Allelic Transcription Factor binding shape transcriptional kinetics in human cell lines

Bowen Jin[1], Hao Feng[3], William S. Bush[2,3*]

[1]Graduate Program in Systems Biology and Bioinformatics, Department of Nutrition, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA.

[2]Cleveland Institute for Computational Biology, Department for Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA.

[3]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA.

*To whom correspondence should be addressed. Email: wsb36@case.edu

Abstract

Gene expression from bulk RNA-seq studies is an average measurement between two chromosomes and across cell populations. Both allelic and cell-to-cell heterogeneity in gene expression results from promoter bursting patterns that repeatedly alternate between an activated and inactivated state. Increased cell-to-cell heterogeneity in gene expression has been shown as a hallmark of aging, which is potentially induced by the change of promoter bursting patterns. Despite their importance in humans, bursting kinetics studies have been studied only in a limited number of genes within selected model organisms due to technical restrictions in measuring multiple transcript levels over time. Here, we construct a transcriptomic kinetics map using single-cell RNA-seq (scRNA-seq) data and systematically investigate the regulatory effect of genetic variants and transcription factor (TF) binding on transcriptional kinetics. We obtain allelic level expression from scRNA-seq data with phased genotypes for two clonal cell lines and estimate transcriptional bursting kinetics for a single chromosome. We found that among transcriptional kinetics, the transcription initiation rate and burst frequency correlate most with eQTL effect sizes from bulk RNA-seq studies, suggesting that eQTLs affect average gene expression mainly through altering the transcription initiation rate and burst frequency. Notably, eQTL studies focused on mean expression changes cannot identify scenarios where burst size and frequency are regulated in opposite directions, which is common amongst genes from lymphoblastoid cell lines. We further found that ~90% of the variance of burst frequency can be explained by TF occupancy within the core promoter and allele-specific binding events of individual TFs are typically associated with the change of transcription initiation rate and burst frequency. Finally, we demonstrate how a genetic variant alters TF binding, resulting in changes to promoter burst kinetics of *HLA-DQA1* to influence multiple hematological traits.

Introduction

Episodic bursting of expression is a prevalent phenomenon in eukaryotic cells (Lammers et al. 2020) compared to prokaryotic cells, which reflects their relative complexity and hierarchical nature of transcriptional regulation. Sandberg et al. (2019) modeled bursting kinetics at each allele independently with scRNA-seq and showed that core promoter elements, including TATA box and initiators, affect bursting kinetics in primary mouse fibroblasts. Such transcriptional bursts can introduce significant cell-to-cell transcriptional variability over a cell population and protein level variability, which can result in phenotypical differences in tissue or organism level. For example, prominent cell-to-cell transcriptional variability has been shown as a hallmark feature of aging (Martinez-Jimenez et al. 2017). Early disturbance in the bursting expression pattern establishes an initial population heterogeneity that allows the selection and propagation of cell-type-specific gene expression (Ohnishi et al. 2014). Hence, elucidating the transcriptional burst provides us with detailed molecular mechanisms of transcription, precise prediction of drug effects, and robust ways to characterize different cell types (Rodriguez and Larson 2020).

The study of episodic bursting expression was limited to a few cell types among selected genes, largely due to technical restrictions. The direct measurement of bursting kinetics depends on single molecule techniques in live cells, such as MS2 tagging and Cas-derived systems, which allow real-time observation of nascent mRNA measurement (Larson et al. 2011; George et al. 2018). However, such techniques can only simultaneously assay tens to hundreds of genes and are limited to cultured cells and tissues. In addition, despite multiple studies that have investigated how gene regulatory factors (transcriptional factors, histone acetylation, DNA methylation, etc.) modulate transcriptional burst, the results are inconsistent among different cell lines and species, probably due to both cell type-specific effects and technical variation (Sánchez and Kondev 2008; Singh et al. 2010; Dar et al. 2012; Fukaya et al. 2016; Faure et al. 2017; Nicolas et al. 2018; Li et al. 2018; Larsson et al. 2019; Bartman et al. 2019; Dobrinic et al. 2021; Gupta et al. 2022). Therefore, a transcriptome-wide and unbiased investigation of how regulatory factors modulate transcriptional bursts in human cells is needed to understand how it influences gene expression at the cell population level and further propagates to tissue or organism-level phenotypes.

Literature exploring allelic bursting kinetics in human cells and their regulatory mechanisms are sparse. One reason is that phased genotypes obtained from familial inference, deep sequencing, or haplotype phasing are not as comprehensive as those from inbred strains of laboratory animals. Therefore in this study, we design a pipeline to characterize bursting kinetics from scRNA-seq data and phased genotype profiles with multiple filtering and validation procedures. scRNA-seq quantifies the mRNA molecules at the single-cell level and reveals complex cellular heterogeneity and dissects cell types. In the study of bursting kinetics, scRNA-seq has unique advantages that allow a snapshot of each cell within the population, capturing the distribution of gene expression over time resulting from the transcriptional burst. Therefore, we can estimate allele-level gene expression by quantifying scRNA-seq reads containing heterozygous variants (i.e. allele-specific expression[ASE]), enabling unbiased measurement distinguishing the bursting kinetics of the two separate promoters present within one subject.

In brief, we first achieve an allele-level single-cell gene expression profile using allele-specific read mapping and inference. We then derive transcriptional kinetics by fitting the allele level expression profile with a two-state model. The two-state model (see Methods) is a stochastic model that represents the episodic bursting feature for gene expression. The estimated kinetics parameters are transcription initiation rate (k+), transcription termination rate (k-), and mRNA transcription rate (r), based on which we can derive bursting kinetics including burst size and frequency. We want to emphasize that the terms "allele-specific expression" and "allelic transcript" were used in this study only to indicate the different observed transcript levels between the two alleles and do not imply any causal relationship between the allele and its expression.

To investigate how regulatory elements modulate transcriptional kinetics, we derive the allelic occupancy profile of 152 TFs, 11 histone markers, and allelic chromatin accessibility profile from ATAC-seq data on GM12878. We hypothesize that transcriptional kinetics are modulated by allelic-specific TF binding or histone modification. We first estimate the variance of transcriptional kinetics explained by the TF binding within the core promoter region and within the distal enhancer region, respectively. Next, we investigate how individual TF and histone markers alter transcriptional kinetics across the genome. Finally, we investigate if eQTLs are associated with the change of transcriptional kinetics due to allele-specific binding (ASB) of TF or allelic open chromatin.

Results

Transcriptomic mapping of transcriptional kinetics for lymphoblastoid cell lines of European and African ancestry

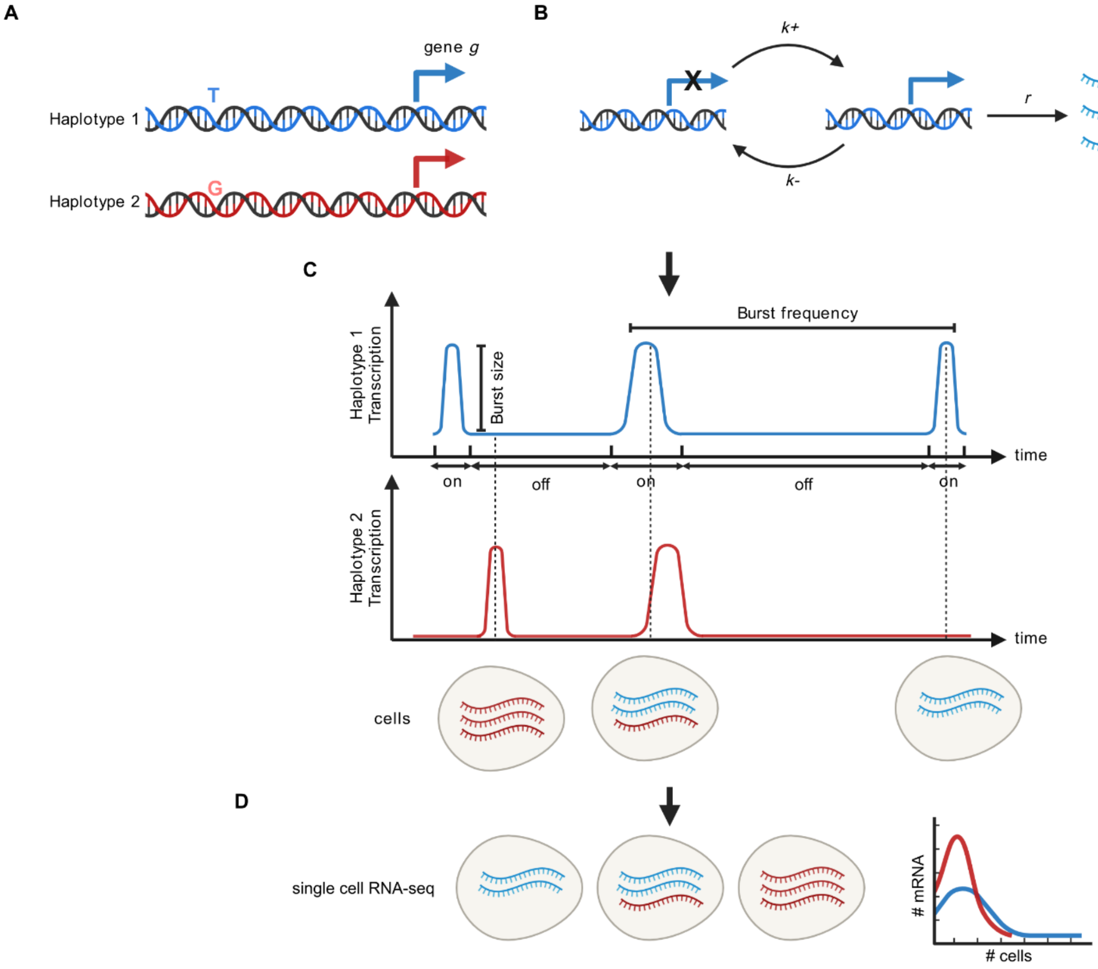


Figure 1. Schema of the allelic transcriptional burst. (A) Gene *g* has two promoters independently transcribing mRNA. (B) Take the first allele as an example. The promoter over time will switch between an on and off state with rates of k+ and k-. In the on state, a promoter can transcribe mRNA with rate *r* and the mRNA will be degraded with rate *d*. (C) The transcriptional burst reflected in single cells are the observed reads from alleles in phase with T or G (colored red and blue, respectively). The transcriptional burst in a time-series experiment will be observed as episodic bursts of the transcript followed by a silenced period. (D) In single-cell studies, promoters with different bursting expression patterns result in different distribution patterns between two promoters.

We derived transcriptional kinetics for GM12878 and GM18502, which are lymphoblastoid cell lines (LCLs) from a female of European descent and a female of African descent, respectively. We obtained allele level expression profiles by combining scRNA-seq data and phased genotype profiles for both samples and fitting them with a beta-binomial distribution (see Methods). Even though the maternal and paternal origin of chromosomes are available for GM12878, we do not have such information for GM18502 and most experiments. Therefore, for simplicity, we assign alleles to Haplotype1and the counterpart to Haplotype2 for the context of all analyses. The associated features for Haplotype1and 2 are color-coded as blue and red, respectively.

Despite the different continental ancestry of the two individuals, the single-cell allelic expression for these two LCLs is highly consistent (Figure S1A, E). For GM12878, 972 of 1,020 genes with an average probability of observing a transcript from Haplotype1 around 0.5 are actually from a bimodal distribution (Figure S1B). Similarly, for GM18502, the majority of the genes (1,463 out of 1,542) expressed with bimodal distributions (Figure S1F). Genes with bimodal distributions are expressed preferably with one of the two alleles within a single cell, which is different from genes with unimodal distribution expressing both alleles within a single cell; even though both distributions can result in biallelic expression in population average.

In addition to the highly consistent allelic expression, GM12878 and GM18502 have highly overlapping allelic transcriptional kinetics. The transcription initiation rate (k+) and transcription termination rate (k-) are not independent of each other and occupy restricted phase space (Figure 2A, E). The average active fraction $\frac{k_+}{(k_+ + k_-)}$ is bound between 0 to 0.6 and most observations are smaller than 0.5, indicating that k+ is smaller than k- (Figure 2B, F). The burst size and burst frequency are empirically independent bursting kinetics derived from k+, k-, and r (Figure 2C, G) The histogram of burst size and burst frequency for GM12878 and GM18502 shows a similar distribution and boundary to those derived from mouse primary fibroblasts (Larsspm et al., 2019) (Supplemental Figure S2). The consistent allelic expression and transcriptional kinetics in GM12878 and GM18502 indicate that they express a set of genes with very similar bursting patterns. While we only have results from two LCLs, we reason that such consistency might be cell-type specific.

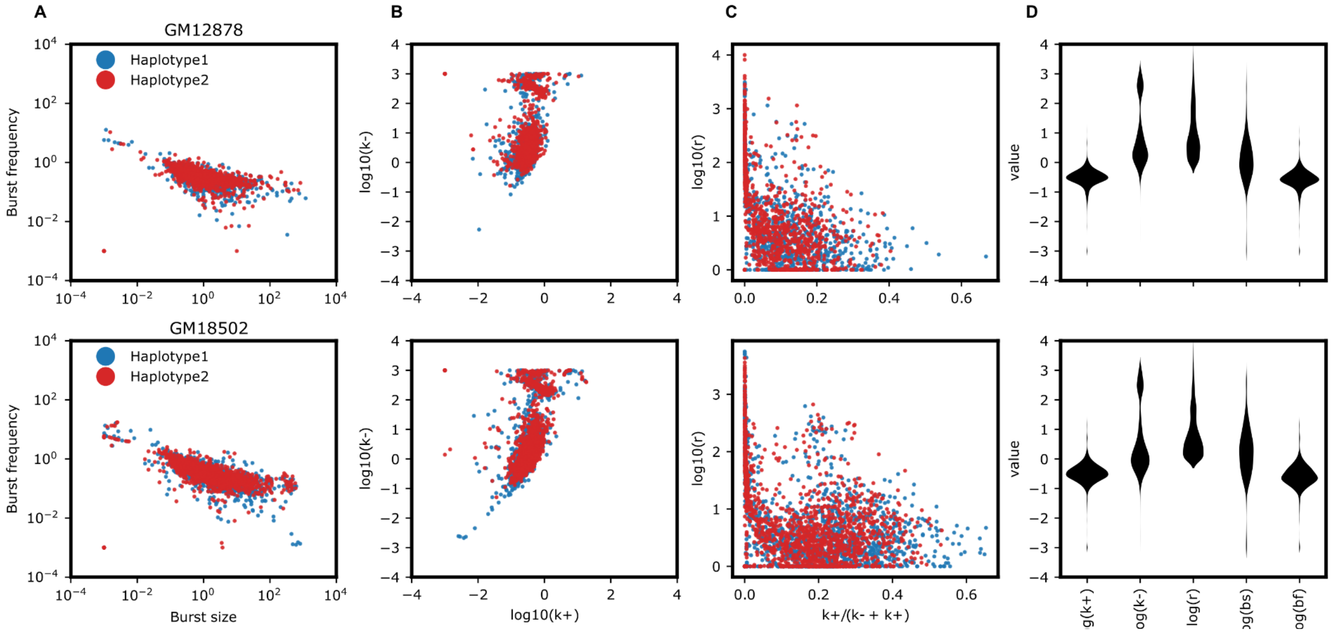


Figure 2. (A-D) The distribution of transcriptional kinetics for GM12878 and GM18502(from up to down). (A) shows the distribution of k_+ and k_- in the log scale. (B) shows the distribution of *r* and $\frac{k_+}{(k_+ + k_-)}$. $\frac{k_+}{(k_+ + k_-)}$ is related to the average active time fraction. (C) shows the distribution of burst size and burst frequency. The burst size and burst frequency are related to k_+ ,k_- , and *r* in the form of $bs = \frac{r}{k_-}$ and $bf = \frac{k_+ \cdot k_-}{k_+ + k_-}$. (D) is the log scale's violin plot of all transcriptional kinetics.

eQTL may exist to alter transcriptional kinetics than mean expression

Previous studies have shown that burst size and burst frequency can be independently regulated by transcriptional machinery and lead to the change of allelic transcription. We want to know if the transcriptional kinetics are influenced by genetic variants or if they are related to eQTL studies that only look at the mean expression. Therefore, we examined the correlation between our transcriptional kinetics and established sets of eQTLs from bulk expression data. We accessed 934,849 significant eQTL from LCLs across four studies in the eQTL Catalog (Kerimov et al. 2021). We retained 48,869 eQTL variants that were heterozygous in GM12878 and were potential eQTLs (alpha=1e-5) for genes with transcriptional kinetics. We normalize the directions for both eQTL effect size and log fold change of transcriptional kinetics from Haplotype2 to Haplotype1. Then we correlated the eQTL effect size with the log-transformed fold change of transcriptional kinetics from Haplotype2 to Haplotype1 (see Methods).

Since bulk gene expression is an averaged expression level across cells and eQTL effect sizes reflect the bulk gene expression change between two alleles, we expect that the mean expression change measured from a single cell experiment will mostly correlate with eQTL effect size. Indeed, the mean and variance change in gene expression between alleles have the highest correlation with the eQTL effect size (Figure 3A, B). Similar results are observed for GM18502 (Supplemental Figure S3A, B).

We found that among transcriptional kinetics, the k+ and burst frequency correlate most with the eQTL effect size (p-value=8.65e-22, Spearman correlation coefficient=0.249; p-value=2.91e-17, Spearman correlation coefficient=0.221), respectively. Similar results are also observed for GM18502 (Supplemental Figure S3C, D). Even though the eQTL is derived primarily from European descent subjects, we still observe a high correlation for k+ and burst frequency with eQTL effect size for GM18502. Based on the correlation results, we hypothesize that eQTLs affect average gene expression mainly through altering k+ and burst frequency.

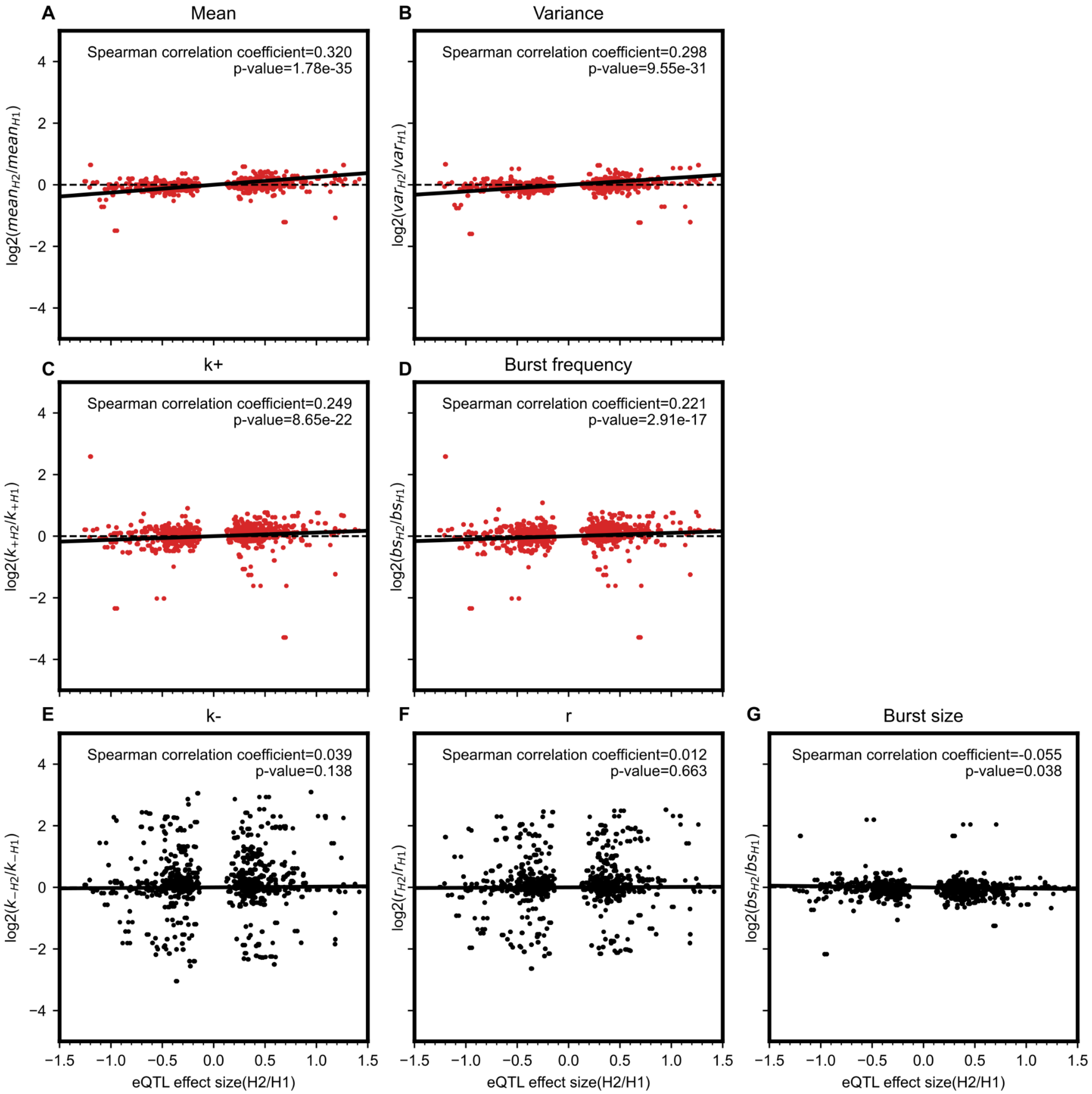


Figure 3. (A-G) Spearman correlation between eQTL effect size and transcriptional kinetics for GM12878. The x-axis is the effect size from the previous eQTL studies, and the y-axis is the $\log_2$ fold change of transcriptional kinetics. The directions for both eQTL effect size and $\log_2$ fold change of transcriptional kinetics are from Haplotype2 to Haplotype1. (A-B) The change of mean and variance in gene expression from scRNA-seq between alleles has the highest correlation with the eQTL effect size. (C-D) k+ and burst frequency have the largest correlation coefficient with the eQTL effect size. (E-G) k-, r, and burst size do not correlate with eQTL effect sizes.

First, we examined if eQTL or transcriptional bursting kinetics analysis can identify the change in expression when burst size and burst frequency are both affected by genetic variants. We simulate an experimentally ideal scenario for both eQTL and transcriptional kinetics analysis with 100 subjects and 1000 cells per subject. We choose the estimated transcriptional kinetics from GM12878 and change them at both burst frequency and burst size. We found that transcriptional kinetics analysis can identify the change in bursting kinetics in all simulated configurations while eQTL analysis cannot when burst size and frequency are changed in opposite directions (Figure 4B). However, we found that it is common for burst size and frequency to change in opposite directions, as shown in GM12878 and GM18502 (Figure 4C).

To further investigate the case where the change of burst size and burst frequency cannot be detected, we simulated it with varied samples and cells where burst size is increased by 2 fold and burst frequency is reduced by half (Figure 4C). We found that even where there are as few as 18 subjects and 600 cells, transcriptional bursting kinetics can still show the change that is detectable with a power greater than 60%. We then investigated why eQTL analysis cannot detect expression changes introduced by underlying altered transcriptional kinetics. When reducing the burst size by half and increasing the burst frequency, the mean expression shows no statistical difference while the variance does (Figure 4D). However, eQTL analysis only reflects the mean expression change driven by the existence of genetic variants. Therefore, when only a small number of cells are available, which is typical for tissue-based gene expression studies, it is more powerful to estimate transcription variance or transcriptional kinetics rather than mean expression.

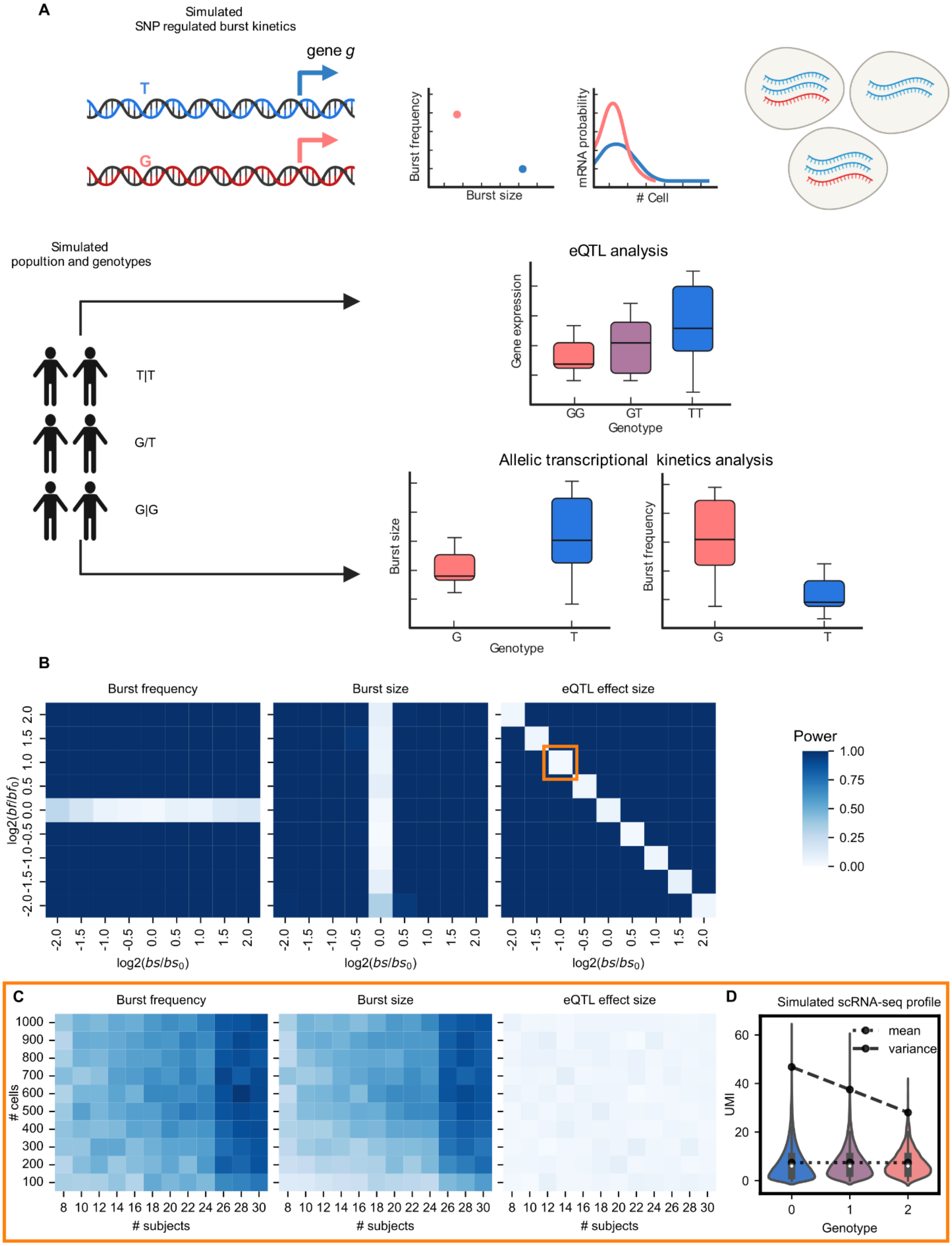


Figure 4. (A) The simulation workflow. A SNP will regulate gene *g* expression by modulating burst size and frequency. We simulated the genotype for n subjects and their allelic expression of gene g for M cells with Poisson-beta distribution with their burst size and frequency. The bulk gene expression for each subject is calculated by averaging through M cells. We then re-estimate the transcriptional kinetics based on each subject's single cell allelic expression. We compare the bursting kinetics for all subjects with the Student *t*-test. On the other side, we perform eQTL analysis on SNP J and bulk

expression of gene g. (B) We simulate the optimal scenario for both eQTL and transcriptional kinetics analysis with 100 subjects and 1000 cells per subject. The x-axis is the $\log_2$ transformed change from bs to bs_0 and the y-axis is the $\log_2$ transformed change from bf to bf_0. Each combination of burst size and burst frequency will be run 100 times independently. The power of either the eQTL test or Student *t*-test of bursting kinetics is the percentage of tests reaching significance level (alpha=0.05). (C) We simulate the scenario where burst size is $0.5bs_0$ and burst frequency $2bf_0$, which is the scenario where eQTL analysis cannot detect the change in gene expression even with 100 subjects and 1000 cells per subject. The x-axis is the number of samples from 8 to 30 and the y-axis is the number of cells from 100 to 1000. Each combination of the number of cells and subjects will be run 100 times independently. The power of either the eQTL test or the Student *t*-test of bursting kinetics is calculated by the percentage of tests reaching the significance level(alpha=0.05). (D) The violin plot of gene expression at single-cell resolution for different dosages of G alleles. The single-cell expression profile is simulated under the experimentally ideal scenario for both eQTL and transcriptional kinetics analysis with 100 subjects and 1000 cells per subject. The amount of transcript from the T allele and G allele is simulated with bursting kinetics (bs_0, bf_0) and ($0.5bs_0$,$2bf_0$).

Transcription Factors bound to the transcription start site explain most transcriptional kinetics

Significant differences in transcriptional kinetics exist between the two chromosomal copies. We hypothesize that transcriptional kinetics are determined by features of the regulatory regions. To examine our hypothesis, we investigate the variance of transcriptional kinetics that can be explained by TF binding or histone markers existing on the same haplotype. In detail, we first obtained the allelic level TF occupancy profile with ChIP-seq reads and phased genotype profiles. We realigned ChIP-seq reads to ASB sites (heterozygous SNPs) located in the regulatory region. The regulatory regions are divided into the distal enhancer region from EnhancerAtlas (Gao and Qian 2020) predicted based on multiple high throughput experimental datasets, and the core promoter region with 1,000 base pairs (bp) upstream and downstream from the transcription start site. The distance between the boundary of the enhancer region and the gene ranges from 1,001bp to 9.7Mbp, with a mean of 1.4M. Then we determined the occupancy for a TF or histone marker on Haplotype1or Haplotype2 by the number of allelic reads mapped to Haplotype1 or Haplotype2. Finally, we estimate the variance of transcriptional kinetics explained by the TF occupancy profile with a linear mixed model (LMM).

In the core promoter profile, ~90% of the variance of burst frequency can be explained by TF occupancy. However, only ~30% variance for burst size is explained (Figure 5). This is consistent with prior findings where burst frequency is highly regulated by transcription factors while burst size is not in mouse embryonic cells (Dobrinic et al., 2021) and *Drosophila* embryos (Fukaya et al., 2016). Compared to the core promoter profile, less variance for bursting kinetics is explained by the distal enhancer profile; it explains ~20% for k+ and ~25% for k-, 37% for r, and 35% for burst frequency. These numbers are on average over 90% in the core promoter profile, therefore, transcriptional kinetics is largely contributed by TFs within the core promoter region.

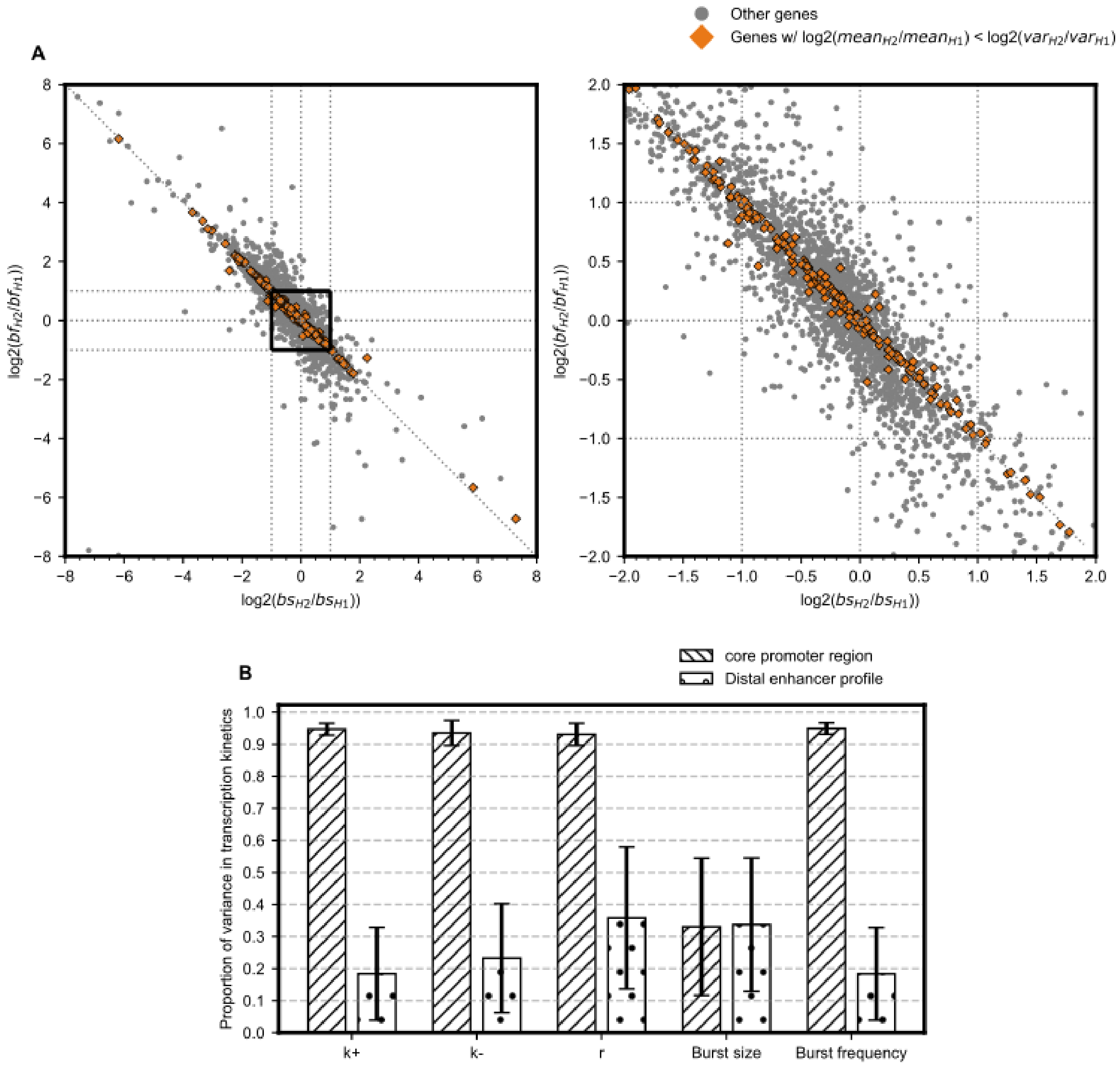


Figure 5. (A) Fold change of bursting kinetics between Haplotype1and Haplotype2 in GM12878 and GM18502. The x-axis is the $\log_2$ transformed change from bs to bs_0 and the y-axis is the $\log_2$ transformed change from bf to bf_0. Left figure is the overall distribution of fold change of bursting kinetics between Haplotype1and Haplotype2 in GM12878 and GM18502. Right figures shows the distribution when $\log_2(bf_{H2}/bf_{H1})$ and $\log_2(bs_{H2}/bs_{H1})$ are bound by (-2, 2). (B) Proportion of variance in transcriptional kinetics explained by TF binding. The bars with back slashes are the estimated variance of transcriptional kinetics explained by TF ASB within the core promoter, while the bars with dots are the estimated variance of transcriptional kinetics explained by TF ASB within the distal enhancer region.

Transcriptional factors regulate gene expression by altering k+ or burst frequency

Given that a significant proportion of the variance in transcriptional kinetics can be explained by the TF ASB profile, we next want to identify the individual effect of TF on transcriptional kinetics.

We combined the core promoter and distal enhancer ASB profiles to investigate the individual TF effect in transcriptional kinetics. Given that both core promoter and distal enhancer regions have ASB profiles, combining them helps increase the number of ASB events and thus the statistical power. Overall, there are 516 genes (519/1,020) with ASB events within either the core promoter or distal enhancer region, and these ASB events come from 112 TFs and Histone markers (112/151).

We performed a paired Wilcoxon test to compare the transcriptional kinetics between Haplotype1 and Haplotype2 of genes across the entire genome. We only include TFs with ASB events in >=10 genes in the Wilcoxon test. The test examines whether the change of a specific TF binding status affects the transcriptional kinetics on average across the genome. We adjust the results for multiple testing with FDR correction. Positive effect size indicates that one allele of a gene with a specific TF binding will have a larger kinetic than the counterpart allele without it and vice versa.

We found that for LCL, the ASB events of multiple TFs are associated with transcriptional kinetics (Figure 6B). The binding of *YY1* is associated with gene expression by increasing k+ and burst frequency (FDR<0.1). The pair-wise comparison between *YY1* bound and unbound alleles are shown in Figure 6C. ASB events of Multiple TFs including *YY1, EED, POLR2AphosphoS5, EBF1, IKZF2, PAX5, MTA2, and FOXK2* are associated with the change of transcriptional kinetics at nominal significance (p-value<0.05). In summary, most of the significant ASB events of TF are associated with the change of k+ and burst frequency. In contrast, we found no significant association between the transcriptional kinetics and allelic open chromatin (Supplemental figure S4).

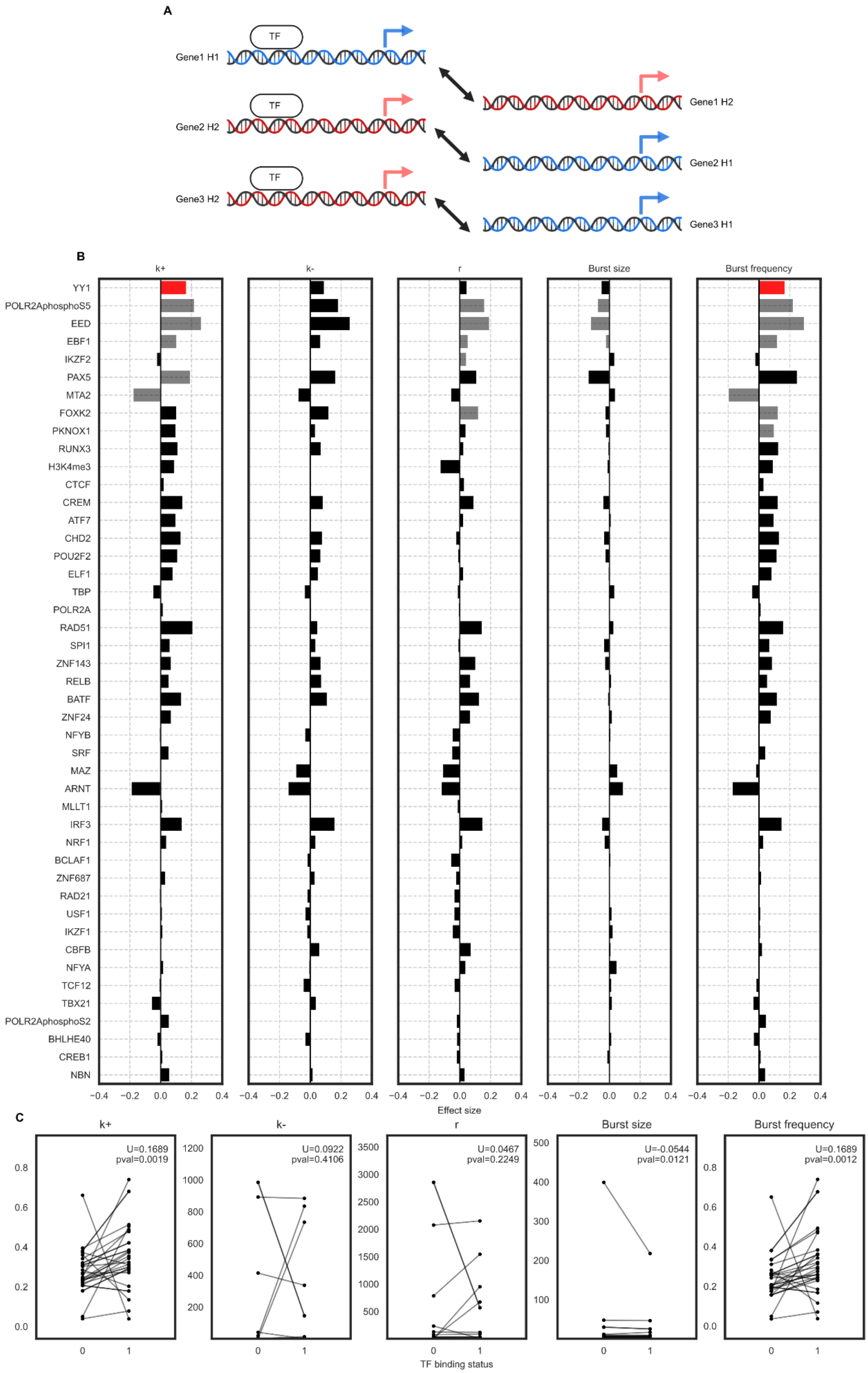

A
TF
Gene1 H1
TF
Gene2 H2
TF
Gene3 H2
Gene1 H2
Gene2 H1
Gene3 H1
B
k+
k-
r
Burst size
Burst frequency
YY1
POLR2AphosphoS5
EED
EBF1
IKZF2
PAX5
MTA2
FOXK2
PKNOX1
RUNX3
H3K4me3
CTCF
CREM
ATF7
CHD2
POU2F2
ELF1
TBP
POLR2A
RAD51
SPI1
ZNF143
RELB
BATF
ZNF24
NFYB
SRF
MAZ
ARNT
MLLT1
IRF3
NRF1
BCLAF1
ZNF687
RAD21
USF1
IKZF1
CBFB
NFYA
TCF12
TBX21
POLR2AphosphoS2
BHLHE40
CREB1
NBN
−0.4 −0.2 0.0 0.2 0.4
Effect size
C
k+
U=0.1689
pval=0.0019
0.0 0.2 0.4 0.6 0.8
k-
U=0.0922
pval=0.4106
200 400 600 800 1000 1200
r
U=0.0467
pval=0.2249
500 1000 1500 2000 2500 3000 3500
Burst size
U=-0.0544
pval=0.0121
100 200 300 400 500
Burst frequency
U=0.1689
pval=0.0012
0.0 0.2 0.4 0.6 0.8
0 1
TF binding status

Figure 6. (A)The work schema for the paired test. The comparison is done across genes with only one allele binding TF and between pairs of alleles. (B) Association between the ASB TF occupancy and kinetics with the Wilcoxon test. Overall, 46 TFs that have ASB for more than 10 genes tested for transcriptional kinetics. We correct each Transcriptional kinetic result for multiple testing.*YY1* binding increases burst frequency and k+ are colored red (FDR < 0.1). Other transcriptional kinetics associated with *YY1*, *EED*, *POLR2AphosphoS5*, *EBF1*, *IKZF2*, *PAX5*, *MTA2*, and *FOXK2* binding at nominal significance are colored gray. Positive effect size indicates that one allele of a gene with TF binding will have a larger transcriptional kinetic than the counterpart allele without the TF and vice versa. (C) Association results between the *YY1* occupancy and transcriptional kinetics with paired Wilcoxon test.

eQTLs induce preferential TF binding that shapes the bursting kinetics of gene expression

In the eQTL section, we have shown that we can potentially detect a genetic variant affecting bursting kinetics through transcriptional kinetics analysis (18 samples with 600 cells per sample to reach a power of 60%; Figure 4C). On the other hand, we have shown that ASB of TF is associated with transcriptional kinetics. In this section, we will combine them to infer potential eQTL mechanisms regulating burst kinetics by altering TF binding affinity and how they associate with phenotypes in the GWAS catalog.

An eQTL and GWAS Catalog variant rs9271588 (Table 1a-c) is carried by GM12878, the T allele of which induces preferential binding of *SRF*, and consequently leads to higher burst frequency and larger transcription variance. rs9271588 is close to the transcription start site of *HLA-DQA1* and overlaps with the *SRF* ChIP-seq peaks (Figure 7A), which is mostly mapped to the rs9271588_T allele. To further confirm that rs9271588 will induce ASB of *SRF*, we use Qbic-pred (Martin et al. 2019) to predict the consequence of rs9271588 and found that TF is bound with the T allele and unbound with the C allele. Therefore, the ASB of *SRF* might be introduced by the rs9271588.

rs9274623 and rs9271586, GWAS catalog loci and eQTLs for *HLA-DQA1* (Table 2a,b), are also associated with multiple TF preferential binding events. While none of the loci are predicted to change TF binding affinity, there is *PORL2* ChIA-PET evidence showing physical contact between these loci and rs9271588. Therefore, the TF preferential binding including *IKZF2*, *PKNOX1*, and *MLLT1* is likely to be introduced due to their physical contact with *SRF,* which is preferential binding at rs9271588.

In GM18502 where no TF ChiP-seq and ASB profile were available, we found that the haplotype in phase with rs9271588_T (Haplotype2) is again concurrent with higher burst frequency and larger transcription variance. Therefore, rs9271588 is also likely to change *SRF* preferential binding in GM18502 and lead to the change in allelic transcription reflected in increased burst frequency and large transcription variance.

The rs9271588_C allele, predicted to have lower TF binding affinity and consequently resulting in low burst frequency and small transcription variance in LCL, is associated with an increased count of cells of myeloid lineage. Lymphoid progenitor cells have similar HLA class II expression profiles to hematopoietic stem cells (Boegel et al. 2018). Therefore, rs9271588_C might have a similar influence on *HLA-DQA1* in hematopoietic stem cells to LCL, which results in cell-to-cell heterogeneity in hematopoietic stem cells and consequently an increased count of cells in myeloid lineage. While the risk alleles of rs9274623 and rs9271586 are not in phase with rs9271588_C, they have been shown associated with open chromatin (Table 2a,b; Goes et al. 2015) and consistent with the ASB results in Figure 7.

Summarizing the above evidence, TF preferential binding is associated with the allelic expression and bursting kinetics. More specifically in the *HLA-DQA1* example, the higher burst frequency and larger transcription variance are associated with a collection of TF preferential binding.

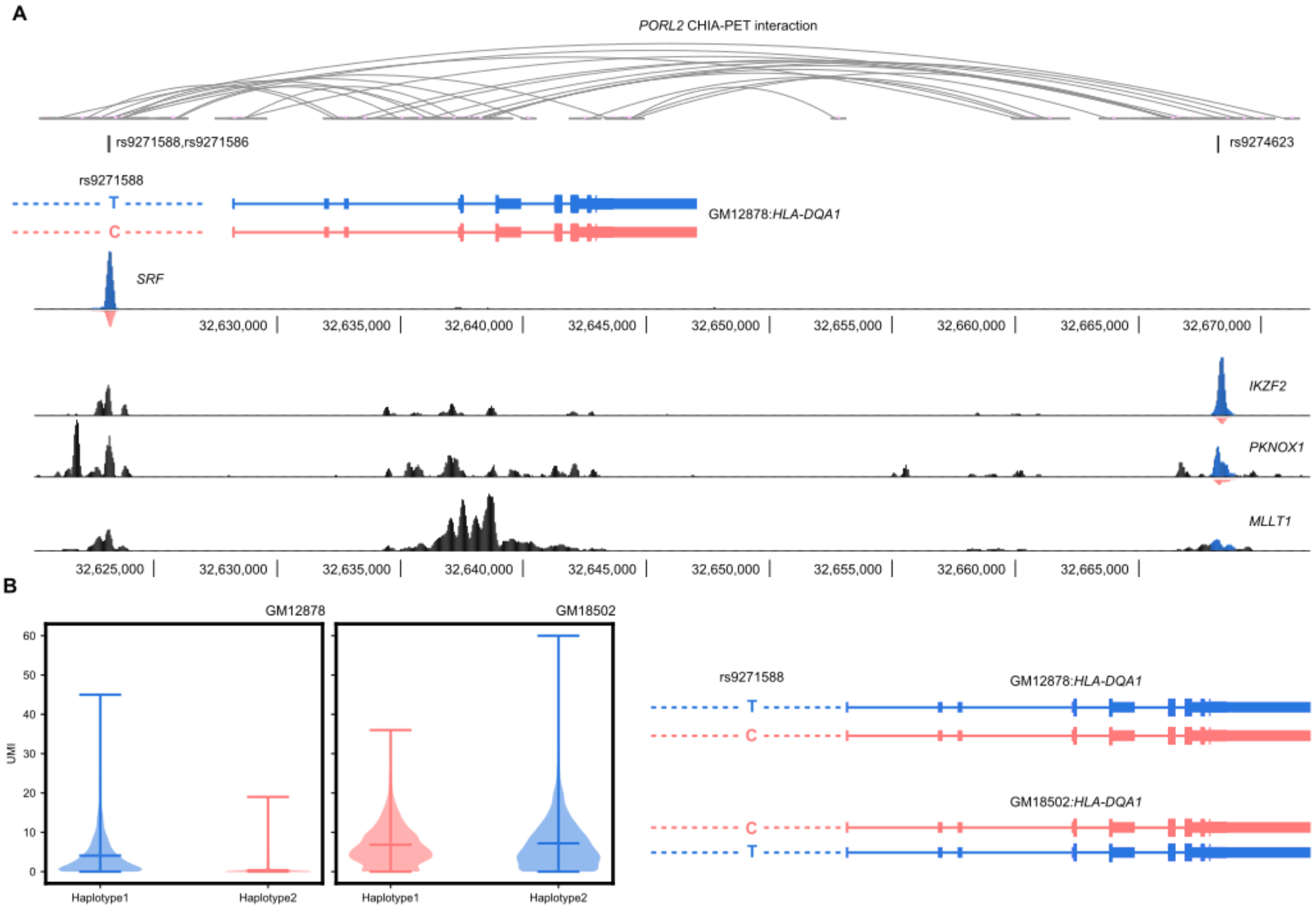


Figure 7. (A) Chromatin context of rs9271588 in GM12878. rs9271588 is an eQTL and GWAS Catalog locus near the transcription start site of *HLA-DQA1*. On rs9271588, the ChIP-seq peaks of *SRF* show *ASB*. Multiple TFs including *IKZF2*, *PKNOX1*, and *MLLT1* show preferential binding in the downstream region of *HLA-DQA1*. The *PORL2* CHIA-PET shows the interaction between the rs9271588 and downstream ASB sites. (B) The histogram of UMI between Haplotype1and Haplotype2 in GM12878 and GM18502. In GM12878, Haplotype1 in phase with rs9271588_T is colored blue and shows a larger mean and variance of transcription while Haplotype2 in phase with rs9271588_C is colored red and shows a smaller mean and variance of transcription. In GM18502, Haplotype1in phase with rs9271588_C is colored blue and shows a smaller mean and variance of transcription while Haplotype2 in phase with rs9271588_C is colored red and shows a larger mean and variance of transcription.

Table 1a: Busting kinetics affected by rs9271588

Gene	TF		Allele	Burst size	Burst frequency	Mean	Variance

HLA-DQA1	*SRF*	GM12878	**T**(Haplotype1)	5.695	0.65	3.680	23.511
			C(Haplotype2)	8.278	0.0368	0.310	1.438
		GM18502	**C**(Haplotype1)	7.554	0.902	6.841	29.773
			T(Haplotype2)	6.168	1.166	7.195	43.320

Table 1b: *HLA-DQA1* association with rs9271588 in the eQTL catalog

Gene	eQTL Locus	Study	Tissue	-log10(p)	Effect Size	SE (Effect Size)
HLA-DQA1	rs9271588_**T**/**C**	GEUVADIS	LCL	8.25	-0.58	0.098
		GTEx		3.69	-0.76	0.2
		GENCORD		2.9	-0.55	0.17

Table 1c: phenotypes for rs9271588,rs9271586 and rs9274623 in the GWAS catalog

Variant and risk allele	P-value	Beta	CI	Trait(s)	Study accession
rs9271588-C	2 x 10^{-58}	0.0576552 unit increase	[0.051-0.065]	neutrophil count, basophil count	GCST004620
	2 x 10^{-66}	0.06164672 unit increase	[0.055-0.069]	granulocyte count	GCST004614
	4 x 10^{-66}	0.06142442 unit increase	[0.054-0.068]	neutrophil count, eosinophil count	GCST004613
	9 x 10^{-68}	0.06247162 unit increase	[0.055-0.07]	myeloid white cell count	GCST004626
	1 x 10^{-58}	0.05764791 unit increase	[0.051-0.065]	neutrophil count	GCST004629

Table 2a: Table 1b: *HLA-DQA1* association with rs9274623 and rs9271586 in the eQTL catalog

Gene	eQTL Locus	Study	Tissue	-log10(p)	Effect Size	SE(Effect Size)
HLA-DQA1	rs9274623_**G**/T	GEUVADIS	LCL	14.35	-1.04	0.13
		GENCORD		7.21	-1.33	0.23
		GTEx		3.05	-0.89	0.26
	rs9271586_**G**/T	GEUVADIS	LCL	8.25	-0.58	0.098
		GTEx		3.69	-0.76	0.2
		GENCORD		2.9	-0.55	0.17

Table 2b: Genes for rs9271588 in the eQTL catalog

Variant and risk allele	P-value	Odds	CI	Trait(s)	Study accession
rs9271586-G	6 x 10^{-115}	-	-	Neutrophil count	GCST90056178
rs9274623-G	6 x 10^{-19}	1.1363636	-	Schizophrenia	GCST003048

Discussion:

We have shown in GM12878 and GM18502, allele-specific expression is prevalent in LCL, specifically in the form of random monoallelic expression in single cells. By associating the allelic level transcriptional kinetics with the allelic-specific TF occupancy profile, we found that the allelic-specific binding of TF is associated with the regulation of transcriptional kinetics and consequently leads to allelic expression. Such allelic expression not only differs in mean transcription but also the transcription variance. Therefore, allelic transcriptional kinetics derived from scRNA-seq data provide higher resolution to study allelic-specific expression than eQTL studies.

Two assumptions are made for the kinetic estimations: the ergodic process and constant mRNA degradation rate. The ergodic process assumption is valid for this study since no perturbation is introduced, and GM12878 is an isogenic cell line. However, the ergodic process can be violated when multiple cell types are involved. We anticipate that along with scRNA-seq data, there will be independent information like surface receptors to help characterize the cell population. The second assumption is the constant mRNA degradation rate. Rabani et al. (2011) have shown that most genes in mammalian cells have a similar degradation rate, and the expression of those genes is little correlated with the change in degradation rates. Therefore, it is valid to estimate kinetics assuming a constant degradation rate and compare the transcriptional kinetics scaled by the degradation rate between genes.

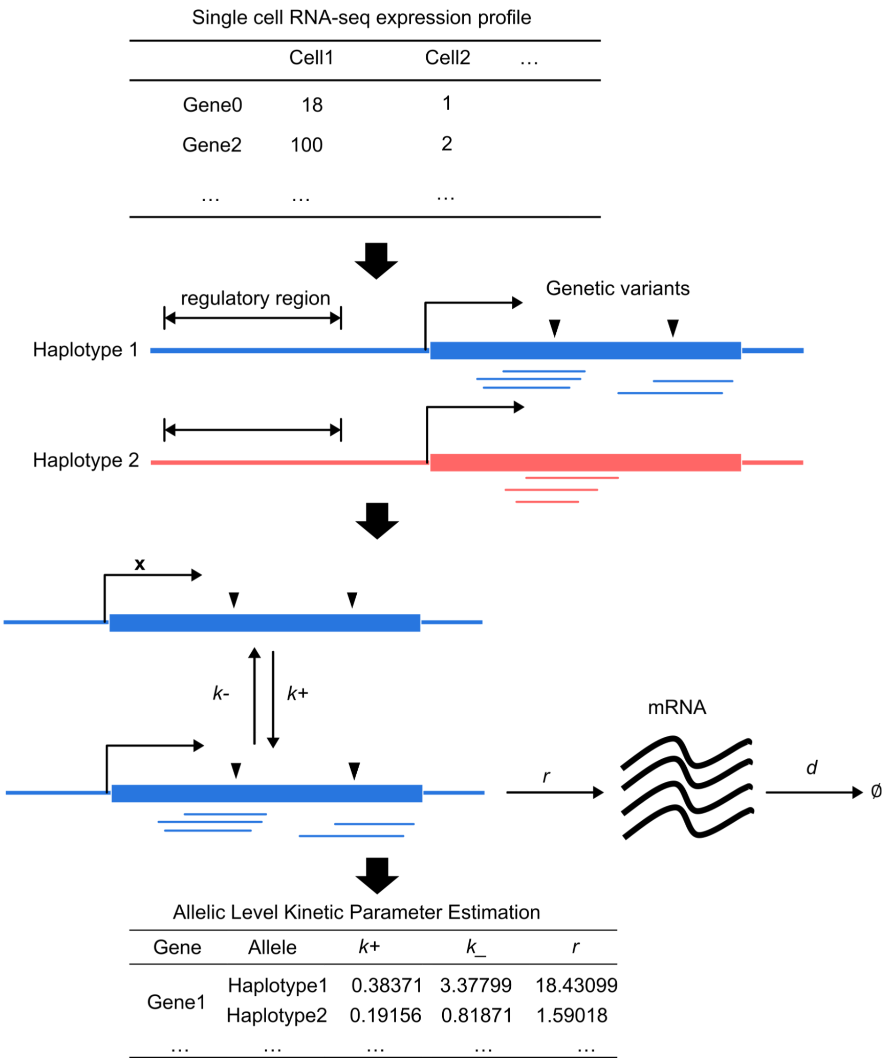
When we investigated the individual TF effect on transcriptional kinetics, we measured the averaged effect of TF binding on the transcriptional kinetics across the genome. However, TF can act collectively to regulate gene expression and is highly orchestrated by the chromatin 3D structure. Figure 7 has demonstrated an example where multiple TFs share the same ASB site. Because of the above possibilities, there can be outliers shown in Figure 6C, of which the change of transcriptional kinetics is not consistent with other pairs. In this case, the binding of *YY1* might be shadowed by other regulatory factors.

The kinetic parameter estimation is sensitive to the protocol of scRNA-seq. Because the scRNA-seq dataset here is generated with the 10x Genomics 3' sequencing, we have less coverage of exome regions with scRNA-seq reads and, therefore, fewer allelic expression events identified. We use a stringent threshold to retain reads with unique alignment and high base-pair sequencing quality at the ASE SNP site. With such a strategy, we can estimate the allelic expression and transcriptional kinetics for 1,020 genes, while the original expression profile is available for 21,524 genes. We recognize that we compromise a larger proportion of lowly expressed genes during the QC process. However, with the latest scRNA-seq protocols, such as the Smart-seq3(Hagemann-Jensen 2020), more genes can likely be included in the estimation of the kinetic parameters.

In this study, we have focused only on the effect of TF on kinetics. However, other notable factors can shape gene expression profiles and regulate the kinetics of genes, such as chromatin looping, DNA methylation, etc. We anticipate that with the development of the multi-omics technique in the single-cell paradigm single-cell profiles with multiple measurements will be available, which can help further explain the mechanism of gene expression and characterize the regulation of transcriptional bursts in depth.

MATERIAL AND METHODS

Kinetic Parameter Estimation pipeline



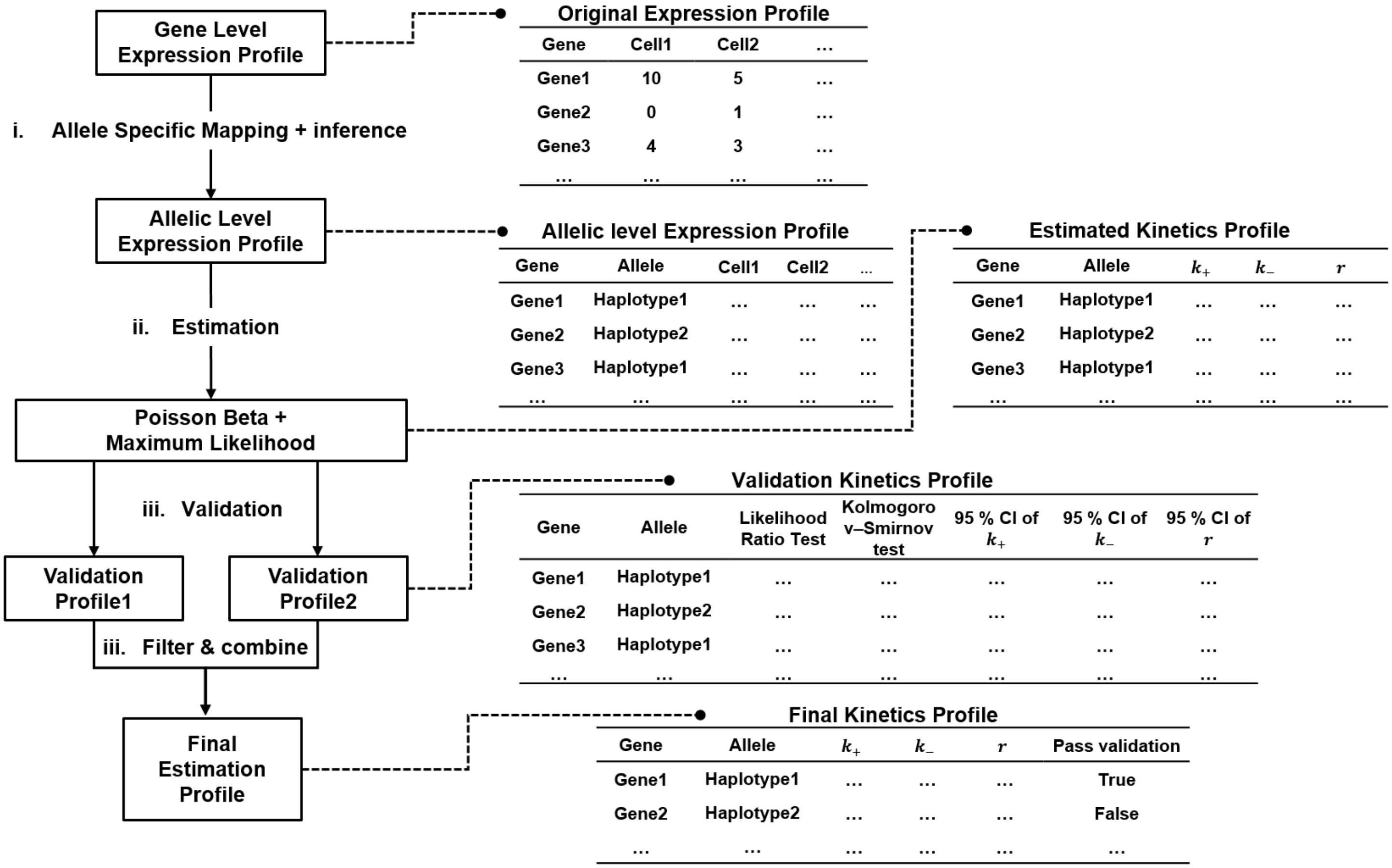
Single cell RNA-seq expression profile

	Cell1	Cell2	…
Gene0	18	1	
Gene2	100	2	
…	…	…	

Allelic Level Kinetic Parameter Estimation

Gene	Allele	*k*+	*k*_	*r*
Gene1	Haplotype1	0.38371	3.37799	18.43099
	Haplotype2	0.19156	0.81871	1.59018
…	…	…	…	…

Figure 8. Diagram for kinetic parameter estimation from the scRNA-seq profile with phased genotype profile. The original scRNA-seq profile is with the unique molecular identifier(UMI) unit. We mapped the scRNA-seq reads back to the chromosomes at heterozygous SNPs and derived the allelic level expression profile. The allelic level expression profile will be fit with the two-state model to obtain the estimation of three kinetic parameters: *k*+, *k*-, and *r*, which are all scaled by the degradation rate *d*.


Gene Level Expression Profile
i. Allele Specific Mapping + inference
Allelic Level Expression Profile
ii. Estimation
Poisson Beta + Maximum Likelihood
iii. Validation
Validation Profile1
Validation Profile2
iii. Filter & combine
Final Estimation Profile

Original Expression Profile

Gene	Cell1	Cell2	...
Gene1	10	5	...
Gene2	0	1	...
Gene3	4	3	...
...	...	...	...

Allelic level Expression Profile

Gene	Allele	Cell1	Cell2	...
Gene1	Haplotype1	...	...	...
Gene2	Haplotype2	...	...	...
Gene3	Haplotype1	...	...	...
...	...	...	...	...

Estimated Kinetics Profile

Gene	Allele	k_+	k_-	r
Gene1	Haplotype1	...	...	...
Gene2	Haplotype2	...	...	...
Gene3	Haplotype1	...	...	...
...	...	...	...	...

Validation Kinetics Profile

Gene	Allele	Likelihood Ratio Test	Kolmogoro v–Smirnov test	95 % CI of k_+	95 % CI of k_-	95 % CI of r
Gene1	Haplotype1	...	...	...	...	...
Gene2	Haplotype2	...	...	...	...	...
Gene3	Haplotype1	...	...	...	...	...
...	...	...	...	...	...	...

Final Kinetics Profile

Gene	Allele	k_+	k_-	r	Pass validation
Gene1	Haplotype1	...	...	...	True
Gene2	Haplotype2	...	...	...	False
...	...	...	...	...	...

Figure 9. Workflow for the allelic-specific mapping, kinetic estimation, and validation process.

1) Obtain Allele-Specific expression from the scRNA-seq profile

The allele-specific mapping is done by combining the scRNA-seq data and phase genotype profiles. We first pile all scRNA-seq reads upon phased heterozygous SNPs with Samtools (Li et al. 2009). The minimum mapping quality for an alignment is 50, and the minimum base quality is 30.

Let *g* denote gene index and *H1* denote Haplotype1. We first process cells with reads mapped to gene *g* and allelic-specific mapping information available. A read mapped to multiple SNPs on one chromosome within the gene *g* will be counted as one single read. For reads mapped to multiple SNPs on both chromosomes, the reads will be discarded in allele-specific mapping. We derive the allele ratio by calculating the proportion of reads from two alleles, and then use the allele ratio to infer the allele count for a gene *g* in a single cell. $m_g^k(H1)$ represents the number of transcripts from Haplotype1 for gene *g* in the *kth* cell. M_g^k is the number of total transcripts for gene *g* in the *kth* cell. We assume that in cell *k* with M_g^k transcripts, the probability of observing $m_g^k(H1)$ follows a beta-binomial distribution(1,2).

$$m_g^k(H1)|M_g^k \sim Bin(M_g^k, \theta_g),\ k \in \{1, 2,... K\}\ (1)$$

$$\theta_g \sim Beta(\alpha_g, \beta_g) \ (2)$$

where α_g *and* β_g are canonical shape parameters of the beta distribution and θ_g is the probability sampled from the beta distribution. Based on α_g and β_g, We can derive the expectation for observing a transcript from Haplotype1 within a single cell(3).

$$P_g = \frac{\alpha_g}{(\alpha_g + \beta_g)} \ (3)$$

Let $\boldsymbol{\eta} = \{(m_g^k(H1), M_g^k),\ k = 1, 2, \ldots K\}$ denote the allelic expression observation in *K* cells. We first obtain the log-likelihood of the compound distribution shown in (4). We use the maximum likelihood estimation (MLE) to estimate the α_g and β_g in cells with ASE reads (5) and obtain the confidence interval for α_g and β_g from the Hessian matrix (6).

$$\boldsymbol{\mathcal{L}}(\alpha_g, \beta_g; \boldsymbol{\eta}) = \sum_{k=1}^{K} ln[f(m_g^k(H1) | M_g^k; \alpha_g, \beta_g)] = \sum_{k=1}^{K} ln[\frac{\Gamma(M_g^k + 1)}{\Gamma(m_g^k(H1)+1)\Gamma(M_g^k - m_g^k(H1)+1)} \frac{\Gamma(m_g^k(H1)+\alpha_g)\Gamma(M_g^k - m_g^k(H1)+\beta_g)}{\Gamma(M_g^k + \alpha_g + \beta_g)} \frac{\Gamma(\alpha_g + \beta_g)}{\Gamma(\alpha_g)\Gamma(\beta_g)}]$$

$$(4)$$

$$\hat{\alpha}_g, \hat{\beta}_g = argmax\, \boldsymbol{\mathcal{L}}(\alpha_g, \beta_g; \boldsymbol{\eta}) \ (5)$$

$$\sigma^2(\boldsymbol{\theta}) = [KI(\boldsymbol{\theta})]^{-1}, I(\boldsymbol{\theta}) = - \{\frac{\partial^2 \boldsymbol{\mathcal{L}}}{\partial\theta_i \partial\theta_j}\}, \boldsymbol{\theta} = (\alpha_g, \beta_g)' \ (6)$$

For cells with reads mapped to gene *g* but no allelic-specific reads available, we use the estimated beta-binomial distribution (1) and (2) infer the Haplotype1 and Haplotype2 transcription count of gene *g*.

To ensure valid inference, we only keep genes with positive α_g, β_g, and confidence intervals. α_g and β_g estimated with an unrealistic confidence interval (i.e., confidence interval covering 0) indicates a failed convergence during maximum likelihood estimation and thus will not be kept.

In summary, with either observed allele-specific reads or inferred allele-specific reads, we divided the overall unique molecular identifier (UMI) to Haplotype1 and Haplotype2 to obtain the allele-specific expression profile for gene *g*.

2) Estimate Kinetic Parameters

The stochastic gene expression is characterized by a two-state model (Figure 8). In the two-state model, a gene switches between active and inactive status with rates of *k*+, *k*-. While in the active status, the gene can transcribe mRNA with a rate of r. The degradation rate of a transcript is *d*.

We use the Poisson-Beta distribution as the steady-state distribution and estimate the kinetic parameters using the maximum likelihood approach. The probability density function for Poisson-beta distribution takes the same form as the analytical solution for the two-state model (Bressloff 2014). We denote *m* as the number of mRNA in the cell, $m_s = r/d$ is the mRNA number at the steady-state, $\xi_0 = k_-/d$, $\xi_1 = k_+/d$, and *F* is the confluent hypergeometric function of the first kind (Shahrezaei and Swain 2008). The analytical solution for the two-state model is shown in Formula 7. The Poisson-beta distribution for a representative steady-state distribution is shown in Formula 8 (Vu et al. 2016), where λ is the hyper-parameter can be interpreted as the stochastic switch rate.

$$p(m) = \frac{m_s^m e^{-m_s}}{m!} \bullet \frac{\Gamma(\xi_0+m)\Gamma(\xi_0+\xi_1)}{\Gamma(\xi_0)\Gamma(\xi_0+\xi_1+m)} F(\xi_1,\ \xi_0 + \xi_1 + m; m_s) \quad (7)$$

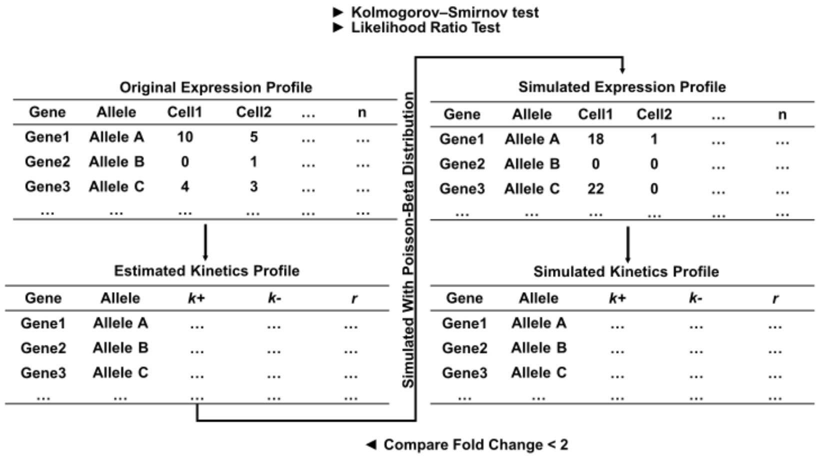
$$p(m) = \int_\lambda \rho_{ss}(\lambda) e^{-\lambda} \frac{\lambda^m}{m!} d\lambda;\ \rho_{ss}(\lambda) = \frac{\lambda^{\xi_1-1}(1-\lambda)^{\xi_0-1}}{B(\xi_1, \xi_0)} \quad (8)$$

The kinetic parameters estimated from all three methods are *k*+, *k*-, and *r,* scaled by *d*. However, *k*+, *k*-, and *r* are not necessarily independent of each other. Therefore, we derive a set of empirically orthogonal bursting kinetics including bursting size (bs) and burst frequency (bf).

$$bs = \frac{r}{k_-};\ bf = \frac{k_+ k_-}{k_+ + k_-} \quad (9)$$

3) Validate, Filter, and Generate Valid Kinetic Parameter Profile

We validate the estimated kinetic parameter profile using two independent approaches and select estimated kinetic parameters whose validation profile has passed either filtering criteria.


► Kolmogorov–Smirnov test
► Likelihood Ratio Test
Simulated With Poisson-Beta Distribution
◄ Compare Fold Change < 2

Original Expression Profile

Gene	Allele	Cell1	Cell2	...	n
Gene1	Allele A	10	5	...	...
Gene2	Allele B	0	1	...	...
Gene3	Allele C	4	3	...	...
...	...	...	...	...	...

Simulated Expression Profile

Gene	Allele	Cell1	Cell2	...	n
Gene1	Allele A	18	1	...	...
Gene2	Allele B	0	0	...	...
Gene3	Allele C	22	0	...	...
...	...	...	...	...	...

Estimated Kinetics Profile

Gene	Allele	*k+*	*k-*	*r*
Gene1	Allele A	...	...	...
Gene2	Allele B	...	...	...
Gene3	Allele C	...	...	...
...	...	...	...	...

Simulated Kinetics Profile

Gene	Allele	*k+*	*k-*	*r*
Gene1	Allele A	...	...	...
Gene2	Allele B	...	...	...
Gene3	Allele C	...	...	...
...	...	...	...	...

Figure 10. The flow chart of the first validation approach

In the first approach, we simulate gene expression based on the kinetic parameter estimated from the original expression profile and re-estimate the kinetic parameters again from the simulated expression profile. This approach examines the correlation between the original and simulated expression profiles and the correlation between the estimated and simulated kinetic parameter profiles. We use the Kolmogorov–Smirnov test and the likelihood ratio test to examine the correlation between the original and simulated expression profiles. When Kolmogorov–Smirnov and Likelihood ratio tests are not significant, indicating that the underlying probability distributions for the original expression profile do not deviate from the simulated expression profile. We also require the estimated kinetic parameter $\hat{\phi}$ and the simulated kinetic parameter to be restricted by

$\left|log_2(\phi_s/\hat{\phi})\right| < 1$, indicating that the simulated kinetic parameters do not differ from the originally estimated kinetic parameters. Estimated kinetic parameters will be kept if Kolmogorov–Smirnov and Likelihood ratio tests are not significant and the estimated kinetic parameter $\hat{\phi}$ and the simulated kinetic parameter are restricted by $\left|log_2(\phi_s/\hat{\phi})\right| < 1$.

In the second approach, we measure the variance of each estimated kinetic parameter and filter out those with large variance. Let $\boldsymbol{M} = \{m_k, k = 1, 2, ... K\}$ denote the observation of transcript in K cells. The posterior distribution $p(k_+, k_-, r|\boldsymbol{M})$ is sampled with the Metropolis-Hastings algorithm (Zoller et al. 2018) in 10,000 iterations. The final estimation and 95% credible interval are derived from the marginal posterior distribution for each kinetic parameter. We require the upper bound $\hat{\phi}_{UB}$ and lower bound $\hat{\phi}_{LB}$ of the confidence intervals for kinetics to be $\left|log_2(\hat{\phi}_{UB}/\hat{\phi})\right| < 1$ and $\left|log_2(\hat{\phi}_{LB}/\hat{\phi})\right| < 1$.

Allelic binding of transcriptional factors and histone modification

We combine the ChIP-seq data and phased genotype from GM12878 to identify allelic-specific TF binding and allelic-specific histone modification events, generalized as ASB in the follows. We used 152 TF and 11 histone ChIP-seq datasets for GM12878 from the ENCODE4 project (Davis et al. 2018). We realign the ChIP-seq reads upon heterozygous SNPs located in the regulatory region and tested if there are allelic-specific binding events with BaalChIP (de Santiage et al. 2017). The regulatory regions are defined as the distal enhancer region from EnhancerAtlas (Gao and Qian 2020) and the core promoter region with 1,000 base pairs (bp) upstream and downstream from the transcription start site. SNPs will be clustered if they are located within the same CHiP-seq peak. If within a single CHiP-seq peak there are multiple SNPs showing both ASB and nonASB events, this CHiP-seq peak will not be considered an ASB event.

We derive the TF occupancy profile from ChIP-seq data by encoding the ASB events. For SNPs that show ASB, based on the phased genotype and allelic reads mapping, we will determine if the TF preferably binds to Haplotype1 or Haplotype2. The preferably bound allele will be encoded as 1 for TF occupancy while the counterpart allele will be encoded as 0. If a single CHiP-seq peak shows conflict ASB events by multiple SNPs, it will not be considered an ASB event.

Associate the transcriptional kinetics with the eQTL effect size

Given the eQTL effect size is defaulted as the change from the alternative allele against the reference allele, we use the following strategy to make the allelic transcriptional kinetics in phase with the eQTL. First, based on the phased genotype, we modified the eQTLs effect size to reflect the change of gene expression from Haplotype2 against Haplotype1. Meanwhile, we derived the log-transformed fold change of transcription burst from Haplotype2 against Haplotype1. With allelic transcriptional kinetics in phase with the eQTL, we associated the transcriptional kinetics with the eQTL effect size by the spearman correlation test.

Power simulation

In Figure 4B, We simulate the scenario for both eQTL and transcriptional kinetics analysis with *N*=100 subjects and *K*=1000 cells per subject. We choose the bursting kinetics bs_0 and bf_0 from *HLA-DQA1* estimated from GM12878. We simulate SNP *J*, the allele of which results in two different sets of bursting kinetics ($bs,\ bf$) and ($bs_0,\ bf_0$) for gene *g*.

We determined the minor allele frequency *p* for the regulatory SNP *J* and simulated the genotype for *N*=100 subjects assuming SNP *J* following Hardy-Weinberg equilibrium. Based on their genotype, we simulate each subject expression of gene *g* for *K*=1000 cells. $m_g^{nk}(H1)$ is the expression of Haplotype 1 allele of gene *g* for the *n*th subject in the *k*th cell and it follows a Poisson-beta distribution with their burst size $bs_g^{nk}(H1)$and burst frequency $bf_g^{nk}(H1)$ (10). Similar annotation applies to $m_g^{nk}(H2)$, $bs_g^{nk}(H2)$, and $bf_g^{nk}(H2)$. The bulk expression g_n for the *n*th subject is simulated by averaging across *K* cells and two alleles H1 and H2 (12). We then re-estimate the transcriptional kinetics for the *k*th subject based on its allelic expression profile $\left\{m_g^k(H1),\ m_g^k(H1)\right\},\ k\epsilon\{1..K\}$ across K cells (13, 14). We compare the bursting kinetics for all subjects with the Student *t*-test. On the other side, we perform eQTL analysis on SNP *J* and the bulk expression of gene *g*. Finally, we compare the results from the eQTL analysis and the re-estimated bursting kinetics. Each combination of burst size and burst frequency will be run 100 times independently. The power of either the eQTL or burst kinetic test is the percentage of tests reaching the significance level (alpha=0.05).

$$m_g^k(H1) \sim PoissonBeta\left(bs_g^n(H1),\ bf_g^n(H1)\right); n\epsilon\{1...N\},\ k\epsilon\{1..K\}\ (10)$$

$$m_g^k(H2) \sim PoissonBeta\left(bs_g^n(H2),\ bf_g^n(H2)\right); n\epsilon\{1...N\},\ k\epsilon\{1..K\}(11)$$

$$g_n = E\left[m_g^{nk}(H1)\right] + E\left[m_g^{nk}(H2)\right],\ k\epsilon\{1..K\}(12)$$

$$\widetilde{bs_g^n}(H1),\ \widetilde{bf_g^n}(H1) \sim \left\{m_g^k(H1)\right\},\ m\epsilon\{1..M\}(13)$$

$$\widetilde{bs_g^n}(H2), \widetilde{bf_g^n}(H2)\ \sim \left\{m_g^k(H2)\right\},\ m\epsilon\{1..M\}(14)$$

Similarly for Figure 4C, we set up the scenario where the allele of SNP *J* will result in a corresponding set of bursting kinetics ($0.5bs_0$, $2bf_0$) and ($bs_0$and bf_0). We varied the number of samples *N* from 8 to 30 and the number of cells *K* from 100 to 1000. Each combination of *K* and *N* will be run 100 times independently. The power of either the eQTL or burst kinetic test is the percentage of tests reaching the significance level (alpha=0.05).

Associate the transcriptional kinetics with the ASB profile

We use a linear mixed model to estimate the variance of transcriptional kinetics explained by TF occupancy, where $\boldsymbol{D}$ is the TF occupancy profile and σ^2 is the variance of a Transcriptional kinetic explained by $\boldsymbol{D}$.

$$y \sim N(0, \sigma^2 \boldsymbol{D}^T \boldsymbol{D}) \quad (15)$$

While LMM requires the residual to follow a normal distribution, we do not know if the transcriptional kinetics are violating the assumption. However, we argue that we only use the LMM to estimate the strength of association between the transcriptional kinetics and TF occupancy. If a large proportion of variance of transcriptional kinetics can be explained by the TF occupancy, we have theoretical evidence to support transcriptional kinetics are likely contributed by gene regulation, which allows us to further investigate the individual TF and their effect on transcriptional kinetics.

We examine the TF occupancy within the regulatory region. The regulatory region is divided into the distal enhancer region and core promoter region. Because transcriptional kinetics is not normally distributed (Figure 1D, E), we use the non-parametric Wilcoxon test to examine the association between ASB of TFs and transcriptional kinetics. Specifically, we normalized the effect size from the Wilcoxon test to be within (-0.5,0.5). A positive Wilcoxon test effect size indicates that the existence of a regulator will increase the transcriptional kinetic and vice versa. We correct the test result for every TF and transcriptional kinetic for multiple testing with the False Discovery Rate(FDR). We select TFs whose association with any transcriptional kinetic is significant after FDR correction.

scRNA-seq Data Source and QC

We use two criteria to filter the scRNA-seq expression profile to remove abnormal cells and still maintain cell-cell variability. Firstly, we preserve cells with <20% of the transcript from mitochondrial genes. In addition, cells with fewer UMI than one standard deviation from the population mean are excluded.

The scRNA-seq data for GM12878 and GM18502 is from Osorio et al. (2019). For GM12878, 8,287 cells are sequenced, and 7,145 pass the quality control(QC). Using these 7,145 cells from GM12878, we obtain the valid estimation of transcriptional kinetics for 1,935 alleles corresponding to 968 genes. For GM18502, 5,991 cells are sequenced, and 5,029 pass the QC. we obtain the valid estimation of transcriptional kinetics for 1,480 genes with 2,959 alleles.

SOFTWARE AVAILABILITY

The code for this study is available on GitHub: https://github.com/bushlab-genomics/ASEkinetics with open access or can be found in the Supplemental code file.

CONFLICT OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work is supported by U01 AG058654 (Haines, Bush, Martin, Farrer, Pericak-Vance) and RF1 AG070935 (Griswold, Bush)

Figures are created with biorender.Com**.**

REFERENCE

Lammers, N.C., Kim, Y.J., Zhao, J., and Garcia, H.G. (2020). A matter of time: Using dynamics and theory to uncover mechanisms of transcriptional bursting. Current Opinion in Cell Biology 67, 147–157.

Ohnishi Y, Huber W, Tsumura A, Kang M, Xenopoulos P, Kurimoto K, Oleś AK, Araúzo-Bravo MJ, Saitou M, Hadjantonakis A-K, et al. 2014. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nat Cell Biol 16: 27–37.

Martinez-Jimenez CP, Eling N, Chen H-C, Vallejos CA, Kolodziejczyk AA, Connor F, Stojic L, Rayner TF, Stubbington MJT, Teichmann SA, et al. 2017. Aging increases cell-to-cell transcriptional variability upon immune stimulation. Science 355: 1433–1436.

Rodriguez J, Larson DR. 2020. Transcription in Living Cells: Molecular Mechanisms of Bursting. Annual Review of Biochemistry 89: 189–212.

Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, Simpson ML, Weinberger LS. 2012. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proceedings of the National Academy of Sciences 109: 17454–17459.

Sánchez Á, Kondev J. 2008. Transcriptional control of noise in gene expression. PNAS 105: 5081–5086.

Nicolas D, Zoller B, Suter DM, Naef F. 2018. Modulation of transcriptional burst frequency by histone acetylation. Proc Natl Bartman CR, Hamagami N, Keller CA, Giardine B, Hardison RC, Blobel GA, Raj A. 2019. Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. Molecular Cell 73: 519-532.e4.Acad Sci USA 115: 7153–7158.

Antolović V, Miermont A, Corrigan AM, Chubb JR. 2017. Generation of Single-Cell Transcript Variability by Repression. Current Biology 27: 1811-1817.e3.

Li C, Cesbron F, Oehler M, Brunner M, Höfer T. 2018. Frequency Modulation of Transcriptional Bursting Enables Sensitive and Rapid Gene Regulation. Cell Systems 6: 409-423.e11.

Faure AJ, Schmiedel JM, Lehner B. 2017. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. Cell Systems 5: 471-484.e4.

Fukaya T, Lim B, Levine M. 2016. Enhancer Control of Transcriptional Bursting. Cell 166: 358–368.

Singh A, Razooky B, Cox CD, Simpson ML, Weinberger LS. 2010. Transcriptional Bursting from the HIV-1 Promoter Is a Significant Source of Stochastic Noise in HIV-1 Gene Expression. Biophysical Journal 98: L32–L34.

Dobrinić P, Szczurek AT, Klose RJ. 2021. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. Nat Struct Mol Biol 28: 811–824.

Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, Segerstolpe Å, Rivera CM, Ren B, Sandberg R. 2019. Genomic encoding of transcriptional bursting kinetics. Nature 565: 251–254.

Martin V, Zhao J, Afek A, Mielko Z, Gordân R. 2019. QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants. Nucleic Acids Research 47: W127–W135.

de Santiago I, Liu W, Yuan K, O'Reilly M, Chilamakuri CSR, Ponder BAJ, Meyer KB, Markowetz F. 2017. BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. Genome Biol 18: 39.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Bressloff PC. Stochastic processes in cell biology. Vol. 41. Berlin: Springer, 2014.

Osorio, D., Yu, X., Yu, P., Serpedin, E., and Cai, J.J. (2019). Single-cell RNA sequencing of a European and an African lymphoblastoid cell line. Sci Data 6, 112.

Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, Gnirke A, Nusbaum C, Hacohen N, Friedman N, et al. 2011. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol 29: 436–442.

Shahrezaei V, Swain PS. 2008. Analytical distributions for stochastic gene expression. PNAS 105: 17256–17261.

Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, Pawitan Y. 2016. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics 32: 2128–2135.

Zoller B, Little SC, Gregor T. 2018. Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. Cell 175: 835-847.e25.

Gao T, Qian J. 2020. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research 48: D58–D64.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46: D794–D801.

Zhou W, Machiela MJ, Freedman ND, Rothman N, Malats N, Dagnall C, Caporaso N, Teras LT, Gaudet MM, Gapstur SM, et al. 2016. Mosaic loss of chromosome Y is associated with common variation near TCL1A. Nat Genet 48: 563–568.

Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, Faridani OR, Sandberg R. 2020. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat Biotechnol 38: 708–714.

Gao T, Qian J. 2020. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research 48: D58–D64.

Dobrinić P, Szczurek AT, Klose RJ. 2021. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. Nat Struct Mol Biol 28: 811–824.

Fukaya T, Lim B, Levine M. 2016. Enhancer Control of Transcriptional Bursting. Cell 166: 358–368.

Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samoviča M, Sakthivel MP, Kuzmin I, Trevanion SJ, et al. 2021. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet 53: 1290–1299.

Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, Kucera KS, Willard HF, Myers RM. 2011. Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. PLOS Genetics 7: e1002228.

Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH. 2011. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science 332: 475–478.

George L, Indig FE, Abdelmohsen K, Gorospe M. Intracellular RNA-tracking methods. Open Biology 8: 180104.

Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samoviča M, Sakthivel MP, Kuzmin I, Trevanion SJ, et al. 2021. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet 53: 1290–1299.

Boegel S, Löwer M, Bukur T, Sorn P, Castle JC, Sahin U. 2018. HLA and proteasome expression body map. BMC Medical Genomics 11: 36.

Gupta A, Martin-Rufino JD, Jones TR, Subramanian V, Qiu X, Grody EI, Bloemendal A, Weng C, Niu S-Y, Min KH, et al. 2022. Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. Proc Natl Acad Sci U S A 119: e2207392119.

Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, Nestadt G, Kenny EE, Vacic V, Peters I, et al. 2015. Genome-wide association study of schizophrenia in Ashkenazi Jews. Am J Med Genet B Neuropsychiatr Genet 168: 649–659.