



# Transcriptome Analysis in Yeast Reveals the Externality of Position Effects

Qian Gui,<sup>†,1</sup> Shuyun Deng,<sup>†,1</sup> ZhenZhen Zhou,<sup>1</sup> Waifang Cao,<sup>1</sup> Xin Zhang,<sup>1</sup> Wenjun Shi,<sup>1</sup> Xiujuan Cai,<sup>1</sup> Wenbing Jiang,<sup>1</sup> Zifeng Cui,<sup>2</sup> Zheng Hu <sup>\*,2</sup> and Xiaoshu Chen <sup>\*,1</sup>

<sup>1</sup>Department of Biology and Medical Genetics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

<sup>2</sup>Department of Obstetrics and Gynecology, Precision Medicine Institute, First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding authors: E-mails: chenxshu3@mail.sysu.edu.cn; huzheng1998@163.com.

Associate editor: Jian Lu

## Abstract

The activity of a gene newly integrated into a chromosome depends on the genomic context of the integration site. This “position effect” has been widely reported, although the other side of the coin, that is, how integration affects the local chromosomal environment, has remained largely unexplored, as have the mechanism and phenotypic consequences of this “externality” of the position effect. Here, we examined the transcriptome profiles of approximately 250 *Saccharomyces cerevisiae* strains, each with GFP integrated into a different locus of the wild-type strain. We found that in genomic regions enriched in essential genes, GFP expression tended to be lower, and the genes near the integration site tended to show greater expression reduction. Further joint analysis with public genome-wide histone modification profiles indicated that this effect was associated with H3K4me2. More importantly, we found that changes in the expression of neighboring genes, but not GFP expression, significantly altered the cellular growth rate. As a result, genomic loci that showed high GFP expression immediately after integration were associated with growth disadvantages caused by elevated expression of neighboring genes, ultimately leading to a low total yield of GFP in the long run. Our results were consistent with competition for transcriptional resources among neighboring genes and revealed a previously unappreciated facet of position effects. This study highlights the impact of position effects on the fate of exogenous gene integration and has significant implications for biological engineering and the pathology of viral integration into the host genome.

**Key words:** position effects, essential genes, fitness.

## Introduction

Gene integration is a major type of genomic alteration commonly observed in both natural (e.g., viral integration into the host genome [Ciuffi 2016], transposons [Yant et al. 2005], and horizontal gene transfer [Keeling and Palmer 2008]) and artificial circumstances (Ivics et al. 2009). Depending on the location of the genomic integration, the activity of the integrated gene varies substantially (Akhtar et al. 2013; Feuerborn and Cook 2015), as does the phenotypic outcome of the integration event (Sturtevant 1925; Kleinjan and van Heyningen 1998; Grewal and Jia 2007); this phenomenon is commonly referred to as the “position effect.” A classic example of the position effect is the translocation of the white gene of fruit fly (*Drosophila melanogaster*) into heterochromatin, giving the original solid red eye a white and red mottled appearance (Sturtevant 1925; Grewal and Jia 2007). Recently, genome-wide studies have provided additional mechanistic details regarding position effects, such as the regulatory role of enhancers, gene order, various epigenetic modifications, chromatin domains, and 3D localization

(Akhtar et al. 2013; Chen et al. 2013; Dey et al. 2015; Chen and Zhang 2016). It is therefore not surprising that position effects have had significant impacts on the evolution of chromosome organization (Batada and Hurst 2007), improvements in genetic engineering (Wilson et al. 1990), and a number of genetic diseases (Milot et al. 1996; Kleinjan and van Heyningen 1998).

Despite extensive efforts to clarify the influence of position effects on the function of focal integrated genes, little is known about how this integration affects other genes. Theoretically, the integration of a gene would significantly alter transcriptional regulation, thereby causing changes in the activity of other genes, especially those sharing local transcriptional resources with the integrated gene. Here, transcriptional resources refer to the transcription factors, coregulators, RNA polymerases (Silveira and Bilodeau 2018), and other possible factors found in a designated area. On the one hand, it is possible that the integrated gene competes with nearby genes for local transcriptional resources, thereby reducing the activities of nearby genes, as observed in the phenomenon of promoter interference (Pande et al. 2018;

Strainic et al. 2000). On the other hand, it is possible that the promoter of the integrated gene recruits additional transcriptional resources that were inaccessible to the nearby genes before integration, thereby increasing the activity of the nearby genes, as observed in the “ripple effect” of endogenous genes (Ebisuya et al. 2008). These two possibilities are hereafter referred to as “transcriptional competition” and “transcriptional synergism,” respectively.

Where any gene integration event falls along the spectrum from transcriptional competition to synergism should depend on both the integrated gene and the genomic context of integration. Here, we focused on the latter and aimed to find general patterns independent of gene function, which is referred to as the “externality of the position effect,” just as the position effect is a general phenomenon independent of the function of the integrated gene. Data relevant to the externality of the position effect are sporadic. For example, a previous study showed that integration into four different loci in the yeast genome resulted in several changes in the transcriptional profile (Chen et al. 2013). In addition, separate integration into 63 loci on yeast chromosome 1 did not cause dramatic changes in the expression level of the HO locus on chromosome 4 (Chen and Zhang 2016). However, these studies and other studies in diploid yeast (Bucci et al. 2018; Harvey et al. 2020; Liu et al. 2020) examined only a few integration events or distal genes and thus might not be sufficiently representative of the impacts on genes close to integration sites.

For systematic profiling of the externality of the position effect, we need a metric reflecting biological context in terms of the transcriptional environment. In this context, one prevailing theory has proposed that essential genes (Winzeler et al. 1999), the deletion of which causes lethality, tend to cluster in the genome, thereby forming regions of open chromatin that promote continuous transcription. This nonrandom distribution ensures that the expression noise of essential genes, which is likely highly deleterious, is minimized (Batada and Hurst 2007; Wang and Zhang 2011; Chen and Zhang 2016). Consistent with the biological relevance of essential gene clusters, it has been shown that the order of genes in the genome and the transcriptional profiles of the wild-type strain are the results of long-term evolutionary optimization (Pal and Hurst 2003; Yang et al. 2017). Based on this theory, we speculated that the density of essential genes is a proxy for the allocation of intracellular transcriptional resources and therefore dictates the prevalence of transcriptional competition and/or transcriptional synergism.

To test our hypothesis, we conducted transcriptome deep sequencing of approximately 250 *Saccharomyces cerevisiae* strains, which were randomly picked from a previously constructed library with *GFP* cassettes individually integrated into various loci across all the chromosomes (Chen and Zhang 2016). We found that integrations into genomic loci enriched in essential genes (based on either linear or 3D proximity) showed decreased transcriptional activity of the integrated gene as well as the adjacent genes, consistent with the model of transcriptional competition. In contrast, integrations into genomic loci where essential genes were few showed increased transcriptional activity of both the integrated and

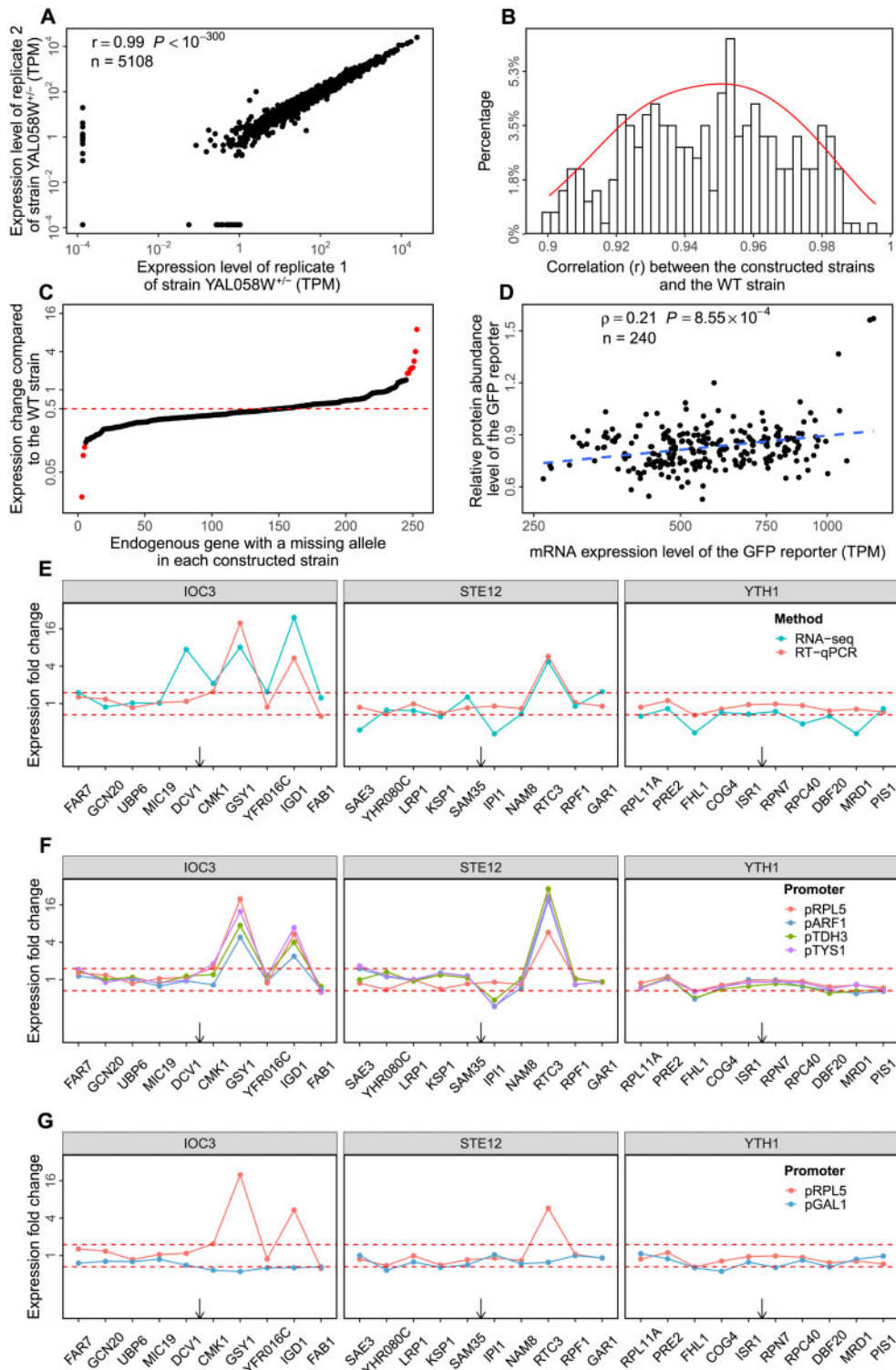
the adjacent genes, consistent with the model of transcriptional synergism. The observed externality of the position effect was at least partially explained by the specific histone methylation status of the surrounding genomic regions. Intriguingly, the changes in expression of the neighboring genes, rather than that of the integrated gene, were correlated with the rate of cellular growth, highlighting the possibility that the externality of the position effect might be phenotypically more important than the position effect itself. Overall, we revealed a previously underappreciated mechanism for the phenotypic consequences of position effects with broad implications in both bioengineering and biomedicine.

## Results

### Transcriptome Sequencing of 240 Yeast Strains with Individual *GFP* Integrations

We previously replaced the kanMX module in heterozygous deletion strains of *S. cerevisiae* at hundreds of different loci across all the chromosomes with an expression cassette comprising the marker gene *URA3* and a *GFP* gene driven by the *RPL5* promoter (*pRPL5-GFP*; *RPL5*, ribosomal 60S subunit protein L5) (Chen and Zhang 2016). To examine the effect of this integrated *GFP* cassette on adjacent genes, we randomly selected over 250 strains from this reconstructed heterozygous deletion library and sequenced the transcriptome of each selected strain (see Materials and Methods, supplementary tables S1 and S2, [Supplementary Material](#) online). In each strain, the deletion of an allele of an endogenous gene has minimal impacts on the position effects (see below) because most yeast genes are haplosufficient (Deutschbauer et al. 2005). We further performed the following quality control steps to ensure that the transcriptome data sets were comparable with each other. First, we confirmed that our experiments were highly reproducible, with strong correlations between the transcriptome profiles from different biological replicates (fig. 1A). Second, strains with severe haploinsufficiency due to the heterozygous deletion of endogenous genes were removed, such that only strains with transcriptome profiles having limited deviation from that of the wild-type strain (Pearson's correlation coefficient > 0.9) were used for the subsequent analysis (fig. 1B). Third, for the heterozygously deleted genes, we estimated the ratio between the gene expression in the constructed *GFP* strains and that in the wild-type strain, which was expected to be 0.5. A constructed strain was excluded from further analyses if this ratio was an outlier among those for all the constructed strains (fig. 1C); this precaution ensured a negligible effect of feedback regulation on the heterozygously deleted gene. Finally, the transcriptome profiles of the constructed strains representing 240 loci integrated with the *GFP* cassette were retained for downstream analyses.

To further corroborate the reliability of our transcriptome data sets, we compared the RNA-seq-based mRNA expression levels of *GFP* with previously measured protein abundance levels based on flow cytometry (Chen and Zhang 2016). We found that mRNA expression was significantly correlated with protein abundance (Spearman's rank



**FIG. 1.** Our transcriptome data set is of high quality. (A) Reproducibility between two biological replicates of strain  $CNE1^{+/-}$ . (B) Distribution of Pearson's correlation coefficients between the transcriptome profile of each constructed strain and the transcriptome profile of the wild-type strain. The red line represents the kernel density estimate of the distribution of correlation coefficients. (C) Changes in the expression of the endogenous gene missing one allele in each constructed strain compared with the wild-type strain. The red dots indicate the constructed strains for which changes in the expression of endogenous genes are outliers among the changes in expression of all the examined endogenous genes missing one allele. The red dotted line indicates the expected value of 0.5. (D) The mRNA expression level of the GFP gene that we tested was significantly correlated with the protein abundance level measured in previous studies (Chen and Zhang 2016). The dotted line represents the fitted linear regression model. (E) The fold changes in expression (compared with expression in the wild-type strain) of five genes each upstream and downstream ( $x$ -axis) of the GFP integration site in three constructed strains as detected by RT-qPCR (red) and RNA-seq (cyan-blue). The red dotted lines indicate expression changes of 1.5-fold upregulation or downregulation (66.7%) compared with the expression in the wild-type strain. The arrows indicate the GFP integration sites. (F) Similar to the red line in (E), except that the RPL5 (red) promoter was replaced with the TDH3 (green), ARF1 (blue) and TYS1 (purple) promoters. (G) Similar to the red line in (E), except that the GAL1 (blue) promoter was used instead of the RPL5 promoter to drive the expression of GFP.

correlation coefficient  $\rho = 0.21$ ,  $P < 9 \times 10^{-4}$ , [fig. 1D](#)). Moreover, the mRNA expression spanned a 5-fold range, whereas the protein abundance spanned a 2.5-fold range ([fig. 1D](#)). Such a slight decrease in the range of protein expression relative to that of mRNA expression is consistent with the known translational buffering of transcriptional variation in yeast ([Artieri and Fraser 2014](#)). Notably, compared with the finding in our previous study that the position effect could change protein abundance by at least 15-fold ([Chen and Zhang 2016](#)), the strains retained after the above quality control procedures in this study did not exhibit the full range of effects possibly due to the position effect.

In addition, we selected three constructed strains and quantified the mRNA expression of ten genes flanking each GFP integration site (five on each side) by real-time quantitative polymerase chain reaction (RT-qPCR). We found that the fold changes in expression relative to that of the wild-type strain as measured by RT-qPCR were highly consistent with those measured by RNA-seq ([fig. 1E](#), see Materials and Methods). We also replaced the RPL5 promoter with other representative promoters (*pTDH3*, *pARF1*, and *pTYS1*) ([Chen and Zhang 2016](#)) and found that these four promoters had the same effect on the expression of the adjacent genes ([fig. 1F](#), see Materials and Methods), indicating that this effect was not promoter-specific. Furthermore, it has been previously shown that the chromatin states reflected by histone modifications such as H3K4me1 can significantly impact the mRNA expression of both the integrated GFP and the endogenous genes at the same locus ([Chen, et al. 2013](#)), and this finding was recapitulated by our transcriptome data (supplementary [fig. S1](#), [Supplementary Material](#) online).

To further distinguish whether the observed expression change of adjacent genes is caused by the integration and expression of GFP or the heterozygous deletion of the endogenous gene, we replaced the RPL5 promoter with the GAL1 promoter (*pGAL1*). Since *pGAL1* is inactive in yeast peptone dextrose (YPD) medium, any expression change of adjacent genes observed for this *pGAL1* strain can be caused only by the heterozygous deletion of the endogenous gene. We found that in this *pGAL1* strain, genes adjacent to the integration site did not show significant expression changes in YPD medium ([fig. 1G](#), see Materials and Methods) but showed significant expression changes (supplementary [fig. S2](#), [Supplementary Material](#) online) in YPG medium (where *pGAL1* is active). Therefore, we confirmed that the heterozygous deletion of the endogenous gene in yeast did not cause significant expression changes in the adjacent genes, and the RPL5 promoter revealed that the expressional effects on adjacent genes were mainly caused by the integration and expression of GFP.

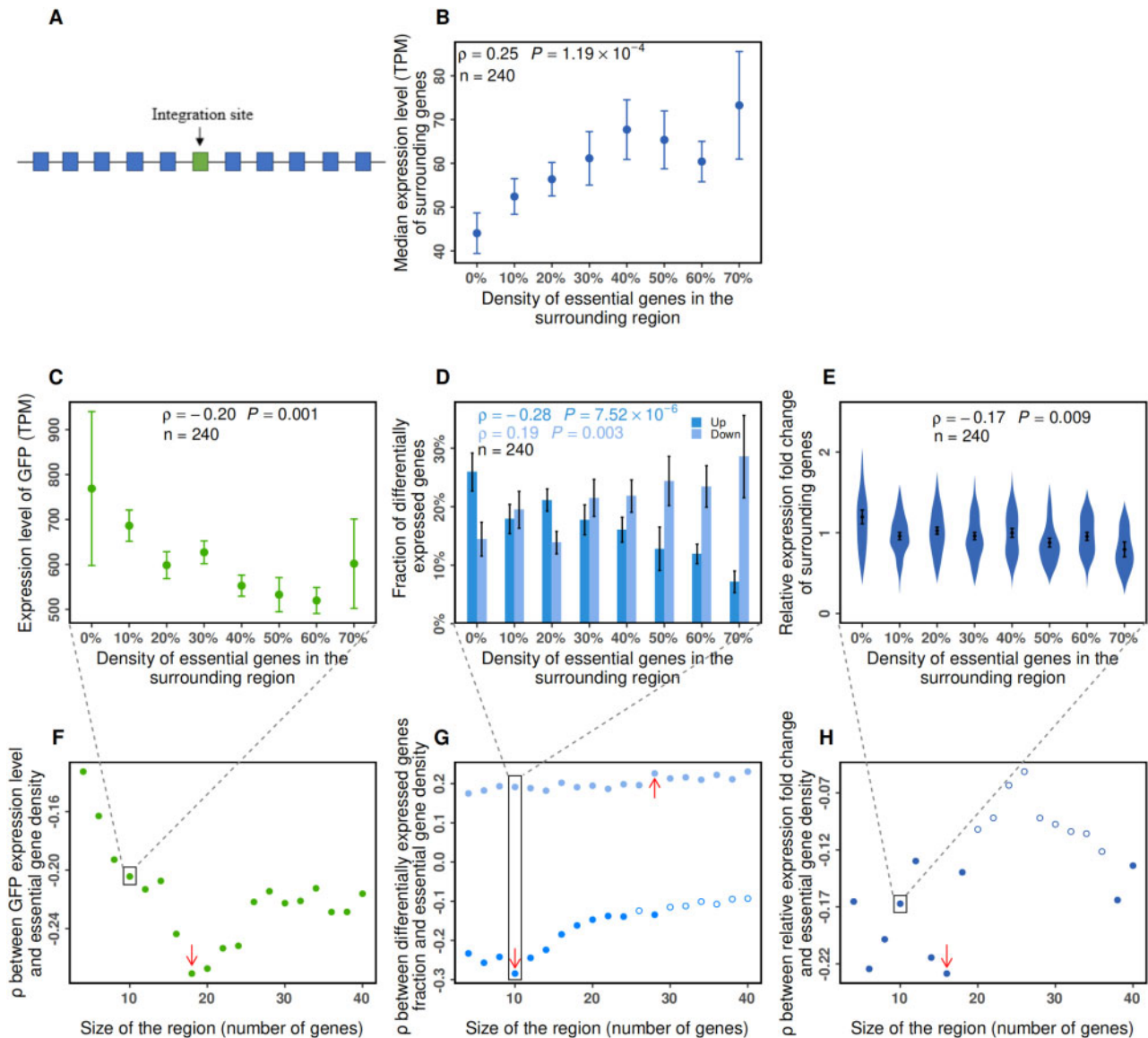
Overall, the above results indicated that our transcriptome data set was of high quality and was sufficiently representative to detect position effects and their potential externality.

### Externality of Position Effects

According to a widely accepted model explaining how position effects drive the nonrandom distribution of genes along chromosomes, clusters of essential genes indicate genomic

regions where genes could have increased transcriptional activity and decreased expression stochasticity ([Batada and Hurst 2007](#); [Wang and Zhang 2011](#); [Chen and Zhang 2016](#)). In support of this model, we found in the transcriptome profile of the wild-type strain that the median expression level of the surrounding genes was significantly correlated with the density of essential genes in the surrounding genomic region, which included five genes upstream and five genes downstream of the GFP integration site ([fig. 2A and B](#), see Materials and Methods). This observation is also consistent with the expression similarity commonly observed among neighboring genes ([Ghanbarian and Hurst 2015](#)). However, when we examined the expression of GFP in the heterozygous deletion strains, we were surprised to find the opposite trend: GFP expression was negatively correlated with the density of essential genes ([fig. 2C](#)). Notably, since strains with extreme GFP expression levels were removed during quality control, the actual opposite trend would be stronger than our results indicate. In addition, we also found that the expression of URA3, which marked the strains we constructed, was negatively correlated with the density of essential genes (supplementary [fig. S3](#), [Supplementary Material](#) online). We were also able to exclude potential confounding factors such as the length of the deleted gene and the GC content of the genomic context (supplementary [fig. S4](#), [Supplementary Material](#) online). Our results therefore suggested that the essentiality of the surrounding genes has an impact on the expression of the focal gene.

How can we explain these opposite observations for the surrounding endogenous genes and the integrated GFP gene? One major difference between the GFP gene and the endogenous genes was that the localization of each endogenous gene was presumably optimized by natural selection, such that endogenous genes with strong promoter or transcriptional activity, and therefore high and stable expression, would likely be located in genomic regions enriched in essential genes. However, for GFP, the expression differences might mostly be the result of strong transcriptional competition when the local density of essential genes is high, since the promoters of the essential genes are most likely more competitive than those of the nonessential genes. Such competitiveness of essential genes presumably evolves to ensure their high expression level and reduced expression noise in individual cells because lowered expression of essential genes is more detrimental than that of nonessential genes ([Batada and Hurst 2007](#); [Chen and Zhang 2016](#); [Wang and Zhang 2011](#)). To test whether genomic regions with a high density of essential genes indeed harbor strong transcriptional competition, we examined the changes in expression after GFP integration in the genes surrounding the integration site. In the absence of position effect externality, GFP integration should have no effect on the neighboring genes. However, we found that as the local density of essential genes increased, the fraction of upregulated genes decreased ([fig. 2D](#), dark blue bars), whereas the fraction of downregulated genes increased ([fig. 2D](#), light blue bars). The above observations remained qualitatively unchanged when different thresholds for differential expression were used (supplementary [fig. S5](#),



**Fig. 2.** Expression of the GFP gene and changes in the expression of adjacent genes are related to essential gene density near the integration site. (A) Linear positional relationship between the GFP integration site and adjacent genes. (B–E) The median expression of adjacent genes before integration (B), the expression level of GFP after integration (C), the fractions of genes with upregulated expression (D, dark blue bars), the fractions of genes with downregulated expression (D, light blue bars), and the median fold changes in expression of adjacent genes (E) were significantly related to the density of essential genes in the surrounding region. The surrounding regions included five genes upstream and five genes downstream of the GFP integration site. The points or bars represent the mean, and the error bars represent the standard errors within each range of essential gene densities. Each violin plot represents the distribution of data. (F–H) The correlation coefficient with the density of essential genes is shown for the expression level of GFP (F), the fractions of genes with upregulated expression (G, dark blue dots), the fractions of genes with downregulated expression (G, light blue dots) and the median fold change in expression of adjacent genes (H) when genomic regions of different sizes (in terms of number of adjacent genes) are considered. The size of the region represents the total number of upstream and downstream adjacent genes. The genes for which expression was increased more than 1.5-fold in the constructed strains compared with the wild-type strain were considered to be upregulated in expression, and genes for which the expression was reduced to  $<66.7\%$  were considered to be downregulated in expression. The solid dots indicate  $P$  values of the correlation coefficient smaller than 0.05 (i.e., statistically significant), and the unfilled dots indicate  $P$  values greater than or equal to 0.05 (i.e., not statistically significant). The red arrow indicates the extreme value of the trend of the continuously significant correlation coefficient. For one boxed dot in each panel, the underlying original correlation is magnified, as shown in C/D/E.

Supplementary Material online). These up- and downregulation events were not sporadic, even considering that the position effect was probably underestimated due to quality control. For the chromosomal regions containing ten genes surrounding the integration site, an average of approximately

four genes were significantly up- or downregulated (fig. 2D). Moreover, the median fold change in expression was also anticorrelated with the local density of essential genes in the genomic region surrounding the integration site (fig. 2E). These results suggested that GFP integration in a

genomic region with a high local density of essential genes led to significant expression downregulation of genes surrounding the integration site, indicating strong transcriptional competition rather than transcriptional synergism. In contrast, the expression of genes surrounding an integration site increased only in genomic regions with no adjacent essential genes. In other words, the transcriptional synergism caused by an integrated highly expressed gene trumped transcriptional competition only when there were no adjacent essential genes, and the dominance of transcriptional competition continued to increase with an increasing local density of essential genes.

In addition, since the expression changes of the surrounding genes were apparently not caused by the native function of GFP, the above observation suggested that the genomic context of gene integration affects the transcriptional activity of not only the focal integrated gene but also the surrounding genes. This externality constitutes a previously unrecognized facet of the position effect. To determine the range of this externality, we examined genomic regions of different sizes (in terms of the number of genes flanking each side of the integration site). We found that in a genomic region containing up to 40 surrounding genes (20 on each side of the integration site), the local density of essential genes was still negatively correlated with GFP expression level, with the strongest correlation found for 18 surrounding genes (fig. 2F). For the fraction of genes with upregulated expression (fig. 2G, dark blue dots), the fraction of genes with downregulated expression (fig. 2G, light blue dots), and the median change in expression of the endogenous genes (fig. 2H), the size of the genomic region that showed significant externality was up to 30, 40, and 20 surrounding genes, with the most significant externality effects at 10, 28, and 16 surrounding genes, respectively. We also tried similar analyses using genomic distance measured by kilo base pairs (kb) rather than the number of adjacent genes and found similar patterns (supplementary fig. S6, [Supplementary Material](#) online). In addition, one might argue that the expression level of the surrounding genes is a better proxy for the competitiveness of the local transcriptional environment. We tested this notion with partial correlation analyses and found that the fraction of essential surrounding genes was superior to the expression level of surrounding genes in explaining the observed externality of position effects (supplementary fig. S7, [Supplementary Material](#) online). This observation therefore suggested that the local density of essential genes might have captured some transcriptional regulatory features independent of expression level, such as persistently open (but not necessarily transcribing) chromatin domains (Batada and Hurst 2007).

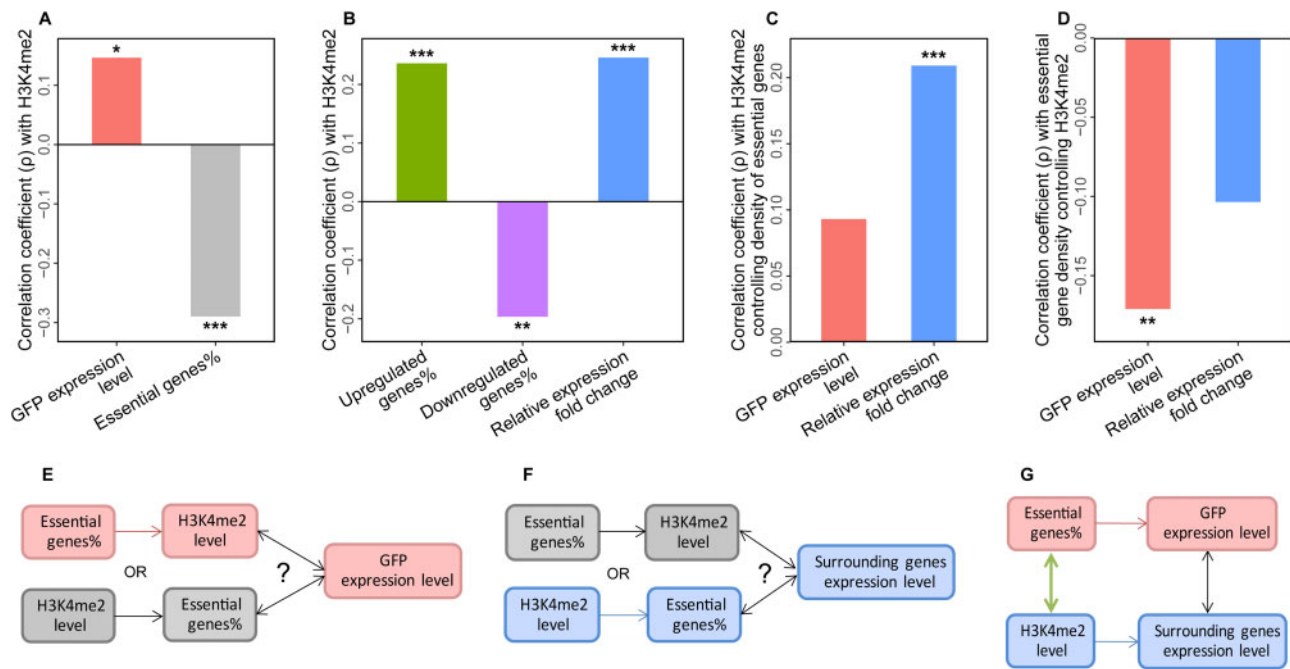
Transcriptional resources are allocated not only linearly along DNA molecules but also three-dimensionally in so-called “transcriptional factories,” as different genomic regions fold into specific focal sites of active transcription (Jackson et al. 1993). If the externality of position effects can indeed be explained by competition for transcriptional resources, we should predict a similar effect for the density of essential genes in 3D proximity to the GFP integration site (supplementary fig. S8A, [Supplementary Material](#) online). We thus tested the above prediction with a 3D model of

the yeast genome (Duan et al. 2010) and found patterns similar to those for linear proximity to the GFP integration site when we considered different numbers of genes (supplementary fig. S8, [Supplementary Material](#) online) and different physical distances (in nanometers or nm, supplementary fig. S9, [Supplementary Material](#) online). Although the majority (~65%) of the three-dimensionally adjacent genes was also linear neighbors of the GFP integration site, excluding these linear neighbors did not change our conclusion regarding the significance of position effect externality (supplementary fig. S10, [Supplementary Material](#) online). Collectively, these results suggested that position effect externality can influence the expression of the genes surrounding the integration site, presumably via competition for transcriptional resources. This phenomenon deserves further investigation to determine its underlying mechanism and its contribution to the phenotypic consequences of the position effect.

### Contribution of Histone Modifications to the Externality of the Position Effect

What is the molecular mechanism underlying the externality of position effects? Theoretically, changes in expression level could be limited by either *cis* or *trans* factors, where *cis* means the epigenetic state of the gene with changed expression and *trans* means the local availability of transcriptional machinery near the integration site. Although it is difficult to probe specifically for the local concentration of PolII or other components of the transcriptional machinery, it is possible to detect whether *cis* factors have contributed to the observed expression change. More specifically, since histone modifications are known to be associated with the expression of reporter genes in yeasts (Soares et al. 2017), we hypothesized that histone modifications are also involved in regulating changes in expression around integration sites. Here, our working model is that histone modifications affect how the local transcriptional environment responds to GFP integration but not that GFP integration alters histone modification and then the expression of adjacent genes. We collected high-throughput sequencing-based profiles for eight major types of histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3, H3K9ac, H3K12ac, and H3K14ac; see Materials and Methods, supplementary tables S1 and S3, [Supplementary Material](#) online) in the wild-type strain, which was previously shown to be mostly undisturbed by integration of expression cassettes (Chen et al. 2013). We then estimated the histone modification levels of the genes around the integration site and determined the relationships of these levels with the local density of essential genes and the expression level of the integrated GFP (see Materials and Methods).

We found that in genomic regions with high GFP expression levels, the surrounding genes tended to have significantly stronger H3K4me1 (supplementary fig. S11, [Supplementary Material](#) online) and H3K4me2 signals (fig. 3A, red bar). Such patterns are consistent with the finding of a previous report that H3K4me1 and H3K4me2 are generally associated with active transcription (Soares et al. 2017). In addition, we found



**Fig. 3.** Expression of the GFP gene and changes in the expression of adjacent genes are related to the H3K4me2 modification level of the surrounding region. (A, B) Correlations with the level of H3K4me2 modification of the genes surrounding the integration site are shown for the expression level of the GFP gene (A, red bar), the local density of essential genes (A, gray bar), the fractions of upregulated genes (B, green bar), the fractions of downregulated genes (B, purple bar), and the median fold change in adjacent gene expression (B, blue bar). (C) Partial correlations between the level of H3K4me2 modification of genes surrounding the integration site after controlling for the local density of essential genes are shown for the expression level of GFP (C, red bar) and the median fold changes in the expression of adjacent genes (C, blue bar). (D) Partial correlations with the local density of essential genes after controlling for the level of H3K4me2 modification are shown for the expression level of the GFP gene (D, red bar) and the median fold change in the expression of adjacent genes (D, blue bar). (A–D) The surrounding regions included five genes upstream and five genes downstream of the GFP integration site. The statistical significance of each correlation is indicated: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . (E) Schematic diagram showing two alternative models for the relationship between the density of essential genes (red background) and the H3K4me2 level (gray background), as well as their (in)direct influence on the GFP expression level. Partial correlation analysis (red bars in C and D) supported the red scenario: the density of essential genes, not the H3K4me2 level, had a more direct effect on the GFP expression level. (F) Similar to (E), except for the determinant of the expression of adjacent genes. Partial correlation analysis (blue bars in C and D) suggested that the H3K4me2 level (blue background), not the density of essential genes (gray background), had a more direct effect on the expression changes of genes surrounding the GFP integration site. (G) Schematic diagram showing correlations between the different factors investigated here, including the proximate influences revealed in (E) and (F) (red and blue arrows, respectively), the known correlation between essential gene density and the H3K4me2 level (green arrow), and the resulting relationship between the expression of GFP and the expression of adjacent genes (black arrow).

that the H3K4me2 signal was positively correlated with the fraction of adjacent genes with upregulated expression, negatively correlated with the fraction of adjacent genes with downregulated expression, and positively correlated with the median fold change in the expression of the adjacent genes (fig. 3B). Analyses of the 3D region of the integration site confirmed the above conclusions (supplementary fig. S12, [Supplementary Material](#) online). These results suggested that the H3K4me2 signal could be used as a marker of the expression of the integrated gene and the expression change of adjacent genes.

The above results confirmed that the expression level of an integrated gene is related to both the aggregation of essential genes (fig. 2) and the presence or absence of a particular histone modification (supplementary fig. S11, [Supplementary Material](#) online). We then used partial

correlation analysis to determine the relative importance of these two factors. If the GFP expression level was not correlated with the H3K4me2 level when the essential gene density was controlled for, then the essential gene density was more likely the proximate cause of the observed expression change after GFP integration (fig. 3E, red background); if, on the other hand, the GFP expression level was not correlated with the essential gene density when the H3K4me2 level was controlled for, then the H3K4me2 level was more likely the proximate cause (fig. 3E, gray background). The results we observed supported the former scenario (fig. 3C, red bar) and did not support the latter (fig. 3D, red bar); that is, the GFP expression level was more directly affected by essential gene density than by H3K4me2 level (fig. 3E and G, red background). These observations suggested that the essential gene density in the neighborhood determined the expression level

of the integrated gene, and this finding was compatible with the model of transcriptional competition underlying position effect externality.

Similarly, to evaluate the relative importance of the aggregation of essential genes and the presence or absence of specific histone modifications in regulating the expression of the adjacent genes, we performed another partial correlation analysis. We found that the expression of the adjacent genes was correlated with the H3K4me2 level when the density of essential genes was controlled for (fig. 3C, blue bar; fig. 3F, gray background), whereas the expression of the adjacent genes was not correlated with the density of essential genes when the H3K4me2 level was controlled for (fig. 3D, blue bar and fig. 3F, blue background). These results indicated that the expression of the neighboring genes was more affected by the H3K4me2 level than by essential gene density (fig. 3F and G, blue background), in contrast to the expression level of the integrated gene. In addition, the H3K4me2 signals were significantly anticorrelated with essential gene density (fig. 3A, gray bar; fig. 3C, green arrow). These results suggested that the native H3K4me2 modification level was presumably evolutionarily optimized to fit the local density of essential genes, thereby creating a correlation between the expression of the integrated gene and that of the adjacent genes (fig. 3C, black arrow).

### Externality Contributes to Fitness Consequences of the Position Effect

To further investigate the contribution of position effect externality to the phenotypic consequences of the position effect, we chose to measure the most important phenotype of yeast, that is, Darwinian fitness, by measuring the growth curve of each constructed strain in YPD medium (see Materials and Methods). Surprisingly, in contrast to a previous study using multiple copies of integrated genes (Dekel and Alon 2005; Kafri et al. 2016), we found a negligible effect of GFP expression on fitness (fig. 4A, see also supplementary fig. S13, [Supplementary Material](#) online, for similar observations based on GFP protein abundance), possibly explainable by the relatively small expression variation among our single-copy GFP genes integrated at different loci. However, we observed a significant negative correlation between the expression changes of the adjacent genes and fitness (fig. 4B), indicating that position effect externality can play a major role in the phenotypic consequences of the position effect.

How would position effect externality impact the evolutionary fate of an integrated gene? Let us consider a scenario where GFP is integrated into a genomic region with a low density of essential genes. On the one hand, GFP would be transcribed at a higher abundance (fig. 2C), which should give rise to a higher total transcriptional yield of GFP (fig. 4C, before 20 h; fig. 4D; see Materials and Methods). On the other hand, position effect externality would dictate that the adjacent genes would likely be upregulated (fig. 2E), thereby decreasing cellular fitness (fig. 4B). As a result, when we estimated the total transcriptional yield of GFP in a population by considering both initial GFP expression and cellular fitness, the strains initially expressing more GFP in essential

gene-depleted regions were predicted to be outperformed by those expressing less GFP in essential gene-rich regions (fig. 4C, after 70 h; see Materials and Methods) because the latter would have a higher cellular growth rate (fitness). This conclusion was also supported by analyses using 3D proximity to the integration site (supplementary fig. S14, [Supplementary Material](#) online).

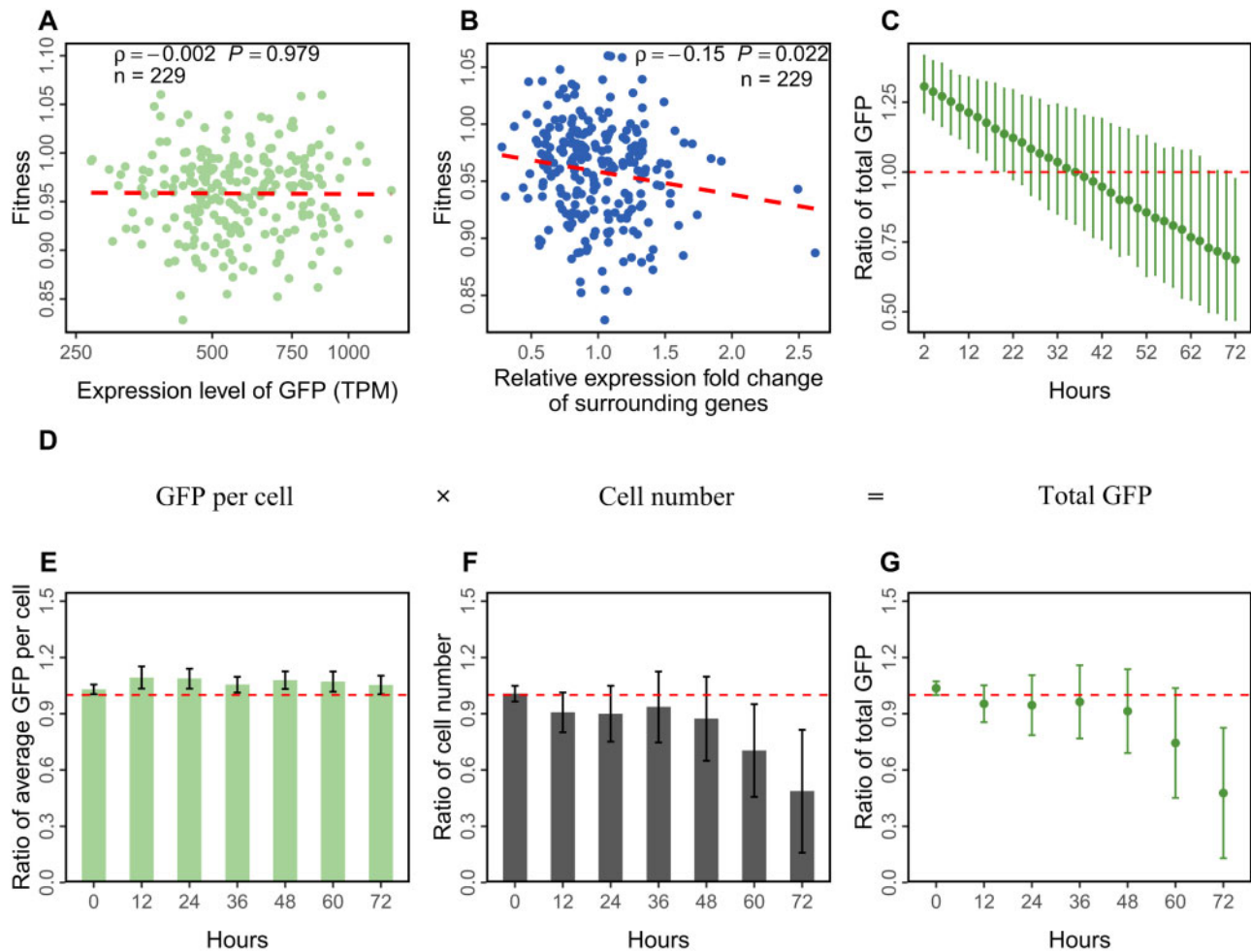
We further confirmed the above model through long-term cultivation of yeast strains and flow cytometry-based measurement of the number of cells and GFP abundance in individual cells (see Materials and Methods). Compared with the strains in which GFP was integrated into a locus with an essential gene density of 60%, the strains with an essential gene density of 10% at the integration site showed a higher GFP protein abundance per cell (fig. 4E), an observation consistent with their mRNA expression levels (fig. 2C). In the early stages of culture (before 12 h), when the two types of strains had similar numbers of cells in their respective populations (fig. 4F), the strains with GFP integrated into essential gene-depleted regions had a higher total yield of GFP protein in a population (fig. 4G). However, after a long period of culture (after 12 h), the strains with GFP integrated into essential gene-depleted regions were gradually outgrown (in terms of the number of cells) by the strains with GFP integrated into essential gene-rich regions (fig. 4F), leading to a lower total yield of GFP protein per population (fig. 4G). At 72 h, the total yield of GFP protein in a population from strains with GFP integrated into essential gene-depleted regions dropped to half that in a population from strains with GFP integrated into essential gene-rich regions (fig. 4G).

Collectively, our observations above revealed a trade-off between the immediate expression of foreign genes integrated into a host genome and their long-term transcriptional yield in a population; this trade-off was caused by the fitness consequences of the integration, which were mediated by the externality of the position effect (fig. 4B) but not by the expression of the focal integrated gene itself (fig. 4A).

### Discussion

In this study, we determined the transcriptome profiles of approximately 250 yeast strains, each with a GFP cassette integrated into a different genomic locus. We found that GFP expression levels were negatively correlated with the local density of essential genes in either linear or 3D proximity to the integration site. An opposite trend in the expression of the neighboring genes around the integration site was also revealed, indicating a previously unappreciated externality of the position effect. Assuming that essential genes are transcriptionally more competitive than nonessential genes, the observed position effects and their externality can be explained by competition among adjacent genes for transcriptional resources. We also found that specific histone modifications were closely related to position effects and their externality. More importantly, the observed externality, but not the expression of the integrated gene, seemed to be one of many factors responsible for the phenotypic consequences of the position effect. Altogether, our results revealed a



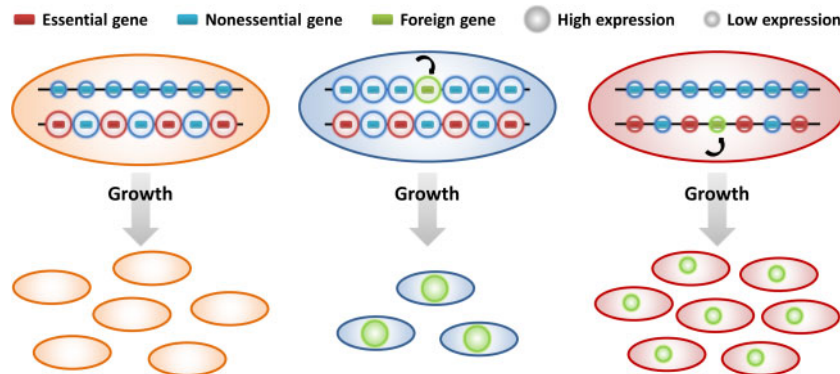


**Fig. 4.** Changes in the expression of adjacent genes, rather than the expression of the GFP gene, are related to fitness in the constructed strains. (A, B) Correlation between the fitness of the constructed strains relative to the wild-type strain and the expression level of the GFP gene (A) and the expression fold change of genes surrounding the integration site (B). The red dotted line represents the fitted linear regression model. (C) Based on fitness modeling, the ratio of the total yield of GFP mRNA in strains with GFP integrated into genomic regions with 10% adjacent essential genes relative to that of GFP mRNA in strains with GFP integrated into genomic regions with 60% adjacent essential genes at different time points of cultivation (from 2 to 72 h). (D) The formula for calculating the y-axis of the C panel. Each component in the formula corresponds to the y-axis of panels E–G. (E–G) Comparing the strains with GFP integrated into genomic regions with 10% adjacent essential genes and those with 60% adjacent essential genes, the ratio of GFP protein abundance per cell (E), the ratio of the number of cells (F), and the ratio of the total yield of GFP protein (G) detected by flow cytometry are shown. (C, E–G) The SD of the ratios was estimated by bootstrapping the genes 1,000 times. The red dotted line represents  $y = 1$ .

previously unappreciated facet of the position effect, which might have a significant impact on synthetic biology, such as genetic engineering aimed at maximizing the transcriptional yield of exogenous genes in a population, and evolutionary biology, such as understanding the evolutionary forces behind gene orders/distributions on chromosomes.

There were potential caveats in our study that warrant discussion. First, yeast contains more than 5,000 verified genes, but only approximately 250 gene loci were tested in our analyses. Although our sampled loci were likely unbiased (supplementary fig. S15, [Supplementary Material](#) online), future large-scale studies covering additional loci should be carried out to examine the generality of our conclusion. We wish to emphasize again that our data were subjected to multiple quality control steps, which should have minimized the regulatory effects exerted by any feedback mechanism; at the

same time, however, these measures may also have reduced the full range of the position effect and its externality. Therefore, any evaluation of effect sizes here should be considered an underestimation. Second, although our experiments have shown that the transcriptional competition caused by the integration of foreign genes is not promoter-specific, the degrees of transcriptional competition caused by different promoters are likely different. Third, it should be reiterated that the externality of the position effects we measured includes the effect of the heterozygous deletion of the endogenous gene. Nevertheless, the deletion effect is likely negligible for our conclusion, as we confirmed that the expression levels of neighboring genes were generally unaltered by heterozygous deletion ([fig. 1G](#)) and because all the mutants grew on YPD, supporting the deleted genes as being haplosufficient for viability ([Deutschbauer et al. 2005](#)). More



**FIG. 5.** Schematic diagram of the externality of the position effect and its influence on the total yield of an integrated foreign gene. Two genomic loci within a cell are shown: One is depleted of essential genes, and the other is enriched in essential genes. Essential genes, nonessential genes, and foreign genes are indicated by red, blue, and green segments, respectively. The expression levels of the genes are indicated by the sizes of the circles around them. Integration of the foreign gene into a locus depleted of essential genes (the blue cell) will lead to relatively high expression of the foreign gene itself and elevated expression of the surrounding genes. However, the expression elevation of the surrounding genes is harmful to the cell, slowing cellular growth relative to that of the wild-type strain (the orange cell). In contrast, integration of the foreign gene into a locus enriched in essential genes (the red cell) will lead to reduced expression of both the foreign gene and the surrounding genes, but this effect leads to a growth advantage for the cell. As a result, given enough time, faster cellular growth in the red cell can compensate for the lowered transcriptional output of the foreign gene per cell, thereby giving rise to a greater total transcriptional output of the foreign gene.

importantly, gene integration should change the local composition of the wild-type genome regardless of whether it occurs in the gene locus. It is this very change in composition at different positions that cause the difference in the expression level of adjacent genes. Fourth, the expression changes of adjacent genes could also be explained by a disruption of local regulation. This is because essential genes are usually old genes with strong expression enabled by a high density of activation marks and transcription factor binding (Zhang et al. 2012; Zhang and Zhou 2019), a state that is likely vulnerable to nearby integration of foreign genes. We nevertheless think this possibility is less likely than our model of transcriptional competition because 1) integration of a single foreign gene has been found to be nonperturbative for the chromatin landscapes in yeast (Chen et al. 2013) and 2) our data suggested the possibility that adjacent genes sometimes tend to be upregulated instead of downregulated (e.g., the first and third groups in supplementary fig. S5B, Supplementary Material online), which is in contrast to the prediction of disruption of the regulatory landscape. Fifth, gene essentiality has been shown to depend on environmental and genetic contexts (Larrimore and Rancati 2019). Therefore, the generality of the results of local regulation found in this study needs further verification using transcriptomes of deletion strains from different genetic backgrounds or growing in different environments. Sixth, our research context is the integration of an exogenous gene into different gene loci in the genome, that is, replacement of an endogenous gene by an exogenous gene. This scenario might be different from gene insertion, which retains the original genomic sequence and could occur during gene translocation and transposon integration. The translocated gene can then be fused with a neighboring gene or sequence (Dougherty et al. 2018). The newly inserted transposon will be suppressed

by small RNA, and this suppression can also spread to adjacent regions (Eickbush and Eickbush 2015). Therefore, the potential insights of endogenous gene translocation and transposon integration provided by our findings require more in-depth research.

The results of our study highlighted how the evolutionary fate of an exogenous gene integrated into the host genome will be affected by the density of essential genes near the integration site. On the one hand, if the integration event occurred at a locus with a high density of essential genes, the expression of the integrated gene and the neighboring genes might be lowered due to the strongly competitive transcriptional environment created by the adjacent essential genes; meanwhile, the cellular fitness would not be strongly influenced. On the other hand, if integration occurred at a locus with a low density of essential genes, the integrated gene and its neighbors might exhibit high expression, but the cellular fitness would be significantly decreased. Most importantly, the changes in the expression of neighboring genes would have a nonnegligible effect on the fitness impact of the position effect, a novel phenomenon here termed the externality of the position effect (fig. 5).

Our results also bear important implications for other types of molecular events that could be generally characterized as gene integration. For example, one common type of exogenous gene integration that occurs naturally is the integration of viral genes into the host genome. Previous studies have mostly focused on only the direct functional consequences of this integration (Ciuffi 2016; Chen et al. 2017), such as intergenic integration destroying *cis*-regulatory elements, intragenic integration altering transcription or even endogenous gene structure. Our study suggested that integration of a transcriptionally active gene can impact genes near the integration site, which would likely further influence host cellular fitness.

## Materials and Methods

### Yeast Cassette Construction

We replaced the kanMX module in the heterozygous deletion strains of yeast (*S. cerevisiae*) at the IOC3, STE12, and YTH1 loci with an expression cassette comprising the marker gene URA3 and a GFP gene driven by three promoters: *pTDH3* (*pTDH3-GFP*; *TDH3*, Glyceraldehyde-3-phosphate dehydrogenase), *pARF1* (*pARF1-GFP*; *ARF1*, ADP-ribosylation factor), and *pTYS1* (*pTYS1-GFP*; *TYS1*, Cytoplasmic tyrosyl-tRNA synthetase). The *pTDH3-GFP-URA3*, *pARF1-GFP-URA3*, and *pTYS1-GFP-URA3* cassettes were obtained as described in our previous study (Chen and Zhang 2016) and amplified with homologous recombination primers corresponding to the loci IOC3, STE12, and YTH1 (supplementary table S4, [Supplementary Material](#) online).

Subsequently, yeast transformations were carried out using a previously published protocol (Gietz and Schiestl 2007) with some adjustments. Specifically, the cells of the corresponding heterozygous deletion strains were cultured at 30 °C in 5 ml of YPD medium (1% yeast extract, 2% peptone, 2% glucose) overnight until saturation. Then, the cells were diluted to an OD<sub>660</sub> of 0.2 and grown for approximately 4 h until the OD<sub>660</sub> reached 0.7. Each culture was harvested by centrifugation at 2,000 × g for 5 min and was used to prepare competent cells with 0.1 M lithium acetate (LiAc, Sigma). Subsequently, 240 μl of polyethylene glycol (50% w/v, Sigma), 30 μl of LiAc (1 M), 30 μl of water, 10 μl of salmon sperm vector DNA (10 mg/ml, Sigma), 5 μg of DNA product, and 50 μl of competent cells were put into a tube and vortexed for 1 min. The above mixture was heat-shocked at 42 °C for 30 min. After being washed once in water, each mixture was spread on synthetic complete medium plates without uracil (SC-ura) and cultured for 2–3 days. Transformants of individual colonies were selected, and correct replacement was confirmed by PCR (supplementary table S4, [Supplementary Material](#) online).

In addition, we constructed a *pGAL1-GFP-Leu* cassette. The GAL1 promoter (phosphorylates alpha-D-galactose to alpha-D-galactose-1-phosphate in the first step of galactose catabolism), the GFP gene and the leucine (Leu) marker gene extracted from the plasmid PYES2 (Invitrogen), the *pTDH3-GFP-URA3* cassette and yeast strain S288C were amplified and then fused together (supplementary table S4, [Supplementary Material](#) online). With the transformation protocol described above, this cassette was integrated into the loci of IOC3, STE12, and YTH1 of yeast strain BY4743. Finally, transformants of individual colonies were selected, and correct replacement was confirmed by PCR (supplementary table S4, [Supplementary Material](#) online).

### RNA Extraction and Sequencing

Each strain of yeast was inoculated into 5 ml of YPD medium and then cultured overnight at 30 °C and 250 rpm. The saturated culture was then returned to OD<sub>660</sub> = 0.2 in 4 ml of YPD, and growth continued at 30 °C until OD<sub>660</sub> = 0.65–0.75. For the constructed strains with the *pGAL1-GFP* cassette, we used 3 ml of YPR (1% yeast extract, 2% peptone, and

2% raffinose) diluted to an OD<sub>660</sub> of approximately 0.1 and then cultured them for 24 h at 30 °C and 250 rpm to perform starvation treatment. The cultures were then diluted again with 4 ml of fresh YPG medium (1% yeast extract, 2% peptone, and 2% galactose) to an OD<sub>660</sub> of approximately 0.1 and incubated at 30 °C and 250 rpm for approximately 3.5 h until the OD<sub>660</sub> was between 0.65 and 0.75.

Total RNA was then extracted from cell lysates using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. The quality and concentration of RNA were determined by a NanoDrop instrument. An A<sub>260</sub>/A<sub>230</sub> ratio > 2 and an A<sub>260</sub>/A<sub>280</sub> ratio in the range of 1.8–2.2 were considered acceptable. Finally, the samples of pRPL5-GFP strains were sequenced on a HiSeq 4000 platform (Illumina) in paired-end 150 bp mode (supplementary table S1, [Supplementary Material](#) online). In addition, we used the PrimeScript RT Reagent Kit with gDNA Eraser (TAKARA) to reverse-transcribe 1 μg of total RNA into cDNA according to the manufacturer's instructions.

### Calculation of RNA Abundance

The yeast S288C reference genome, version R64-2-1, and corresponding genome annotation were obtained from the Saccharomyces Genome Database (SGD) (Cherry, et al. 2012). To estimate the RNA abundance in each strain, we mapped the adaptor-trimmed and quality-filtered (Bolger et al. 2014) short reads to the reference genome with HISAT2 (Kim et al. 2019). Then, transcripts per million reads (TPM) values (Wagner et al. 2012) were estimated by StringTie (Pertea et al. 2015) based on the mapping results and used as gene expression levels (supplementary table S2, [Supplementary Material](#) online). To avoid division by zero, genes with TPM values of 0 were assigned values of one-tenth of the minimum nonzero TPM value in all the yeast transcriptional profiles.

### RT-qPCR Primer Design and Measurement

We searched the cDNA sequences of 30 related genes and the control gene ACT1 (Actin) in the S288C reference genome, version R64-2-1, and used NCBI Primer Blast to design RT-qPCR primers (supplementary table S4, [Supplementary Material](#) online). The RT-qPCR products were limited to 100–200 bp, and the melting temperature of each RT-qPCR primer was between 58 °C and 62 °C. The concentration of cDNA (0.2 μl) was then measured by RT-qPCR with 10 μM forward and reverse primers in 96-well plates using iTaq Universal SYBR Green Supermix (Bio-Rad) on a LightCycler 96 real-time PCR System (Roche). The cycling parameters for amplification were 95 °C for 30 s and 40 cycles of 95 °C for 5 s and 60 °C for 30 s. The signals were normalized to that of Actin and quantified by the  $\Delta\Delta C_t$  method (Livak and Schmittgen 2001). The resulting expression levels are presented as the mean ± SD of four independent experiments, each performed in triplicate.

### Determination of Linear Gene Clusters

According to the yeast S288C genome annotation, 5,108 verified genes were selected for subsequent data analysis. Fitness

data in rich medium were obtained from a previous report (Winzeler et al. 1999). First, we grouped the yeast gene components into overlapping windows of a specific number of consecutive genes, with a step size of 2 and a window size of 2, 4, 6, . . . or 40. Subsequently, we counted the fraction (density) of essential genes in each window, the median expression level of the genes in this window in the wild-type strains, and the fraction of the genes that were upregulated (greater by >1.5-fold or 2-fold) and downregulated (less by <66.7% or 50%), as well as the median change in gene expression within that window after integration of the GFP gene. In particular, when the window size was equal to ten, the loci where the essential gene density was 0%, 10%, 20%, 30%, 40%, 50%, 60%, and 70% were associated with 27, 38, 44, 40, 33, 25, 26, and 7 constructed strains, respectively.

Similarly, we set overlapping windows of a specific number of base pairs (the step size was equal to 5 kb, and the window sizes were set to 10, 15, 20, . . . , or 80 kb) in the yeast genome to calculate the above properties again.

### Determination of 3D Gene Clusters

We used the haploid yeast 3D chromosomal architecture that was inferred through chromosome conformation capture-on-chip (4C) coupled with massively parallel sequencing (Duan et al. 2010). Notably, the 3D model of yeast chromosomes that we used was measured in the haploid strain rather than in the diploid strain; however, dramatic differences between the 3D genomes of haploid and diploid cells are unlikely (Tan et al. 2018). A file containing a list of chromosomal interactions identified from *HindIII* libraries was downloaded from the original report to infer the spatial distances between gene pairs. Each gene window was defined as a fixed number of genes (set to 2, 4, 6, . . . , or 40) that were closest (with the highest interaction probability) to the focal gene. Similar to the analysis at the linear level, we calculated the density of essential genes, the median expression level of the genes in the wild-type strain, the fraction of genes that were upregulated and downregulated, and the median change in gene expression within each window after integration of the GFP gene. In particular, when the window size was equal to 10, the loci where the essential gene density was 0%, 10%, 20%, 30%, 40%, 50%, 60%, and 70% were associated with 23, 41, 57, 42, 31, 22, 17, and 7 constructed strains, respectively.

In addition, we set the distance from the GFP integration site to 1, 2, 3, . . . , or 15 nm as the window size to calculate the above properties again.

### Calculation of Histone Modification Levels

We downloaded high-throughput sequencing data for eight types of histone acetylation and histone methylation (H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3, H3K9ac, H3K12ac, and H3K14ac) in wild-type yeast from the Sequence Read Archive (supplementary table S1, Supplementary Material online). By applying a pipeline similar to the one we used to quantify RNA abundance, we calculated the histone modification levels (i.e., abundance of short

reads from high-throughput sequencing targeting specific modifications) of the 200-bp region starting 200 bp upstream of the start codon of each endogenous gene (supplementary table S3, Supplementary Material online). We then calculated the average level of histone modification in the 10 gene windows at the linear level and the 3D level.

### Fitness Measurement

To measure the growth rates of the constructed *GFP* strains, we cultured the cells in YPD medium at 30 °C overnight, and then, 5 µl of the saturated culture was transferred to 145 µl of YPD in 96-well plates. Each 96-well plate was shaken on an Epoch 2 microplate reader (BioTek) at 30 °C for 12 h, and OD600 readings were taken every 10 min. After repeating this experiment at least three times for each strain, we calculated the doubling time for each strain according to Murakami and Kaeberlein (2009) (supplementary table S5, Supplementary Material online). Based on a comparison of the doubling times of the constructed strain ( $t_{cs}$ ) and the wild-type strain ( $t_{wt}$ ), the relative fitness ( $w$ ) of each *GFP* strain was calculated as follows:

$$w = \frac{t_{wt}}{t_{cs}}$$

Additionally, fitness was used to estimate the relative total yield of GFP mRNA in the strains with GFP integrated into loci with 10% essential genes compared with those with 60% essential genes by the following equation:

$$\begin{aligned} \frac{\text{TPM}_{10\%} \times 2^{t/t_{10\%}}}{\text{TPM}_{60\%} \times 2^{t/t_{60\%}}} &= \frac{\text{TPM}_{10\%} \times 2^{t/\left(\frac{t_{wt}}{w_{10\%}}\right)}}{\text{TPM}_{60\%} \times 2^{t/\left(\frac{t_{wt}}{w_{60\%}}\right)}} \\ &= \frac{\text{TPM}_{10\%}}{\text{TPM}_{60\%}} 2^{t(w_{10\%} - w_{60\%})/1.5}, \end{aligned}$$

where TPM is the mRNA expression level of GFP inferred from RNA-seq,  $t$  is culture time (in hours), and  $w$  is fitness.  $2^{t/t_{10\%}}$  and  $2^{t/t_{60\%}}$  represent the fold increase in the number of cells after  $t$  hours of culture for strains with GFP integrated into loci where the densities of essential genes were 10% and 60%, respectively. The doubling time of the wild-type strain ( $t_{wt}$ ) was approximately 1.5 h (supplementary table S5, Supplementary Material online).

### Experimental Determination of Total GFP Protein Yield

Five yeast strains with GFP integrated into loci with an essential gene density of 10% and five yeast strains with GFP integrated into loci with an essential gene density of 60% were randomly selected and subjected to continuous culture experiments to estimate the total yield of GFP protein. The resuscitated cells of two biological replicates of each strain were aspirated into YPD and cultured at 30 °C for 72 h. During this process,  $1 \times 10^7$  cells in each sample were transferred to a new 700 ml of YPD every 12 h to ensure that the density of the transferred cells did not affect the growth rate comparison (Ferrezuelo et al. 2012). At the same time, the cell

density was measured with an ultraviolet spectrophotometer to ensure that each sample was in the exponential growth phase (Ginovart et al. 2017). In addition, an equal volume of cell culture medium (~10,000 cells) was taken from each sample to record the abundance of GFP protein in each cell and the number of test cells by an Attune N × T flow cytometer (Life Technologies) with a 533/30 nm optical filter for GFP acquisition. After a set of cytometric events considered to be single fluorescing cells were filtered for each sample, cells were gated based on cell size and shape (forward and side scatter pulse area, FSC-A and SSC-A). Then, the GFP abundance (BL1-A) of each cell in each well was recorded.

The ratio of the number of cells at each time point ( $R_C$ ) of the strains in which GFP was integrated into essential gene-depleted regions compared with the strains in which GFP was integrated into essential gene-enriched regions was calculated by the following formula:

$$R_C = \frac{\sum_{i=1}^5 C_{i0} \times C_{i12} \times \cdots \times C_{it}}{\sum_{j=1}^5 C_{j0} \times C_{j12} \times \cdots \times C_{jt}},$$

where  $C_i$  is the number of cells detected by the Attune NxT flow cytometer in strain  $i$  (with GFP integrated into a locus with an essential gene density of 10%),  $C_j$  is the number of cells in strain  $j$  (with GFP integrated into a locus with an essential gene density of 60%), and  $t$  is the culture time (0, 12, 24, 36, 48, 60, and 72 h).

The ratio of the total yield of GFP protein at each time point ( $R_G$ ) of the strains in which GFP was integrated into essential gene-depleted regions compared with the strains in which GFP was integrated into essential gene-enriched regions was calculated by the following formula:

$$R_G = \frac{\sum_{i=1}^5 C_{i0} \times C_{i12} \times \cdots \times C_{it} \times G_{it}}{\sum_{j=1}^5 C_{j0} \times C_{j12} \times \cdots \times C_{jt} \times G_{jt}},$$

where  $G_{it}$  is the average GFP abundance detected by flow cytometry in strain  $i$  (with GFP integrated into a locus with an essential gene density of 10%) at time  $t$  and  $G_{jt}$  is the average GFP abundance detected in strain  $j$  (with GFP integrated into a locus with an essential gene density of 60%) at time  $t$ .

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We thank Jian-Rong Yang, Feng Chen, and Xiaoyu Zhang for helpful discussions and the anonymous reviewers for their insightful and helpful comments on this work. The overall research project was supported by grants from the National Key R&D Program of China (project number

2017YFA0103504, awarded to X.C. and project number 2018ZX10301402, awarded to Z.H.), the National Special Research Program of China for Important Infectious Diseases (project number 2018ZX10302103, awarded to X.C.), and the National Natural Science Foundation of China (project number 31771406, awarded to X.C.).

## Data Availability

The data generated in this study are available from NCBI GEO under accession number GSE142839.

## References

- Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LF, van Lohuizen M, van Steensel B. 2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* 154(4):914–927.
- Artieri CG, Fraser HB. 2014. Evolution at two levels of gene expression in yeast. *Genome Res.* 24(3):411–421.
- Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet.* 39(8):945–949.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
- Bucci MD, Weisenhorn E, Haws S, Yao Z, Zimmerman G, Gannon M, Taggart J, Lee T, Klionsky DJ, Russell J, et al. 2018. An autophagy-independent role for atg41 in sulfur metabolism during zinc deficiency. *Genetics* 208(3):1115–1130.
- Chen H-C, Martinez JP, Zorita E, Meyerhans A, Filion GJ. 2017. Position effects influence HIV latency reversal. *Nat Struct Mol Biol.* 24(1):47–54.
- Chen M, Licon K, Otsuka R, Pillus L, Ideker T. 2013. Decoupling epigenetic and genetic effects through systematic analysis of gene position. *Cell Rep.* 3(1):128–137.
- Chen X, Zhang J. 2016. The genomic landscape of position effects on protein expression level and noise in yeast. *Cell Syst.* 2(5):347–354.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40(Database issue):D700–D705.
- Ciuffi A. 2016. The benefits of integration. *Clin Microbiol Infect.* 22(4):324–332.
- Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436(7050):588–592.
- Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 169(4):1915–1925.
- Dey SS, Foley JE, Limsirichai P, Schaffer DV, Arkin AP. 2015. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Mol Syst Biol.* 11(5):806.
- Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* 28(10):1566–1576.
- Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* 465(7296):363–367.
- Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol.* 10(9):1106–1113.
- Eickbush TH, Eickbush DG. 2015. Integration, regulation, and long-term stability of R2 retrotransposons. *Mobile DNA.* III:1125–1146.
- Ferrezuelo F, Colomina N, Palmisano A, Gari E, Gallego C, Csikasz-Nagy A, Aldea M. 2012. The critical size is set at a single-cell level by growth rate to attain homeostasis and adaptation. *Nat Commun.* 3:1012.
- Feuerborn A, Cook PR. 2015. Why the activity of a gene depends on its neighbors. *Trends Genet.* 31(9):483–490.

- Ghanbarian AT, Hurst LD. 2015. Neighboring genes show correlated evolution in gene expression. *Mol Biol Evol.* 32(7):1748–1766.
- Gietz RD, Schiestl RH. 2007. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc.* 2(1):35–37.
- Ginovart M, Carbo R, Blanco M, Portell X. 2017. Digital image analysis of yeast single cells growing in two different oxygen concentrations to analyze the population growth and to assist individual-based modeling. *Front Microbiol.* 8:2628.
- Grewal SI, Jia S. 2007. Heterochromatin revisited. *Nat Rev Genet.* 8(1):35–46.
- Harvey ZH, Chakravarty AK, Futia RA, Jarosz DF. 2020. A prion epigenetic switch establishes an active chromatin state. *Cell* 180(5):928–940.e14.
- Ivics Z, Li MA, Mates L, Boeke JD, Nagy A, Bradley A, Izsvak Z. 2009. Transposon-mediated genome manipulation in vertebrates. *Nat Methods.* 6(6):415–422.
- Jackson DA, Hassan AB, Errington RJ, Cook PR. 1993. Visualization of focal sites of transcription within human nuclei. *EMBO J.* 12(3):1059–1065.
- Kafri M, Metzl-Raz E, Jona G, Barkai N. 2016. The cost of protein production. *Cell Rep.* 14(1):22–31.
- Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet.* 9(8):605–618.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 37(8):907–915.
- Kleinjan DJ, van Heyningen V. 1998. Position effect in human genetic disease. *Hum Mol Genet.* 7(10):1611–1618.
- Larrimore KE, Rancati G. 2019. The conditional nature of gene essentiality. *Curr Opin Genet Dev.* 58–59:55–61.
- Liu J, Shively CA, Mitra RD. 2020. Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. *Nucleic Acids Res.* 48(9):e50.
- Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the  $2^{-\Delta\Delta CT}$  method. *Methods* 25(4):402–408.
- Milot E, Fraser P, Grosveld F. 1996. Position effects and genetic disease. *Trends Genet.* 12(4):123–126.
- Murakami C, Kaeberlein M. 2009. Quantifying yeast chronological life span by outgrowth of aged cells. *J Vis Exp.* doi: 10.3791/1156
- Pal C, Hurst LD. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet.* 33(3):392–395.
- Pande A, Brosius J, Makalowska I, Makalowski W, Raabe CA. 2018. Transcriptional interference by small transcripts in proximal promoter regions. *Nucleic Acids Res.* 46(3):1069–1088.
- Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 33(3):290–295.
- Silveira MA, Bilodeau S. 2018. Defining the transcriptional ecosystem. *Mol Cell.* 72(6):920–924.
- Soares LM, He PC, Chun Y, Suh H, Kim T, Buratowski S. 2017. Determinants of histone H3K4 methylation patterns. *Mol Cell.* 68(4):773–785.e776.
- Strainic MC, Sullivan JJ, Collado-Vides J, deHaseth PL. 2000. Promoter interference in a bacteriophage lambda control region: effects of a range of interpromoter distances. *J Bacteriol.* 182(1):216–220.
- Sturtevant AH. 1925. The effects of unequal crossing over at the bar locus in *Drosophila*. *Genetics* 10(2):117–147.
- Tan L, Xing D, Chang CH, Li H, Xie XS. 2018. Three-dimensional genome structures of single diploid human cells. *Science* 361(6405):924–928.
- Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131(4):281–285.
- Wang Z, Zhang J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc Natl Acad Sci USA.* 108(16):E67–E76.
- Wilson C, Bellen HJ, Gehring WJ. 1990. Position effects on eukaryotic gene expression. *Annu Rev Cell Biol.* 6:679–714.
- Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285(5429):901–906.
- Yang Y-F, Cao W, Wu S, Qian W. 2017. Genetic interaction network as an important determinant of gene order in genome evolution. *Mol Biol Evol.* 34(12):3254–3266.
- Yant SR, Wu X, Huang Y, Garrison B, Burgess SM, Kay MA. 2005. High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell Biol.* 25(6):2085–2094.
- Zhang JY, Zhou Q. 2019. On the regulatory evolution of new genes throughout their life history. *Mol Biol Evol.* 36(1):15–27.
- Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays* 34(11):982–991.