNA-seq [26] 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. 2E), median δ_e for all chromosomally linked gene pairs (Fig. 2F), and median δ_e for the left-most bin (Fig. 2F) all increase.

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. S2), 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 [27]. This impact, however, is unobserved in our study (Fig. S4), 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.

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 [28], whereas the diameter of the nucleus is ~8 μm [28]. 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. 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 [29]. The smaller the 3D distance between two genomic regions, the higher the interaction frequency between them[30]. 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 [31] of mouse neural progenitor cells (NPC). For any two linked loci *A* and *B* as depicted in the left diagram of Fig. 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. 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. 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. 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 [32], 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) [33] 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 [34]. 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. 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. 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. 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.

Because we hypothesize that the linkage effect on expression co-fluctuation is via 3D chromatin proximity that leads to chromatin co-accessibility (Fig. 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. 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 [35].

To test the hypothesis that chromatin co-accessibility leads to expression co-fluctuation (even for unlinked alleles) (Fig. 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 [36], chromatin co-accessibility, which is affected by 3D chromatin proximity (Fig. 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*-$r_e = r_e(A_1, B_2)/2 + r_e(A_2, B_1)/2$ 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. 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).

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. 3A). However, 3D proximity can lead to promoter co-accessibility by several means, which have been broadly summarized into three categories of mechanisms [28]: 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 [28]. 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 [37], evident in Fig. 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 [28], and any contact-based mechanism is unlikely to be long-range (e.g., topologically associating domains) [36]. 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 [38] and hence no linkage effect is expected (top cell in Fig. 3H). 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 [38] and hence cannot be chromosome-wide (bottom cell in Fig. 3H). 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. 3H). The typical transcriptional burst interval is 18-50 minutes in mammalian cells [39, 40]. 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 [32]. 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 [32, 41], 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 [41, 42]. 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 [43], have a diffusion coefficient of ~0.01$\mu m^2/s$ [44]. 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.

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 [45]. 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 [46]. We found that, as the correlation between $[A]_t$ and $[B]_t$ increases, mean [AB] of the 10,000 cells rises (Fig. 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 [47, 48], suggesting a demand for expression co-fluctuation of members of the same protein complex.

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 [49]. Because genes may be linked due to their origins from tandem duplication, 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. 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. 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. 4D) or different (Fig. 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. 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. 4F; see Methods). Therefore, the first hypothesis is supported. Under this hypothesis, the result in Fig. 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 [50, 51]. 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 [52] 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. 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. 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 [53]. The trends shown in Fig. 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 [42]. Again, the trends in Fig. 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.

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. 2D) is much smaller than what was previously estimated for a pair of linked florescent protein genes [21], due in a large part to the inherently large error in quantifying mRNA concentrations by single-cell RNA-seq [54]. The small size of δ_a (Fig. 3E) is likely caused at least in part by the low efficiency of ATAC-seq in detecting open chromatins (see Methods). The positive correlation between *trans*-r_a and *trans*-r_e (Fig. 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 [21]. Regardless, the effects appear visible to natural selection, as reflected in the preferential chromosomal linkage of genes encoding members of the same protein complex.

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 [55, 56], the two concentrations correlate well at the steady state [21]. 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 [21]. It is interesting to note that in Raj et al.'s study [21], 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 [57]. Corroborating this finding is the recent report [58] 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 [59]. 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 [60], 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 [56, 61], while mRNA concentration fluctuation explains protein concentration fluctuation quite well [21, 58]. 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 [62]. 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. 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) [4, 63]. Third, the protein complex stoichiometry often differs among environments, which makes co-expression of complex components disfavored in the face of environmental changes [64, 65]. 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 [62], 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. 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 [56, 66-73], 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 [66]. 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. S3), 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 [74], 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 [38, 75]. 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 [76]. Note that some other mechanisms such as cell cycle [77] 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 [78], 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 [79], 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.

METHODS

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 [31] 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[34], and the data from 16 NPC cell populations were analyzed. All analyses were performed using custom programs in R or python.

Protein complex data and pre-processing

The mouse protein complex data were downloaded from the CORUM database (http://mips.helmholtz-muenchen.de/corum/) [80]. The coordinates for all mouse protein-coding genes were downloaded from Ensembl BioMart (GRC38m.p5) [49]. 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.

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 [81], one kind of Markov-Chain Monte-Carlo sampling [82]. 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. 4F.

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) [34]. 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.

$$AT = \sum_{i=1}^{N} A_i \text{ and } BT = \sum_{i=1}^{N} B_i . \quad (1)$$

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

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

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

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

$$E(A_i B_j) = E(A_i)E(B_j). \quad (3)$$

Thus,

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

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

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

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

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 [35], 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.

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

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

$$E(A) = r + s,$$
$$E(B) = q + s,$$
$$E(AB) = s,$$
$$Var(A) = (r + s)(p + q),$$
$$Var(B) = (q + s)(p + r). \quad (7)$$

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. S5A). 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. S5B).

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 [54]. 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. S5C) or cell populations (Fig. S5D).

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[6], the only eukaryote with genome-wide protein expression noise data [15]. 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.

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

where we used $K = 10^5$ based on the empirical values of association constants of protein complexes [46]. We then took the average [AB] among all cells to acquire the mean complex concentration.

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 [52] 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 [83]. 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/) [84]:
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

DATA AND SOFTWARE AVAILABILITY

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

ACKNOWLEDGEMENTS

We thank members of the Zhang lab for valuable comments. This work was supported by U.S. National Institutes of Health research grant GM120093 to J.Z.

REFERENCES

1. Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. (2002) Stochastic gene expression in a single cell. Science 297 (5584), 1183-1186.
2. Raser, J.M. and O'shea, E.K. (2005) Noise in gene expression: origins, consequences, and control. Science 309 (5743), 2010-2013.
3. Blake, W.J., Kærn, M., Cantor, C.R. and Collins, J.J. (2003) Noise in eukaryotic gene expression. Nature 422 (6932), 633.
4. Batada, N.N. and Hurst, L.D. (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39 (8), 945-9.
5. Lehner, B. (2008) Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol Syst Biol 4, 170.
6. 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 (16), E67-76.
7. 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. and Vijg, J. (2006) Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature 441 (7096), 1011-4.
8. 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 (21), 13783-8.
9. Veening, J.-W., Smits, W.K. and Kuipers, O.P. (2008) Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol 62, 193-210.
10. Zhang, Z., Qian, W. and Zhang, J. (2009) Positive selection for elevated gene expression noise in yeast. Mol Syst Biol 5, 299.
11. Turing, A.M. (1952) The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences 237 (641), 37-72.
12. 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 (7194), 544.
13. Huang, S. (2009) Non-genetic heterogeneity of cells in development: more than just noise. Development 136 (23), 3853-3862.

14. 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 (5991), 533-8.
15. 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 (7095), 840-6.
16. Stewart-Ornstein, J., Weissman, J.S. and El-Samad, H. (2012) Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol Cell 45 (4), 483-493.
17. 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 (23), 2336-2345.
18. Raj, A. and van Oudenaarden, A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135 (2), 216-26.
19. Sanchez, A., Choubey, S. and Kondev, J. (2013) Regulation of noise in gene expression. Annu Rev Biophys 42, 469-91.
20. 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 (8), e1001621.
21. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. and Tyagi, S. (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4 (10), e309.
22. 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 (11), 1430-1435.
23. 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 (1), 171.
24. 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 (5), 1202-1214.
25. 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 (1), 77.
26. 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 (10), 877.
27. Yan, C., Wu, S., Pocetti, C. and Bai, L. (2016) Regulation of cell-to-cell variability in divergent gene expression. Nat Commun 7, 11099.
28. Dekker, J. and Mirny, L. (2016) The 3D genome as moderator of chromosomal communication. Cell 164 (6), 1110-1121.
29. 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 (3), 268-276.
30. Dekker, J., Marti-Renom, M.A. and Mirny, L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics 14 (6), 390.
31. Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E. and Dekker, J. (2016) Structural organization of the inactive X chromosome in the mouse. Nature 535 (7613), 575-9.

32. Phillips, R., Theriot, J., Kondev, J. and Garcia, H. (2012) Physical biology of the cell, Garland Science.
33. Buenrostro, J.D., Wu, B., Chang, H.Y. and Greenleaf, W.J. (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 109 (1), 21.29. 1-21.29. 9.
34. 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 (3), 377-386.
35. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y. and Greenleaf, W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523 (7561), 486-490.
36. 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 (7398), 376-380.
37. Hahn, S. and Kim, D. (2013) Physical origin of the contact frequency in chromosome conformation capture data. Biophys J 105 (8), 1786-1795.
38. Mahmutovic, A., Fange, D., Berg, O.G. and Elf, J. (2012) Lost in presumption: stochastic reactions in spatial models. Nature methods 9 (12), 1163.
39. 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.
40. 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 (6028), 472-474.
41. Hager, G.L., McNally, J.G. and Misteli, T. (2009) Transcription dynamics. Mol Cell 35 (6), 741-753.
42. 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 (6814), 873.
43. 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 (3), 192.
44. Bernas, T., Brutkowski, W., Zarębski, M. and Dobrucki, J. (2014) Spatial heterogeneity of dynamics of H1 linker histone. Eur Biophys J 43 (6-7), 287-300.
45. Veitia, R.A. (2010) A generalized model of gene dosage and dominant negative effects in macromolecular complexes. The FASEB Journal 24 (4), 994-1002.
46. 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 (suppl_1), D750-D753.
47. 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 (7119), 643.
48. Budnik, B., Levy, E., Harmange, G. and Slavov, N. (2018) Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. bioRxiv, 102681.
49. 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 (D1), D635-D642.
50. Cherry, J.L. (2010) Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol 2, 757-69.
51. 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 (5), e1000944.
52. 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.
53. 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.
54. 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 (3), 496-510.
55. Liu, Y., Beyer, A. and Aebersold, R. (2016) On the dependency of cellular protein levels on mRNA abundance. Cell 165 (3), 535-550.
56. Kustatscher, G., Grabowski, P. and Rappsilber, J. (2017) Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol Syst Biol 13 (8), 937.
57. 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 (6018), 764-768.
58. 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.
59. Gedeon, T. and Bokes, P. (2012) Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations. Biophys J 103 (3), 377-385.
60. Teichmann, S.A. and Veitia, R.A. (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes. Genetics 167 (4), 2121-2125.
61. Franks, A., Airoldi, E. and Slavov, N. (2017) Post-transcriptional regulation across human tissues. PLoS Comput Biol 13 (5), e1005535.
62. Hurst, L.D., Pál, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. Nature Reviews Genetics 5 (4), 299-310.
63. Chen, X. and Zhang, J. (2016) The genomic landscape of position effects on protein expression level and noise in yeast. Cell Syst 2 (5), 347-54.
64. 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 (1), 47.
65. Slavov, N., Semrau, S., Airoldi, E., Budnik, B. and van Oudenaarden, A. (2015) Differential stoichiometry among core ribosomal proteins. Cell reports 13 (5), 865-873.
66. Levesque, M.J. and Raj, A. (2013) Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nature methods 10 (3), 246.
67. Fukuoka, Y., Inaoka, H. and Kohane, I.S. (2004) Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics 5 (1), 4.
68. 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 (3), 767-775.

69. Sémon, M. and Duret, L. (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol 23 (9), 1715-1723.
70. 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 (3), 825-831.
71. Spellman, P.T. and Rubin, G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1 (1), 5.
72. Ghanbarian, A.T. and Hurst, L.D. (2015) Neighboring genes show correlated evolution in gene expression. Mol Biol Evol 32 (7), 1748-1766.
73. Liao, B.-Y. and Zhang, J. (2008) Coexpression of linked genes in Mammalian genomes is generally disadvantageous. Mol Biol Evol 25 (8), 1555-1565.
74. 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 (6161), 948-953.
75. Elf, J. and Barkefors, I. (2018) Single-molecule kinetics in living cells. Annu Rev Biochem.
76. Carrera, J. and Covert, M.W. (2015) Why build whole-cell models? Trends Cell Biol 25 (12), 719-722.
77. 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 (8), 809.
78. Raj, A., Rifkin, S.A., Andersen, E. and Van Oudenaarden, A. (2010) Variability in gene expression underlies incomplete penetrance. Nature 463 (7283), 913-918.
79. Baker, M. (2011) Synthetic genomes: The next step for the synthetic genome. Nature 473 (7347), 403-408.
80. 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 (suppl_1), D497-D501.
81. Geman, S. and Geman, D. (1987) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In Readings in Computer Vision, pp. 564-584, Elsevier.
82. Gilks, W.R. (2005) Markov chain monte carlo. Encyclopedia of Biostatistics.
83. Medrzycki, M., Zhang, Y., Cao, K. and Fan, Y. (2012) Expression analysis of mammalian linker-histone subtypes. Journal of visualized experiments: JoVE (61).
84. 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 (suppl_1), D724-D728.

FIGURE LEGENDS

Fig. 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. 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. *P*-values from binomial tests on independent gene pairs are presented. (C) Fraction of gene pairs with positive δ_e in each chromosome. 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. 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.

Fig. 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. *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, $md \pm 1.5(qu_3 - qu_1)$, 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. Spearman's ρ computed from unbinned data and associated *P*-value determined by a Mantel test are presented. (H) 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.

Fig. 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.

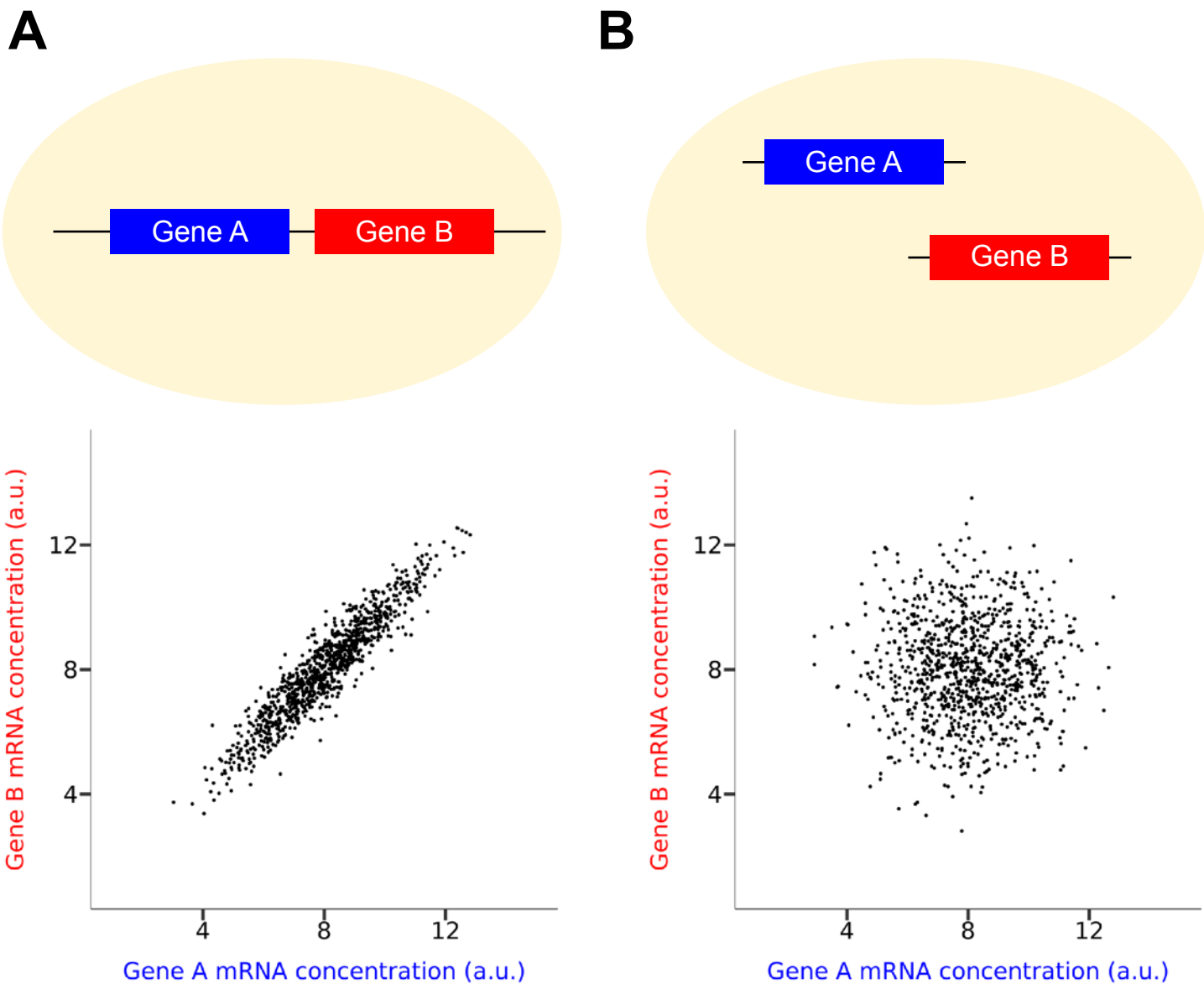

Figure 1

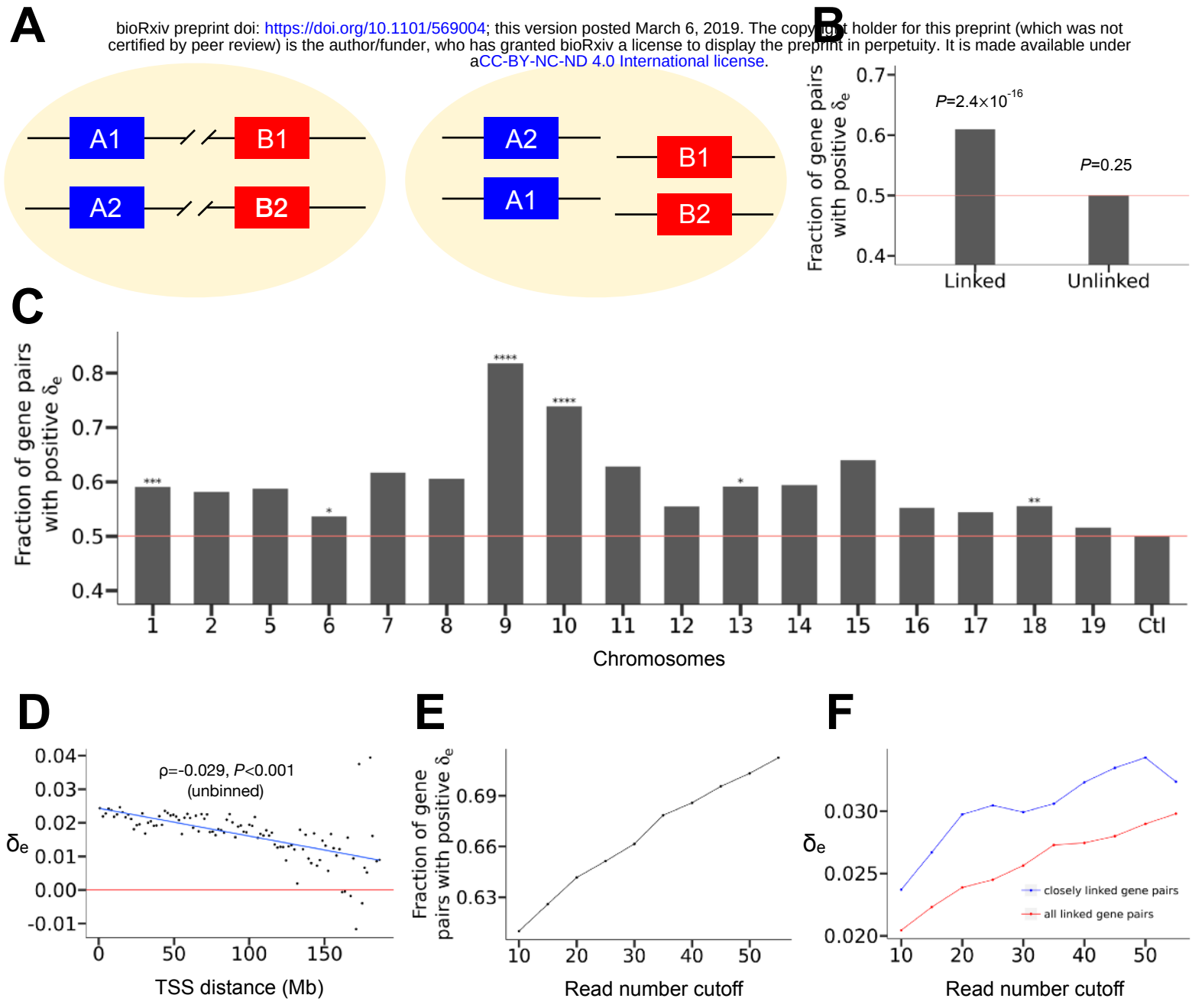

Figure 2

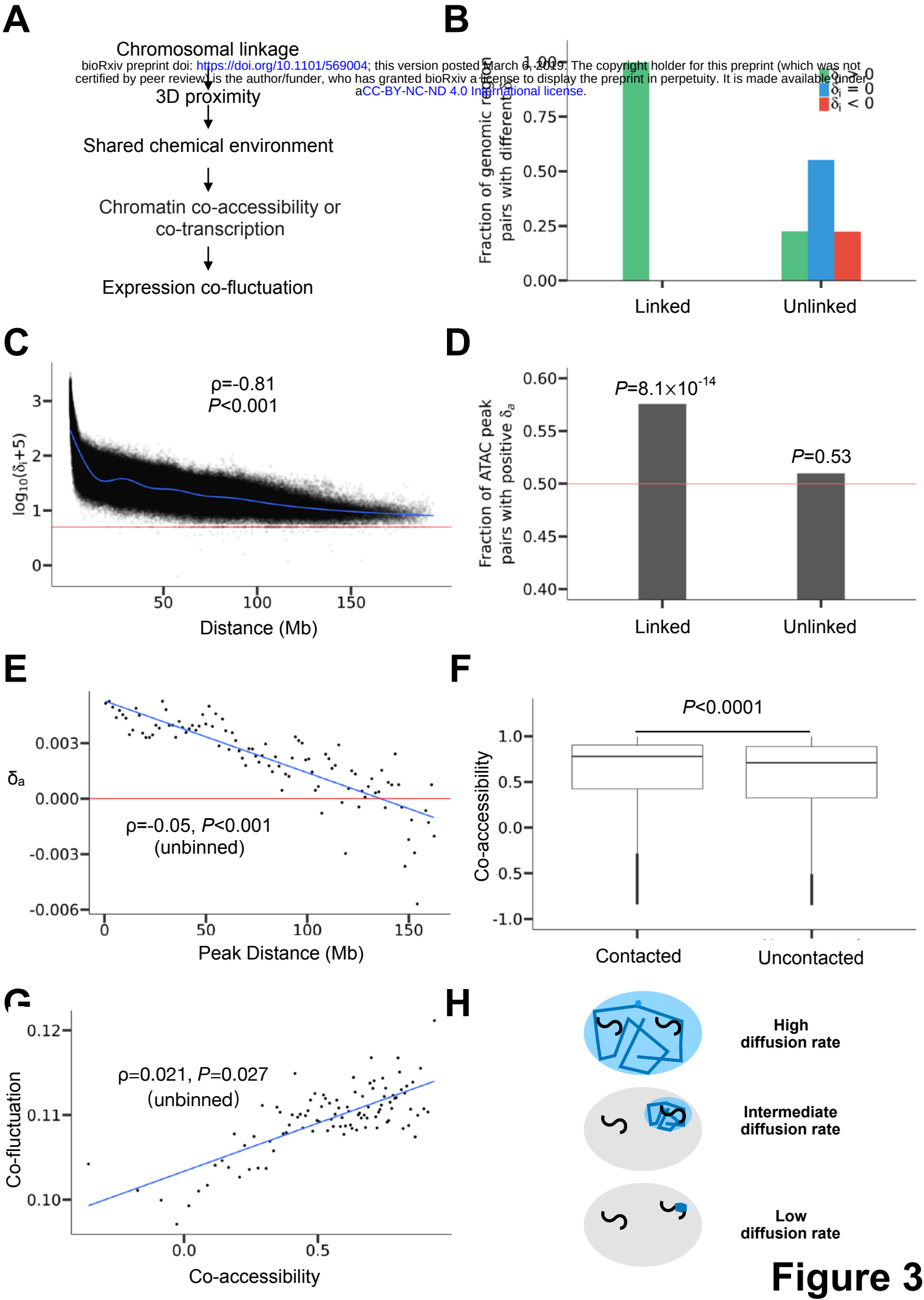

Figure 3

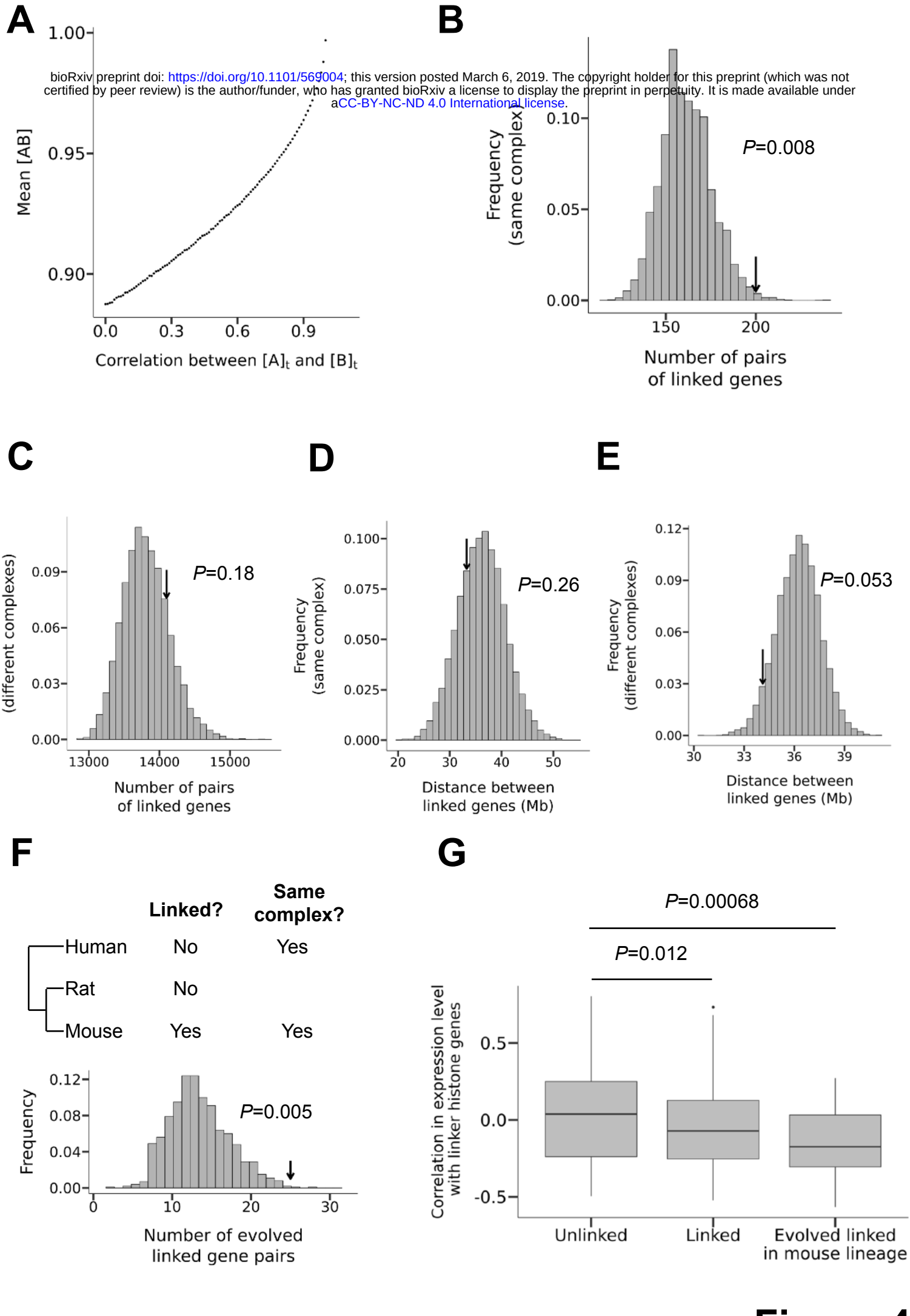

Figure 4

LEGENDS OF SUPPLEMENTARY FIGURES

Fig. S1. 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. S2. 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.

Fig. S3. 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. S4. δ_e for pairs of neighboring genes with different orientation types. 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, $md \pm 1.5(qu_3-qu_1)$, 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. S5. Chromatin co-accessibility between two ATAC peaks quantified using single-cells vs. using cell populations. (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%.

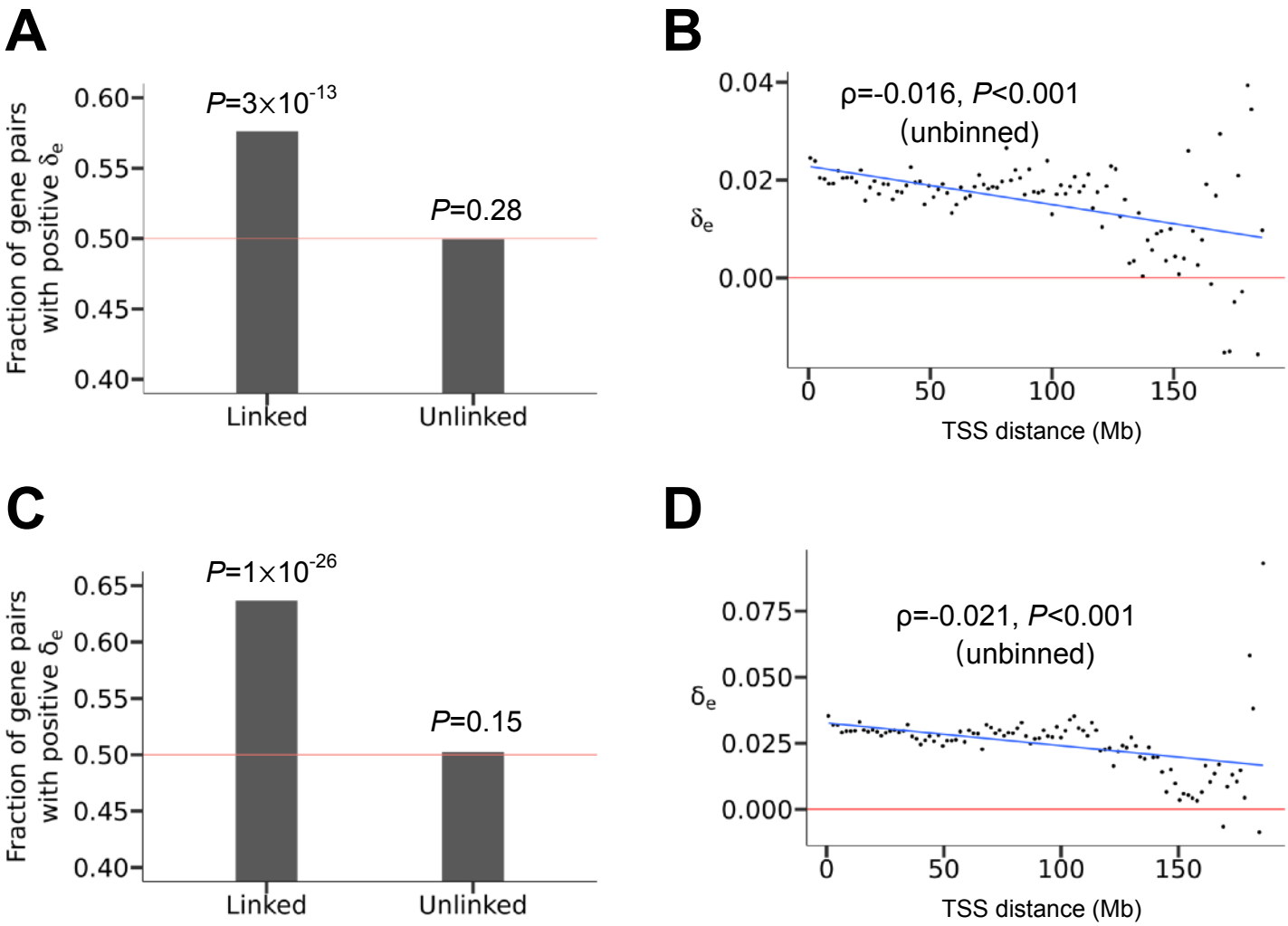

Figure S1

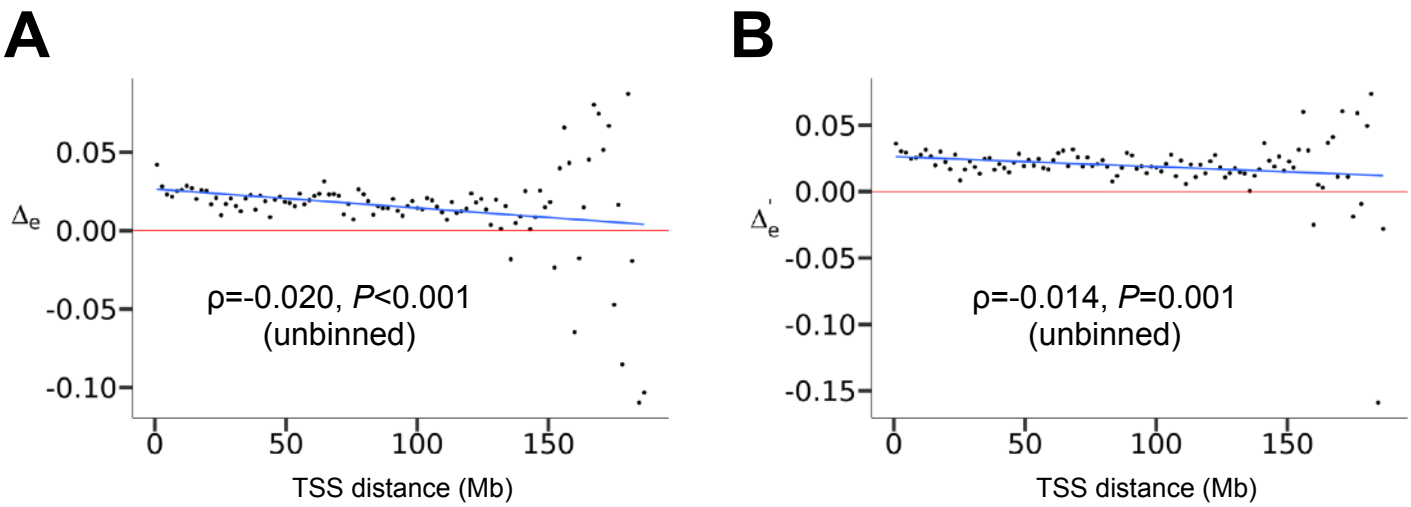

Figure S2

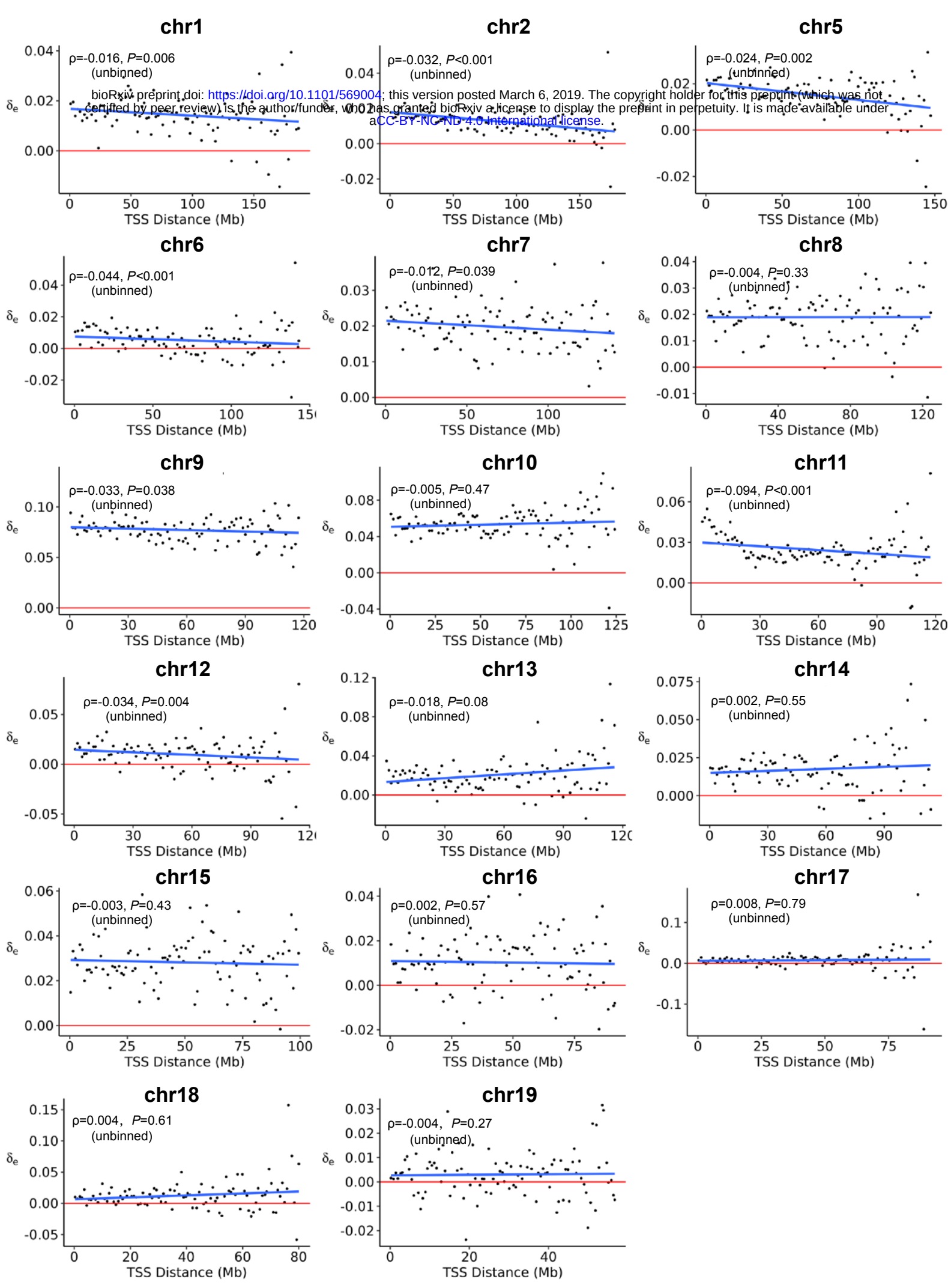

Figure S3

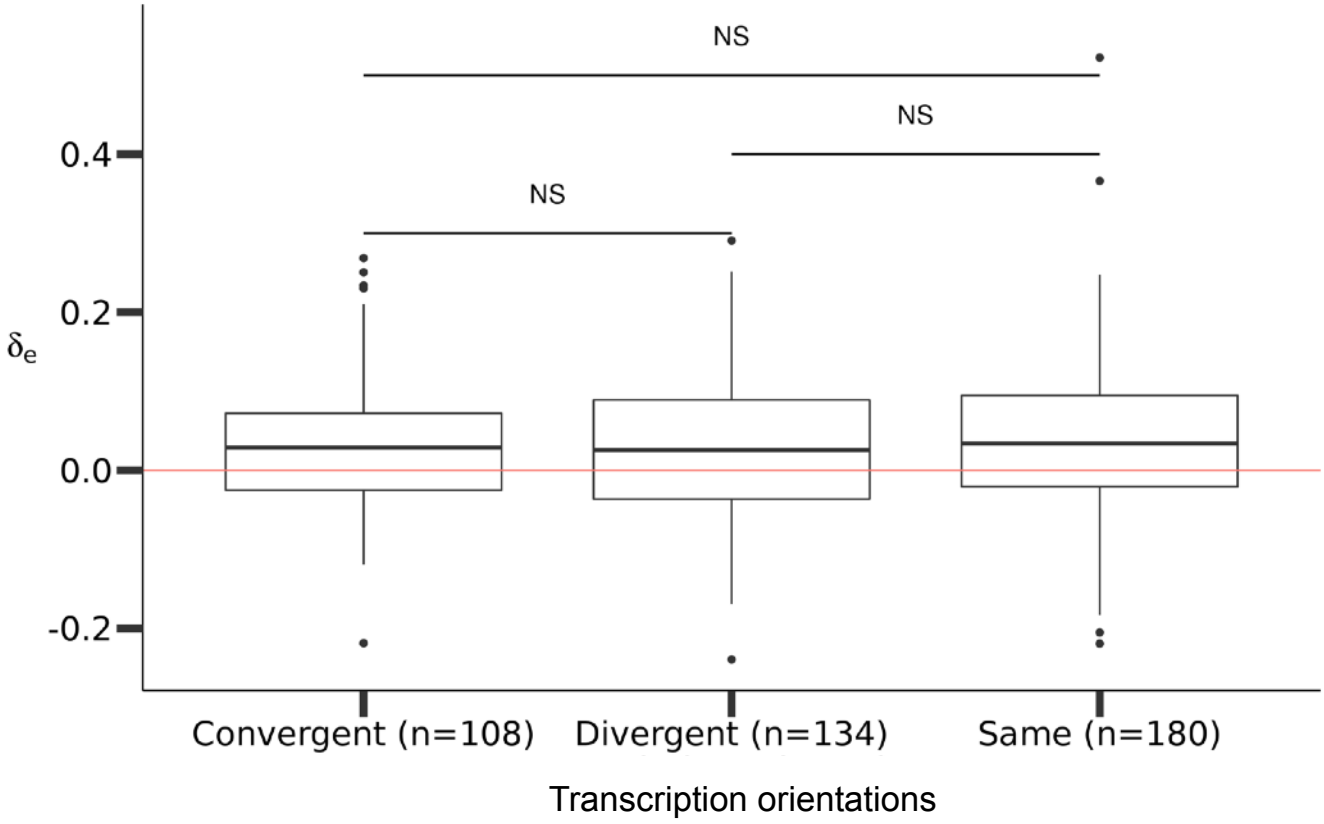

Figure S4

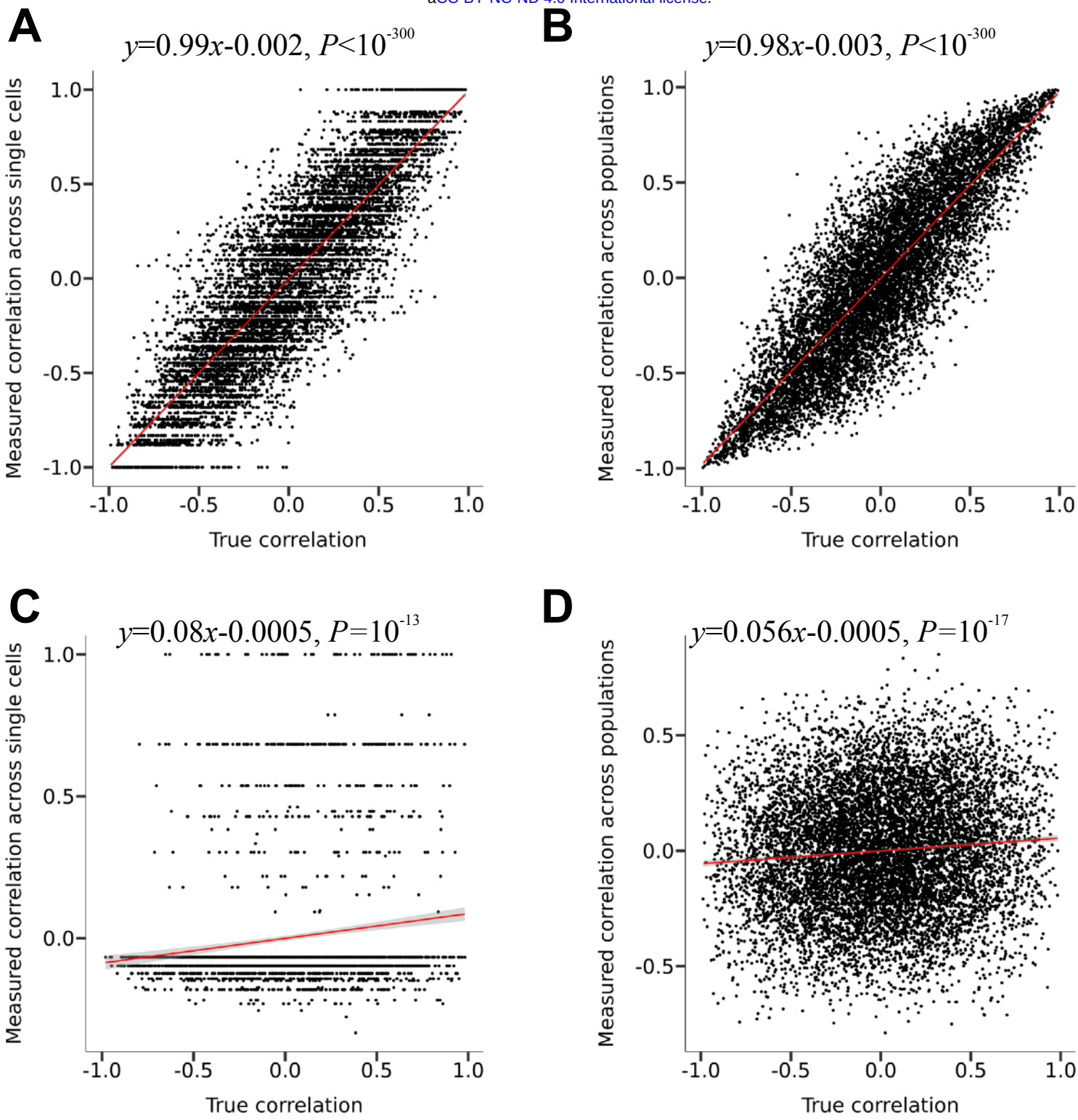

Figure S5