Open access • Posted Content • DOI:10.1101/694869

STARRPeaker: Uniform processing and accurate identification of whole human STARR-seq active regions — Source link

Dong-Hoon Lee, Manman Shi, Jennifer R. Moran, Martha Wall ...+8 more authors

Institutions: Yale University, University of Chicago, Westlake University

Published on: 08 Jul 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: STARR-seq

STARRPeaker: Uniform processing and accurate identification of STARR-seq active regions

Donghoon Lee[1,2], Manman Shi[3], Jennifer Moran[3], Martha Wall[3], Jing Zhang[1,2], Jason Liu[2], Dominic Fitzgerald[3], Yasuhiro Kyono[3], Lijia Ma[3,4], Kevin P White[3,5]*, Mark Gerstein[1,2,6,7]*

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[3] Institute for Genomics and System Biology, University of Chicago, Chicago, IL 60637, USA

[4] School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China

[5] Tempus Labs, Inc., Chicago, IL 60654, USA

[6] Department of Computer Science, Yale University, New Haven, CT 06520, USA

[7] Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

* Corresponding authors

E-mail: pi@gersteinlab.org

Abstract

Background: High-throughput reporter assays, such as self-transcribing active regulatory region sequencing (STARR-seq), allow for unbiased and quantitative assessment of enhancers at a genome-wide scale. Recent advances in STARR-seq technology have employed progressively more complex genomic libraries and increased sequencing depths, to assay larger sized regions, up to the entire human genome. These advances necessitate a reliable processing pipeline and peak-calling algorithm.

Results: Most STARR-seq studies have relied on chromatin immunoprecipitation sequencing (ChIP-seq) processing pipelines. However, there are key differences in STARR-seq versus ChIP-seq. First, STARR-seq uses transcribed RNA to measure the activity of an enhancer, making an accurate determination of the basal transcription rate important. Second, STARR-seq coverage is highly non-uniform, overdispersed, and often confounded by sequencing biases, such as GC content and mappability. Lastly, here, we observed a clear correlation between RNA thermodynamic stability and STARR-seq readout, suggesting that STARR-seq may be sensitive to RNA secondary structure and stability. Considering these findings, we developed a negative-binomial regression framework for uniformly processing STARR-seq data, called STARRPeaker. In support of this, we generated whole-genome STARR-seq data from the HepG2 and K562 human cell lines and applied STARRPeaker to call enhancers.

Conclusions: We show STARRPeaker can unbiasedly detect active enhancers from both captured and whole-genome STARR-seq data. Specifically, we report ~33,000 and ~20,000 candidate enhancers from HepG2 and K562, respectively. Moreover, we show that STARRPeaker outperforms other peak callers in terms of identifying known

47 enhancers with fewer false positives. Overall, we demonstrate an optimized processing
48 framework for STARR-seq experiments can identify putative enhancers while
49 addressing potential confounders.

50 **Keywords**: STARR-seq, peak caller, enhancer, non-coding, regulatory element

51

52 Background

53 The transcription of eukaryotic genes is precisely coordinated by an interplay between
54 *cis*-regulatory elements. For example, enhancers and promoters serve as binding
55 platforms for transcription factors (TFs) and allow them to interact with each other via
56 three-dimensional looping of chromatin. Their interactions are often required to initiate
57 transcription [1,2]. Enhancers, which are often distant from the transcribed gene body
58 itself, play critical roles in the upregulation of gene transcription. Enhancers are cell-type
59 specific and can be epigenetically activated or silenced to modulate transcriptional
60 dynamics over the course of development. Enhancers can be found upstream or
61 downstream of genes, or even within introns [3–5]. They function independent of their
62 orientation, do not necessarily regulate the closest genes, and sometimes regulate
63 multiple genes at once [6,7]. In addition, several recent studies have demonstrated that
64 some promoters – termed E-promoters – may act as enhancers of distal genes [8,9].

65

66 Consensus sequences (or canonical sequences) have been identified at certain protein
67 binding sites, splice sites, and boundaries of protein-coding genes. However, there are
68 no known consensus sequences that characterize enhancer function, making it
69 challenging to identify enhancers based on sequence alone in an unbiased fashion. The

non-coding territory occupies over 98% of the genome landscape, making the search space very broad. Moreover, the activity of enhancers depends on the physiological condition and epigenetic landscape of the cellular environment, complicating a fair assessment of enhancer function.

Previously, putative regulatory elements were computationally predicted, indirectly, by profiling DNA accessibility (using DNase-seq, FAIRE-seq, or ATAC-seq) as well as histone modifications (ChIP-seq) that are linked to regulatory functions [10–12]. More recently, researchers have developed high-throughput episomal (exogenous) reporter assays to directly measure enhancer activity across the whole genome, specifically massively parallel reporter assays (MPRA) [13,14] and self-transcribing active regulatory region sequencing (STARR-seq) [15,16]. These assays allow for quantitative assessment of enhancer activity in a high-throughput fashion.

In STARR-seq, candidate DNA fragments are cloned downstream of a reporter gene into the 3′ untranslated region (UTR). After transfecting the plasmid pool into host cells, one can measure the regulatory potential by high-throughput sequencing of the 3′ UTR of the expressed reporter gene mRNA. These exogenous reporters enable accurate and unbiased assessment of enhancer activity at the whole-genome level, independent of chromatin context. Unlike MPRA – which utilizes barcodes – STARR-seq produces self-transcribed RNA fragments that can be directly mapped onto the genome (we call this STARR-seq output hereafter). The activities of enhancers are measured by comparing the amount of RNA produced from the relative amount of genomic DNA in the STARR-seq library (we call this STARR-seq input hereafter). STARR-seq has several technical advantages over MPRA. Library construction is relatively simple

because barcodes are not needed. In addition, candidate enhancers are cloned instead of synthesized, allowing the assay to test extended sequence contexts (>500 bp) for enhancer activity, which studies have shown to be critical for functional activity [17]. Importantly, STARR-seq can be scaled to the whole-genome level for unbiased scanning of functional activities. However, scaling STARR-seq to the human genome is still very challenging, primarily due to its massive size. A more complex genomic DNA library, a higher sequencing depth, and increased transfection efficiency are required to cover the whole human genome [16], which could ultimately introduce biases. Furthermore, inserting a large fragment of DNA into the 3' UTR of the reporter gene could inadvertently introduce regulatory sequences that might affect mRNA abundance and stability, which could lead to both false positives and false negatives. MPRA is more robust in this regard because the activity of each candidate enhancer is quantified by multiple molecular barcodes associated with the fragment, making it less prone to such artifacts than STARR-seq.

The processing of STARR-seq data is somewhat similar to that of ChIP-seq, where protein-crosslinked DNA is immunoprecipitated and sequenced. A typical ChIP-seq processing pipeline identifies genomic regions over-represented by sequencing tags in a ChIP sample compared to a control sample. STARR-seq data is compatible with most ChIP-seq peak callers. Hence, previous studies on STARR-seq have largely relied on peak-calling software developed for ChIP-seq such as MACS2 [16,18,19]. However, one must be cautious using ChIP-seq peak callers, at least without re-tuning the default parameters optimized for processing TF ChIP-seq [20].

In this paper, we describe key differences in the processing of STARR-seq versus ChIP-seq data. Due to increased complexity of the genomic screening library and sequencing depth requirements, STARR-seq coverage is highly non-uniform. This leads to a lower signal-to-noise ratio than a typical ChIP-seq experiment and makes estimating the background model more challenging, which could ultimately lead to false-positive peaks. In addition, STARR-seq measures more of a continuous activity, similar to quantification in RNA-seq, than a discrete binding event. Therefore, STARR-seq peaks should be further evaluated using a notion of activity score. These differences necessitate a unique approach to processing STARR-seq data.

We propose an algorithm optimized for processing and identifying functionally active enhancers from STARR-seq data, which we call STARRPeaker. This approach statistically models the basal level of transcription, accounting for potential confounding factors, and accurately identifies reproducible enhancers. We applied our method to two whole human STARR-seq datasets and evaluated its performance against previous methods. We also compared an R package, BasicSTARRseq, developed to process peaks from the first STARR-seq data [15], which models enrichment of sequencing reads using a binomial distribution. We benchmarked our peak calls against known human enhancers. Thus, our findings support that STARRPeaker will be a useful tool for uniformly processing STARR-seq data.

Results and Discussion

Precise measurement of STARR-seq coverage

We binned the genome using a sliding window of length, *l*, and step size, *s*. Based on the average size of the STARR-seq library, we defined a 500 bp window length with a 100 bp step size to be the default parameter. Based on the generated genomic bins, we calculated the coverage of both STARR-seq input and output mapped to each bin. For calculating the sequence coverage, other peak callers and many visualization tools commonly use the start position of the read [15,21,22]. However, given that the average size of the fragments inserted into the STARR-seq libraries were approximately 500 bp, we expected that the read coverage using the read start position may shift the estimate of the summit of signal and dilute the enrichment. Some peak callers have used read densities of forward and reverse strands separately to overcome this issue [23,24]. To precisely measure the coverage of STARR-seq input and output, we first inferred the size of the fragment insert from paired-end reads and used the center of the fragment insert, instead of start position of the read, to calculate coverage. For inferring the size of the fragment insert, we first strictly filtered out reads that were not properly paired and chimeric. Chimeric alignments are reads that cannot be linearly aligned to a reference genome, implying a potential discrepancy between the sequenced genome and the reference genome and indicative of a structural variation or a PCR artifact [25]. We also filtered out read pairs that had a fragment insert size greater than l_{max} and less than l_{min}. By default, we filtered out fragment insert sizes less than 200 bp and greater than 1,000 bp. After filtering out spurious read-pairs, we estimated the center of the fragment insert and counted the fragment depth for each genomic bin. To assess the benefit of using fragment-based coverage, we compared the coverage calculated using the center of fragment insert to an alternate model using the start position of the sequencing read.

We found that the position of the peaks shifted up approximately 200 bp when we used the alternate model (Figure 1A, Supplementary Figure 1A). Such a shift caused by the read-based coverage could lead to the omission of TF binding sites located at the boundary. Moreover, we observed that the read-based coverage diluted the overall STARR-seq signal; as a result, peaks calculated based on the alternate model had lower fold enrichment and were less confident and broader in size (Figure 1B-D, Supplementary Figure 1B-D). Overall, the fragment-based coverage offered more concentrated and robust peak signal compared to the read-based coverage counting scheme. The benefit of using the center of the fragment is highlighted in Figure 1E, where we find more concise and precise peak with a higher fold enrichment using fragment-based coverage.

Controlling for potential systemic bias in the STARR-seq assay

To unbiasedly test for the regulatory activity, a model needs to control for potential systemic biases inherent to generating STARR-seq data. STARR-seq measures the ratio of transcribed RNA to DNA for a given test region and determines whether the test region can facilitate transcription at a higher rate than the basal level. This is based on the assumption that (1) the basal transcriptional level stays relatively constant across the genome and (2) the transcriptional rate is a reflection of the regulatory activity of the DNA insert. However, these assumptions may not always be true, and one needs to consider potential systemic biases that can interfere with the quantification of regulatory activity when analyzing the data.

We next tested whether potential sequencing biases and other covariates confounded STARR-seq readouts (Figure 2). We found that STARR-seq RNA coverage was significantly correlated with GC content (PCC 0.61; P-val 1E-299) and mappability (PCC 0.45; P-val 2.9E-148). This could be attributed to intrinsic sequencing biases in library preparation. A genome-wide reporter library is made from randomly sheared genomic DNA, but DNA fragmentation is often non-random [26]. Studies also have suggested that epigenetic mechanisms and CpG methylation may influence fragmentation [27]. Furthermore, the isolated polyadenylated RNAs are reverse transcribed and PCR amplified before sequencing, and this process can further confound the sequenced candidate fragments.

Notably, we found that STARR-seq coverage was also significantly confounded by RNA thermodynamic stability (PCC -0.55; P-val 0). Unlike ChIP-seq, where both the experiment and input controls derive from the same DNA origin, STARR-seq experiments measure the regulatory potential from the abundance of transcribed RNA, which adds a layer of complexity. For example, RNA structure and co-transcriptional folding might potentially influence the readout of STARR-seq experiments [28]. Single-stranded RNA starts to fold upon transcription and the resulting RNA structure might influence the measurement of regulatory activity. Previously, researchers suggested a potential linkage between RNA secondary structure and transcriptional regulation [29]. In addition, the resulting transcribed RNA undergoes a series of post-transcriptional regulation, and RNA stability might play a critical role. Moreover, previous reports have shown that the degradation rates – the main determinant of cellular RNA levels [30] –

vary significantly across the genome and that RNA stability correlates with functionality [31,32].

Based on these findings, we built a regression-based model that accounts for various confounding variables of test sequence fragments to unbiasedly identify potential enhancer regions from STARR-seq data. Note that many of the covariates have appreciable correlation with each other. However, we did find, using stepwise forward selection, that each of them contributes substantially and independently to the model fit as assessed by Akaike information criterion (AIC) and Bayesian information criterion (BIC) (Supplementary Figure 2).

Accurate modelling of STARR-seq using negative binomial regression

To model the fragment coverage data from STARR-seq using discrete probability distribution, we assumed that each genomic bin is independent and identically distributed, as specified in the Bernoulli trials [33]. That is, each test fragment can only map to a single fixed-length bin. Therefore, we only considered a non-overlapping subset of bins for modeling and fitting the distribution. We also excluded bins not covered by any genomic input or those in which the normalized input coverage was less than a minimum quantile t_{min}, since these regions do not have sufficient power to detect enrichment. We selected the bin size and the minimum coverage based on the experimental design of STARR-seq. We simulated and fitted various discrete probability distributions to STARR-seq output coverage. We observed that the STARR-seq output coverage data was overdispersed and fit the best with a negative binomial distribution

(Figure 3A). Moreover, a Q-Q plot of simulated coverage further demonstrated that the negative binomial model provides the best fit for the data (Figure 3B).

We observed a slight negative enrichment in the STARR-seq output coverage, suggesting that some candidate fragments can repress the basal transcriptional activity. However, these regions may contain sequences that can destabilize mRNAs. Therefore, additional experiments are necessary to demonstrate that STARR-seq can reliably detect silencers. In the meantime, we suggest opting for a system specifically designed for identifying silencers for this task [34].

Peak-calling algorithm

To accurately model the ratio of STARR-seq output fragment coverage (RNA) to input fragment coverage (DNA) while controlling for potential confounding factors, we applied a negative binomial regression. The overview of our model is outlined in Figure 4. Our model starts by fitting an analytical distribution to the observed fragment coverage across fixed non-overlapping genomic bins. In doing so, we use covariates to model expected counts in the form of multiple regression. Subsequently, once a model is fitted, we evaluate the likelihood of obtaining the observed fragment counts and assign p-values using the null negative binomial distribution. In this testing phase, we use flexible genomic bins with a sliding window in order to find enrichment peaks at a higher resolution. Genomic bins with significant enrichments are selected based on their adjusted p-values using multiple testing correction. Finally, peak locations are fine-tuned to the summit of the direct fragment coverage. Note that the adjusted p-value should be

regarded as the likelihood of a candidate region being an enhancer while the fold enrichment can be directly interpreted as a quantitative measure of enhancer activity.

Let Y be a vector of STARR-seq output (RNA) coverage, then y_i for $1 \leq i \leq n$ denotes the number of RNA fragments from a STARR-seq experiment mapped to the i-th bin from the total of n genomic bins. Let t_i be the number of input library fragments (DNA) mapped to the i-th bin. We define X to be the matrix of covariates, where $\overrightarrow{x_i}$ is the vector of covariates corresponding to the i-th bin and x_{ij} is the j-th covariate for the i-th bin.

Negative binomial distribution

A negative binomial distribution, which arises from a Gamma-Poisson mixture, can be parametrized as follows [35–37] (see Methods for derivation).

$$f_Y(y_i|\mu_i,\theta) = \frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\cdot\Gamma(\theta)}\cdot\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\cdot\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i} \quad (1)$$

A negative binomial is a generalization of a Poisson regression that allows the variance to be different from the mean, shaped by the dispersion parameter θ. There are two alternative forms of parametrization for a negative binomial – NB1 and NB2 – which were first introduced by Cameron and Trivedi [36]. The difference between NB1 and NB2 is in the conditional variance of y_i. Assuming y_i has mean λ_i, the general variance function follows the form $\omega_i = \lambda_i + \alpha\lambda_i^{\,p}$, where α is a scalar parameter. NB1 uses $p = 1$, whereas NB2 uses the quadratic form of variance with $p = 2$. We use the most common

implementation of the negative binomial, NB2, hereafter. The variance for the NB2 model is given as

$$\sigma^2 = \mu + \frac{\mu^2}{\theta} \qquad \text{(Error! Bookmark not defined.)}$$

We assume that the majority of genomic bins will have a basal level of transcription, the expected fragment counts at each i-th bin, $E(y_i)$, represents the mean incidence, μ_i, and the count of RNA fragments Y follows the traditional negative binomial (NB2) distribution.

$$\begin{aligned} E(y_i) &= \mu_i \\ Y &\sim NB(\mu, \theta) \end{aligned} \qquad (2)$$

Negative binomial regression model

The regression term for the expected RNA fragment count can be expressed in terms of a linear combination of explanatory variables, a set of m covariates ($\vec{x}$). We use the input library variable t_i as one covariate. For simplicity, we denote t_i as x_{0i} hereafter.

$$\begin{aligned} \ln \mu_i &= \beta_0 x_{0i} + \beta_1 x_{1i} + \cdots + \beta_m x_{mi} \\ \mu_i &= \exp(\beta_0 x_{0i} + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}) \\ \mu_i &= \exp(\vec{x_i}^{\mathrm{T}} \beta) \end{aligned} \qquad (3)$$

Alternatively, instead of using the input library variable t_i as one covariate, we can directly use it as an offset variable. Generally, a fractional observation cannot be modeled using discrete probability. However, an offset variable in a generalized linear model can be used to correct the response term to behave like a fraction. One advantage of using the input variable as an "exposure" to the RNA output coverage is that it allows us to directly model the basal transcription rate (the ratio of RNA to DNA) as a rate response variable. More details on this alternative parametrization are included in the Methods section. In our STARRPeaker model, we used four covariates; fragment coverage of input genomic libraries, GC content, mappability, and the thermodynamic stability of genomic libraries.

Maximum-likelihood estimation

We fit the model and estimate regression coefficients using the maximum likelihood method, where log-likelihood function is shown as follows.

$$\mathcal{L}_{NB}(\mu|y,\theta) = \sum_{i=1}^{n} y_i \ln\left(\frac{\mu_i}{\theta+\mu_i}\right) + \theta \ln\left(\frac{\theta}{\theta+\mu_i}\right) + \ln\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\cdot\Gamma(\theta)}\right) \quad \text{(Error! Bookmark not defined.)}$$

Substituting μ_i with the regression term, the log-likelihood function can be parametrized in terms of regression coefficients, β.

$$\mathcal{L}_{NB}(\beta|y,\theta) = \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\vec{x_i}^T\beta}}{\theta + e^{\vec{x_i}^T\beta}}\right) + \theta \ln\left(\frac{\theta}{\theta + e^{\vec{x_i}^T\beta}}\right) + \ln\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\cdot\Gamma(\theta)}\right) \quad \text{(Error! Bookmark not defined.)}$$

We can determine the maximum likelihood estimates of the model parameters by setting the first derivative of the log-likelihood with respect to β, the gradient, to zero, and there is no analytical solution for $\hat{\beta}$. Numerically, we iteratively solve for the regression coefficients β and the dispersion parameter θ, alternatively, until both parameters converge.

Estimation of P-value

The P-value is defined as the probability of observing equal or more extreme value than the observed value at the i-th bin, y_i, under the null hypothesis.

$$P\text{-}value_i = \Pr(Y \geq y_i|H) \quad (4)$$

As defined earlier, we assume the random variable Y comes from a negative binomial distribution with the fitted mean at the i-th bin, μ_i, as the expected value, and θ as the dispersion parameter. Then, we can estimate the P-value from the cumulative distribution function CDF, which is the sum of the probability mass function f_Y from 0 to $y_i - 1$.

$$\Pr(Y \geq y_i|H) = 1 - CDF(y_i - 1) = 1 - \sum_{k=0}^{y_i-1} f_Y(k|\mu_i, \theta) \quad (5)$$

322

323 Substituting (1) gives

324

$$P\text{-}value_i = 1 - \sum_{k=0}^{y_i-1} \frac{\Gamma(k+\theta)}{\Gamma(k+1)\cdot\Gamma(\theta)} \cdot \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} \cdot \left(\frac{\mu_i}{\theta+\mu_i}\right)^{k} \quad (6)$$

325

326 Finally, we calculate the false discovery rate using *Benjamini & Hochberg* method [38].

327

328 **Application of STARRPeaker**

329 We applied STARRPeaker to two whole human genome STARR-seq experiments,
330 K562 and HepG2, utilizing origin of replication (ORI)-based plasmids [39]. Based on
331 peaks identified from these datasets, we evaluated the quality and characteristics of the
332 identified enhancers as well as the performance of the peak caller by comparing to
333 external enhancer resources.

334

335 *Initial evaluation of STARRPeaker enhancers*

336 We processed two biological replicates from each cell type independently and assessed
337 the correlation between each pair. Overall, we observed high correlation between two
338 replicates (PCC=0.99 for both HepG2 and K562; see Supplementary Figure 3). By
339 intersecting peaks from two replicates, we identified 32,929 and 20,471 reproducible
340 candidate enhancers from HepG2 and K562, respectively (Supplementary Table S1).

Although the total number of peaks varied between HepG2 and K562, we observed a comparable number of peaks within the accessible region of the genome. We found 12,019 (36.34%) and 11,420 (55.57%) candidate enhancers from HepG2 and K562, respectively, within the open chromatin defined by ENCODE DNase-seq hotspots. Consistent with previous findings [39], a substantial fraction of candidate enhancers was epigenetically silenced at the chromatin level. However, as demonstrated previously using a histone deacetylase inhibitor (HDAC) [16], these poised enhancers can become functional under a more transcriptionally permissive environment. Therefore, episomal reporter assays like STARR-seq have the unique advantage of detecting potential enhancer activity independent from chromatin context. We would like to note that it is important to identify poised enhancers located in heterochromatic regions of the genome, which could become functional during developmental or pathological time courses.

Assessment of robustness and reproducibility of the method

A reliable peak-calling method should be able to identify peaks from suboptimal datasets. To evaluate the robustness of STARRPeaker, we used subsets of the whole-genome STARR-seq library to call peaks and compared the results. We subsampled randomly at various rates from 20 to 80% of the total dataset and compared the quality of peaks. We found that STARRPeaker was able to reliably identify the peaks using approximately 60% of the original sequencing library (Supplementary Figure 4). However, the quality of the peak calls started to deteriorate when 40% or less were used.

Evaluation of potential orientation bias in candidate enhancers

In general, enhancers are thought to function independent of orientation [40]. However, the fragment counts in one orientation could be skewed over the other due to orientation-specific activities, PCR, or sequencing artifacts. To test for potential orientation-based biases, we ran a binomial test on the candidate enhancers we identified. We observed a small fraction of candidate enhancers showing strand bias [3.19% for HepG2 rep1 (n=1,605); 3.76% for HepG2 rep2 (n=1,991); 7.77% for K562 rep1 (n=2,347); 5.25% for K562 rep2 (n=2,195); FDR ≤ 0.01] (Supplementary Figure 5). Less than one third of the enhancers (n=690) showed strand-specific activity in both replicates. Thus, we conclude that there is insufficient evidence to show that orientation-dependent biases are present in our STARR-seq data. Furthermore, this finding provides further support that enhancers function independent of orientation.

Performance comparison to other peak-calling algorithms

We evaluated the performance of STARRPeaker by comparing it to previously used methods, namely BasicSTARRseq and MACS2.

First, we qualitatively assessed the peak-calling algorithms using a simulated dataset where the ground truth exists. We created a STARR-seq dataset that consists of four spike-in controls (hybrid of DNA input library and RNA output library of known specific location). All three methods successfully identified the four control peaks with high confidence (Supplementary Figure 6). However, we noticed that BasicSTARRseq peaks were fragmented due to its limitation of fixed peak size. Moreover, the peaks were

shifted toward the enrichment of sequencing reads. Furthermore, BasicSTARRseq identified a false-positive peak, and as a result, identified a total of eight regions instead of four.

Second, we quantitatively assessed the peak-calling algorithms using the whole human genome STARR-seq dataset. After uniformly calling peaks from each method using the recommended default settings, we evaluated the quality of the candidate enhancers identified. We found that both BasicSTARRseq and MACS2 called significantly more peaks (4 to 20-fold higher) than STARRPeaker (Supplementary Table S4). While it is uncertain how many true enhancers were present in each sample, we had to ensure that we made a fair comparison across different methods due to the tradeoff between sensitivity and specificity. An increase in sensitivity is generally achieved at the expense of a decrease in specificity, as described in receiver operating characteristic curves. In our context, a method having higher specificity suffers from having less overlap with open chromatin and previously identified enhancers from other assays. Therefore, we used a uniform P-value threshold of 0.001 and subsampled the peaks before the comparison. After uniformly processing the dataset using each method, we measured the level of epigenetic profile enrichment around the peaks. We observed higher enrichment of DNase-hypersensitive sites, as well as more distinct double-peak patterns of H3K27ac and H3K4me1, using STARRPeaker compared to BasicSTARRseq or MACS2 (Figure 5, Supplementary Figure 7). Furthermore, STARRPeaker peaks had significantly higher enrichment of TF binding events (based on the number of TF ChIP-seq binding sites) compared to the peaks identified using other methods.

Comparison to previously characterized enhancers

First, we compared the peaks identified by STARRPeaker to previously characterized enhancers from HepG2 or K562 cell lines by CAGE [41], MPRA [17,42], and STARR-seq [19] (Figure 6, Supplementary Table S2). Overall, we observed a higher fraction of STARRPeaker peaks overlapping with external datasets compare to other methods. Moreover, we found higher overlaps when peaks from both replicates were merged, due to fewer but more precise candidate enhancers from merging replicates. However, we noticed reduced agreement across different types of enhancer assays. Low overlap between assays may arise from different formats or layouts of reporter plasmids, such as differing enhancer cloning sites or promoters, or differences in the complexity of the screening library. Furthermore, CAGE is an entirely different assay from episomal reporter assays like MPRA and STARR-seq, with enhancers defined based on bidirectional transcripts originating from an eRNA.

Second, we examined the nine distal enhancers from the GATA1 and MYC loci characterized in-depth by CRISPRi tiling screen (Supplementary Figure 8). We found that STARRPeaker accurately called peaks for 6 of 9 enhancers from both replicates. For the remaining three regions, we observed insufficient enrichment of STARR-seq output and, therefore, we concluded that this is not a shortcoming of the peak caller.

Application to external STARR-seq datasets

To ensure that STARRPeaker can be generally applied to different variants of STARR-seq assays, we tested STARRPeaker on previously published STARR-seq datasets.

First, we applied STARRPeaker to the whole-genome ORI-STARR-seq dataset on HeLa-S3 [39] and assessed the quality of the peaks identified. Consistent with the previous claim that IFN-I signaling may induce false-positive enhancers, we identified more peaks in untreated HeLa-S3 samples (n=28,381) compared to inhibitor-treated samples (n=16,150). Furthermore, peaks from untreated samples had lower enrichment of chromatin accessibility (DNase-seq) than those from inhibitor-treated samples, supporting that TBK1/IKK/PKR inhibition reduces false-positive enhancer signals related to IFN-I signaling (Supplementary Figure 9A). Moreover, STARRPeaker covered 77.5% (n=7,451) of published peaks, which were called using BasicSTARRseq and then further shortlisted using a stringent threshold (P-value 1E-5 with corrected enrichment ≥ 4). Furthermore, STARRPeaker found 6,540 additional peaks from a HeLa-S3 sample that was highly enriched with chromatin accessibility signals (Supplementary Figure 9B). Second, we tested if STARRPeaker can be reliably applied to captured STARR-seq datasets (Cap-STARR-seq). We applied STARRPeaker to a previously characterized GM12878 STARR-seq dataset based on an ATAC-seq-capture technique called HiDRA [43] and compared its performance with published results. The HiDRA dataset was reported to have ~65,000 regions with enhancer function. In the STARRPeaker run, we identified only 20,852 regions with significant enhancer activities from the five replicates they produced. Approximately 73.6% of peaks overlapped with the published results (n=15,347). While it is debatable to claim that one method is superior to the other, this result demonstrates that STARRPeaker can be reliably used against the Cap-STARR-seq dataset.

Third, we further evaluated the performance of the peak-calling methods by applying STARRPeaker and two other peak-calling methods to another published Cap-STARR-seq dataset [19]. The dataset covers approximately 91% of the surrounding 3 Mb of the MYC locus. Consistent with the earlier analysis, we observed that STARRPeaker is highly specific and identifies fewer candidate enhancers (n=26) compared to the other methods (BasicSTARRseq n=223; MACS2 n=136). Furthermore, a four-way comparison (STARRPeaker, BasicSTARRseq, MACS2, and published peaks) showed that all of the STARRPeaker peaks overlapped with peaks from other methods but not the other way around (Supplementary Figure 10). These results indicate that STARRPeaker is more robust and reliable at identifying reproducible candidate enhancers from various STARR-seq datasets than previous methods.

Conclusions

In summary, we developed a reliable peak-calling analysis pipeline named STARRPeaker that is optimized for large-scale STARR-seq experiments. To illustrate the utility of our method, we applied it to two whole human genome STARR-seq datasets from K562 and HepG2 cell lines, utilizing ORI-based plasmids. STARRPeaker has several key improvements over previous approaches including (1) precise and efficient calculation of fragment coverage; (2) accurate modeling of the basal transcription rate using negative binomial regression; and (3) accounting for potential confounding factors, such as GC content, mappability, and the thermodynamic stability of genomic libraries. We demonstrate the superiority of our method over

previously used peak callers, supported by strong enrichment of epigenetic marks relevant to enhancers and overlap with previously known enhancers.

To fully understand how noncoding regulatory elements can modulate transcriptional programs in human, STARR-seq active regions must be further characterized and validated within different cellular contexts. For example, recent applications of CRISPR-dCas9 to genome editing have allowed researchers to epigenetically perturb and test these elements in their native genomic context [44,45]. The next step for CRISPR-based functional screens is to overcome the current limitation of small scale by leveraging barcodes and single-cell sequencing technology [46]. In the meantime, we envision that the STARRPeaker framework could be utilized to detect and quantify enhancers at the whole-genome level, thereby aiding in prioritizing candidate regions in an unbiased fashion to maximize functional characterization efforts.

Methods

Cell culture

We cultured K562 cells (ATCC) in IMDM (Gibco #12440) supplemented with 10% fetal bovine serum (FBS) and 1% pen/strep and maintained in a humidified chamber at 37°C with 5% CO_2. We cultured HepG2 cells (ATCC) in EMEM (ATCC #30-2003) supplemented with 10% FBS and 1% pen/strep, maintained in a humidified chamber at 37°C with 5% CO_2.

Generating an ORI-STARR-seq input plasmid library

We sonicated human male genomic DNA (Promega #G1471) using a Covaris S220 sonicator (duty factor – 5%; cycle per burst – 200; 40 sec) and ran it on a 0.8% agarose gel to size-select 500 bp fragments. After gel purification using a MinElute Gel Extraction kit (Qiagen), we end-repaired, ligated custom adaptors, and PCR-amplified DNA fragments using Q5 Hot Start High-Fidelity DNA polymerase (NEB) (98°C for 30 sec; 10 cycles of 98°C for 10 sec, 65°C for 30 sec, and 72°C for 30 sec; 72°C for 2 min) to add homology arms for Gibson assembly cloning.

We used AgeI-HF (NEB) and SalI-HF (NEB) to linearize the hSTARR-seq_ORI plasmid (gift from Alexander Stark; Addgene plasmid #99296) and cloned the PCR products into the vector using Gibson Assembly Master Mix (NEB); we set up 60 replicate reactions to maintain complexity. We purified the assembly reactions using SPRI beads (Beckman Coulter), dialyzed them using Slide-A-Lyzer MINI dialysis devices (ThermoScientific), and concentrated them using an Amicon Ultra-0.5 device (Amicon). We transformed the reaction into MegaX DH10BTM T1 electrocompetent cells (Thermo Fisher Scientific) (with 25 replicate transformations to maintain complexity) and let them grow in 12.5L LB-Amp medium until they reached an optical density of ~1.0. We extracted the plasmids using a Plasmid Gigaprep Kit (Qiagen) and dialyzed the plasmid prep using Slide-A-Lyzer MINI dialysis devices before electroporation.

Electroporation-mediated transfection of ORI-STARR-seq input plasmid library into K562 and HepG2 cell lines

We electroporated the ORI-STARR-seq library using an AgilePulse Max (Harvard Apparatus) and generated two biological replicates for each cell line. For K562 cells, we

electroporated 5.6 mg of input plasmid library into 700 million cells per biological replicate by delivering three 500 V pulses (1 ms duration with a 20 ms interval). For HepG2 cells, we electroporated 8 mg of input plasmid library into one billion cells in one replicate, and 5.6 mg into 700 million cells in another replicate by delivering three 300 V pulses (5 ms duration with a 20 ms interval).

Generation of an Illumina sequencing library

Output RNA library: We harvested cells 24 hr after electroporation, and extracted total RNA using an RNeasy Maxi kit (Qiagen). We further isolated polyA-plus mRNA using Dynabeads® Oligo (dT) kit (ThermoFisher Scientific), treated it with TURBO DNase (Invitrogen), and purified the reaction using an RNeasy MinElute Kit (Qiagen). We synthesized cDNA using SuperScript III (ThermoFisher Scientific) with a custom primer that specifically recognizes mRNAs that had been transcribed from the ORI-STARR-seq library. After reverse transcription, we treated the reactions with a cocktail of RNase A and RNase T1 (ThermoFisher Scientific). We split cDNA samples into 160 replicate sub-reactions, and PCR-amplified each sub-reaction with a primer with a unique index (helping to identify PCR duplicates) using Q5 Hot Start High-Fidelity DNA polymerase (NEB) with the following program: 98°C for 30 s; cycles of 98°C for 10 s, 65°C for 30 s, 72°C for 30 s (until they reached mid-log amplification phase; we cycled 18 cycles for K562 Rep.1; 16 cycles for K562 Rep. 2; 18 cycles for HepG2 Rep. 1; and 15 cycles for HepG2 Rep2); 72°C for 2 min). After PCR, we re-combined all sub-reactions into one and purified it with Agencourt Beads. We generated 100 bp paired-end reads for each

biological replicate on an Illumina Hiseq4000 at the University of Chicago Genome Facility.

Input DNA library: We PCR-amplified a total of 200 ng of input plasmid library (in 16 replicate reactions) using Q5 Hot Start High-Fidelity DNA polymerase (NEB) with the following program: 98°C for 30 s; 4 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 20 s; 8 cycles of 98°C for 10 s and 72°C for 50 s; 72°C for 2 min). After PCR, we combined all products into one and purified it with Agencourt Beads. We generated 100 bp paired-end reads on an Illumina Hiseq4000 at the University of Chicago Genome Facility.

Sequencing and preprocessing

For each of 160 replicates, paired-end sequencing reads were aligned to the human reference genome GRCh38 downloaded from the ENCODE portal (ENCSR425FOI) using BWA-mem (v0.7.17). Alignments were filtered against unmapped, secondary alignments, mapping quality score less than 30, and PCR duplicates using SAMtools (v1.9) and Picard (v2.9.0). All of the replicates were pooled and sorted for downstream analysis.

Negative binomial distribution

A negative binomial distribution, which arises from Gamma-Poisson mixture, can be parametrized for y>=0 as follows.

$$Pr(Y = y_i|\mu_i, \theta) = f_Y(y_i; \mu_i, \theta) = \binom{y_i + \theta - 1}{y_i} \cdot \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta} \cdot \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i}$$

where

$$\binom{y_i + \theta - 1}{y_i} = \frac{\Gamma(y_i + \theta)}{y_i! \cdot \Gamma(\theta)} = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)}$$

Substituting gives:

$$f_Y(y_i; \mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)} \cdot \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta} \cdot \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i}$$

Rearranging gives:

$$f_Y(y_i; \mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)} \cdot \left(\frac{1}{1 + \frac{\mu_i}{\theta}}\right)^{\theta} \cdot \left(\frac{\frac{\mu_i}{\theta}}{1 + \frac{\mu_i}{\theta}}\right)^{y_i}$$

$$f_Y(y_i; \theta, \mu_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)} \cdot \left(\frac{\mu_i}{\theta}\right)^{y_i} \left(\frac{1}{1 + \frac{\mu_i}{\theta}}\right)^{\theta + y_i}$$

$$f_Y(y_i; \theta, \mu_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)} \cdot \left(\frac{\mu_i}{\theta}\right)^{y_i} \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta + y_i}$$

$$f_Y(y_i; \theta, \mu_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)} \cdot \frac{{\mu_i}^{y_i} \theta^{\theta}}{(\theta + \mu_i)^{\theta + y_i}}$$

Alternative parametrization of negative binomial regression using a rate model

Alternative parametrization allows STARR-seq data to be modelled as a rate model. In contrast to using input coverage as one of the covariates, we can consider it as "exposure" to output coverage. This "trick" allows us to directly model the basal

578 transcription rate (the ratio of RNA to DNA) as a rate response variable. We defined the
579 transcription rate (RNA to DNA ratio) as a new variable, π_i.
580

$$\frac{y_i}{t_i} = \pi_i$$

581
582 If we assume the majority of genomic bins will have the basal transcription rate, we can
583 model the transcription rate at each i-th bin following the traditional negative binomial
584 (NB2) distribution.
585

$$\pi_i \sim NB\left(\frac{\mu_i}{t_i}, \theta\right)$$

586
587 The expected basal transcription, $E(\pi_i)$, becomes the mean incidence rate of y_i per unit
588 of exposure, t_i.
589

$$E\left(\frac{y_i}{t_i}\right) = \frac{\mu_i}{t_i}$$

590
591 By normalizing μ_i by t_i, we are modeling a rate instead of a discrete count using the
592 negative binomial distribution. The regression term for the expected transcription rate
593 can be expressed in terms of a linear combination of explanatory variables, j covariates
594 ($\vec{x}$).
595

$$\ln \frac{\mu_i}{t_i} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}$$

Rearranging in terms of the expected value of y, or μ, gives

$$\ln \mu_i - \ln t_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}$$

$$\ln \mu_i = \ln t_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}$$

$$\mu_i = \exp\left(\ln t_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}\right)$$

The natural log of t_i on the RHS ensures μ_i is normalized in the model, acting as an offset variable. In STARRPeaker software, we allow users to optionally choose this alternative rate model (implemented as “mode 2”) instead of the default covariate model described in the main text. This alternate model is useful if constant basal transcription is expected throughout the genome or if covariates are available for directly modelling the basal transcription rate π.

BasicSTARRseq

We used BasicSTARRseq R package version 1.10.0 downloaded from Bioconductor (https://bioconductor.org/packages/release/bioc/html/BasicSTARRseq.html). We used default setting as described in the software manual, except for disabling deduplication (minQuantile = 0.9, peakWidth = 500, maxPval = 0.001, deduplicate = FALSE, model = 1), to call peaks.

MACS2

We used MACS2 version 2.1.1 [23] at the recommended default setting, except for allowing duplicates in read (--keep-dup all), since our STARR-seq dataset was multiplexed. We called peaks with an FDR cutoff of 0.01, as recommended by the author of the software.

Calculating folding free energy

We used the LinearFold [47] algorithm to estimate the folding energy of each genomic bin iteratively across the whole genome. Specifically, we used the Vienna RNAfold thermodynamic model [48] with parameters from Mathews et al. 2004 [49]. We implemented a parallel processing scheme to leverage multicore processors to expedite the calculation of folding free energy.

Declarations

Availability of data and source codes

We implemented the method described in this article as a Python software package called STARRPeaker. The software package can be downloaded, installed, and readily used to call peaks from any STARR-seq dataset. The STARRPeaker package, as well as source code and documentation, is freely available at: http://github.com/gersteinlab/starrpeaker. All raw data used in the analysis as well as derived resources are available to download from the ENCODE portal (https://www.encodeproject.org/) with accession code ENCSR135NXN for HepG2 and ENCSR858MPS for K562. DNase-seq and ChIP-seq data used for the analysis is also publicly available from the ENCODE portal. The specific accession codes used for the

638 analysis are listed in Supplementary Table S3. GC content was downloaded from the
639 UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/gbdb/hg38/bbi/gc5BaseBw/),
640 and the mappability track was created using gem-library software [50] with a k-mer size
641 of 100 bp and the reference human genome build hg38.
642
643 **Competing Interests**
644 The authors declare that they have no competing interests
645
646 **Funding**
647 We acknowledge support from the NIH and from the AL Williams Professorship funds.
648
649 **Author Contributions**
650 D.L., M.S., K.W., and M.G. conceived the project. D.L. and M.G. drafted the manuscript.
651 D.L. developed the STARRPeaker software package. M.S., J.M., M.W., D.F., Y.K., and
652 L.M. performed experimental works. M.W. and Y.K. performed experimental validations.
653 D.L., J.Z., and J.L. performed the downstream analyses. M.G. and K.W. provided
654 funding and supervised the project.
655
656 **Acknowledgements**
657 We thank Jinrui Xu and Joel Rozowsky for thoughtful discussion about ChIP-seq
658 processing, Michael Rutenberg Schoenberg and Zhen Chen for thoughtful discussion
659 about RNA-folding biology, and all other members of the Gerstein and White
660 laboratories for advice and critical feedback on the manuscript.

661

662 **References**

663 1. Muerdter F, Boryń ŁM, Arnold CD. STARR-seq — Principles and applications.
664 Genomics. Academic Press; 2015;106:145–50.
665 2. Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory
666 code. Trends Genet. 2013;29:11–22.
667 3. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range
668 Shh enhancer regulates expression in the developing limb and fin and is associated
669 with preaxial polydactyly. Hum Mol Genet. 2003;12:1725–35.
670 4. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by
671 remote SV40 DNA sequences. Cell. 1981;27:299–308.
672 5. Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T. Elimination of a long-range
673 cis-regulatory module causes complete loss of limb-specific Shh expression and
674 truncation of the mouse limb. Development. 2005;132:797–803.
675 6. Melo CA, Drost J, Wijchers PJ, van de Werken H, de Wit E, Vrielink JAFO, et al.
676 eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. Mol
677 Cell. 2013;49:524–35.
678 7. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene
679 promoters. Nature. 2012;489:109–13.
680 8. Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-
681 Rivera A, Souaid C, et al. Genome-wide characterization of mammalian promoters with
682 distal enhancer functions. Nat Genet. Nature Publishing Group; 2017;49:1073–81.
683 9. Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, et al. A tiling-deletion-based genetic

684 screen for cis-regulatory element identification in mammalian cells. Nat Methods.
685 2017;14:629–35.
686 10. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and
687 characterization. Nat Methods. Nature Research; 2012;9:215–6.
688 11. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised
689 pattern discovery in human chromatin structure through genomic segmentation. Nat
690 Methods. Nature Publishing Group; 2012;9:473–6.
691 12. Sethi A, Gu M, Gumusgoz E, Chan L, Yan K-K, Rozowsky J, et al. A cross-organism
692 framework for supervised enhancer prediction with epigenetic pattern recognition and
693 targeted validation. bioRxiv. Cold Spring Harbor Laboratory; 2018;385237.
694 13. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively
695 parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. Nature
696 Publishing Group; 2012;30:265–70.
697 14. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic
698 dissection and optimization of inducible enhancers in human cells using a massively
699 parallel reporter assay. Nat Biotechnol. Nature Publishing Group; 2012;30:271–7.
700 15. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide
701 quantitative enhancer activity maps identified by STARR-seq. Science (80-).
702 2013;339:1074–7.
703 16. Liu Y, Yu S, Dhiman VK, Brunetti T, Eckart H, White KP. Functional assessment of
704 human enhancer activities using whole-genome STARR-sequencing. Genome Biol.
705 2017;18:219.
706 17. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, et al. A systematic

evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. bioRxiv. Cold Spring Harbor Laboratory; 2019;576405.

18. Johnson GD, Barrera A, McDowell IC, D'Ippolito AM, Majoros WH, Vockley CM, et al. Human genome-wide measurement of drug-responsive regulatory activity. Nat Commun. 2018;9:5317.

19. Rathert P, Roth M, Neumann T, Muerdter F, Roe J-S, Muhar M, et al. Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. Nature. 2015;525:543–7.

20. Koohy H, Down TA, Spivakov M, Hubbard T. A comparison of peak callers used for DNase-Seq data. Helmer-Citterich M, editor. PLoS One. 2014;9:e96303.

21. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov F V, Hodges E, et al. Site identification in high-throughput RNA-protein interaction data. Bioinformatics. Oxford University Press; 2012;28:3013–20.

22. Strbenac D, Armstrong NJ, Yang JYH. Detection and classification of peaks in 5' cap RNA sequencing data. BMC Genomics. BioMed Central; 2013;14 Suppl 5:S9.

23. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. BioMed Central; 2008;9:R137.

24. Kharchenko P V, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. Nature Publishing Group; 2008;26:1351–9.

25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

26. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov M V,

730 Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci
731 Rep. 2014;4:4532.
732 27. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al.
733 Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl
734 Acad Sci U S A. 2013;110:6376–81.
735 28. Lai D, Proctor JR, Meyer IM. On the importance of cotranscriptional RNA structure
736 formation. RNA. 2013;19:1461–73.
737 29. Ringnér M, Krogh M. Folding free energies of 5'-UTRs impact post-transcriptional
738 regulation on a genomic scale in yeast. PLoS Comput Biol. 2005;1:e72.
739 30. Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, et al. Metabolic
740 labeling of RNA uncovers principles of RNA production and degradation dynamics in
741 mammalian cells. Nat Biotechnol. 2011;29:436–42.
742 31. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, et
743 al. Decay rates of human mRNAs: Correlation with functional characteristics and
744 sequence attributes. Genome Res. 2003;13:1863–72.
745 32. Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, Wakamatsu A, et al. Genome-wide
746 determination of RNA stability reveals hundreds of short-lived noncoding transcripts in
747 mammals. Genome Res. 2012;22:947–56.
748 33. Papoulis A, Athanasios. Probability, random variables and stochastic processes.
749 New York McGraw-Hill, 1984, 2nd ed. 1984;
750 34. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet.
751 Nature Publishing Group; 2020;52:1–10.
752 35. Hilbe JM. Negative Binomial Regression [Internet]. Cambridge: Cambridge

753 University Press; 2011.

754 36. Cameron ACA, Trivedi PK. Regression Analysis of Count Data [Internet].

755 Cambridge: Cambridge University Press; 2013.

756 37. Hilbe JM. Modeling Count Data [Internet]. Cambridge: Cambridge University Press;

757 2014.

758 38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and

759 Powerful Approach to Multiple Testing. J R Stat Soc Ser B. John Wiley & Sons, Ltd

760 (10.1111); 1995;57:289–300.

761 39. Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, et al.

762 Resolving systematic errors in widely used enhancer activity assays in human cells. Nat

763 Methods. Nature Publishing Group; 2018;15:141–9.

764 40. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five

765 essential questions. Nat Rev Genet. NIH Public Access; 2013;14:288–95.

766 41. Kawaji H, Kasukawa T, Forrest A, Carninci P. The FANTOM 5 collection, a data

767 series underpinning mammalian transcriptome atlases in diverse cell types. Sci Data.

768 2017;2016–8.

769 42. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, et al. A

770 systematic comparison reveals substantial differences in chromosomal versus episomal

771 encoding of enhancer activity. Genome Res. 2017;27:38–52.

772 43. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-

773 resolution genome-wide functional dissection of transcriptional regulatory regions and

774 nucleotides in human. Nat Commun. Nature Publishing Group; 2018;9:1–15.

775 44. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq:

776 Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic
777 Screens. Cell. Cell Press; 2016;167:1853-1866.e17.
778 45. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed Engineering and Analysis of
779 Combinatorial Enhancer Activity in Single Cells. Mol Cell. Cell Press; 2017;66:285-
780 299.e5.
781 46. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A
782 Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens.
783 Cell. Cell Press; 2019;176:377-390.e19.
784 47. Huang L, Zhang H, Deng D, Zhao K, Liu K, Hendrix DA, et al. LinearFold: linear-
785 time approximate RNA folding by 5'-to-3' dynamic programming and beam search.
786 Bioinformatics. Narnia; 2019;35:i295–304.
787 48. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et
788 al. ViennaRNA Package 2.0. Algorithms Mol Biol. BioMed Central; 2011;6:26.
789 49. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH.
790 Incorporating chemical modification constraints into a dynamic programming algorithm
791 for prediction of RNA secondary structure. Proc Natl Acad Sci U S A. National Academy
792 of Sciences; 2004;101:7287–92.
793 50. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast
794 Computation and Applications of Genome Mappability. Ouzounis CA, editor. PLoS One.
795 2012;7:e30377.
796

797 ## Supplementary Tables

798 **Table S1** contains significant peaks called by STARRPeaker.

Table S2 contains overlap of various peak callers (STARRPeaker, BasicSTARRseq, and MACS2) to published enhancers identified using other types of enhancer assays.

Table S3 contains a list of data sources and accession numbers used for the analysis.

Table S4 compares peaks identified by various peak callers (STARRPeaker, BasicSTARRseq, and MACS2).

Supplementary Figures

Supplementary Figure 1 Comparison of STARR-seq output coverage calculated using the center of the fragment to using the start position of the sequencing read. **(A)** Distribution of shift in final peak locations resulting from using two alternative coverage counting schemes in HepG2. Comparison of **(B)** overall fold enrichment level, **(C)** p-value, and **(D)** size of resulting peaks.

Supplementary Figure 2 Contribution of covariates and model selection. **(A)** Q-Q plots of various models with different sets of covariates showing the goodness of fit. **(B)** Both AIC and BIC measure relative qualities of statistical models considering the trade-off between the goodness of fit and the simplicity of the model. AIC: Akaike information criterion; BIC: Bayesian information criterion.

Supplementary Figure 3 Correlation between replicates for **(A)** HepG2 or **(B)** K562 cell lines.

821 **Supplementary Figure 4** Comparison of peaks called from subsamples of the original
822 STARR-seq library, highlighting the robustness of STARRPeaker.
823
824 **Supplementary Figure 5** Orientation biases analysis for **(A-B)** HepG2 or **(C-D)** K562
825 cell lines. The ratio between forward and reverse stranded fragments was tested for
826 statistical significance using a binomial test. Orange dots represent peaks with
827 significant strand bias (FDR q-value < 0.01).
828
829 **Supplementary Figure 6** Comparison of peaks identified by various methods using a
830 simulated STARR-seq dataset containing four spike-in control regions.
831
832 **Supplementary Figure 7** Enrichment of epigenetic signals around peaks in K562. All
833 peaks were centered at the summit, uniformly thresholded using P-value < 0.001, and
834 10,000 peaks were randomly selected. Aggregated read depth at 2,000 bp upstream
835 and downstream were plotted for **(A)** DNase I hypersensitive sites (DHS), **(B)** H3K27ac,
836 **(C)** H3K4me1, and **(D)** aggregated TF ChIP-seq profile. For DNase-seq, enrichment
837 indicates unique read depth. For histone ChIP-seq, enrichment indicates fold change
838 over control. For TF ChIP-seq aggregate, enrichment indicates the number of TFs
839 binding.
840
841 **Supplementary Figure 8 (A-C)** Genome browser session comparing STARRPeaker to
842 other peak-calling methods at validated enhancers from CRISPRi.
843

844 **Supplementary Figure 9** Application of STARRPeaker on an external HeLa-S3 dataset.
845 **(A)** Comparison of chromatin accessibility (DNase-seq) for STARRPeaker peaks
846 between untreated and inhibitor-treated samples. **(B)** Comparison of STARRPeaker
847 peaks to published results. STARRPeaker found 6,540 additional peaks that are
848 enriched with chromatin accessibility signals from a HeLa-S3 sample.
849
850 **Supplementary Figure 10** Venn diagram for four-way comparison of peaks identified
851 by various methods using a published dataset from Rathert et al. 2015.
852

853 ## Figure legends

854 **Figure 1** Comparison of STARR-seq output coverage calculated using the center of the
855 fragment to using the start position of the sequencing read. **(A)** Distribution of the shift in
856 final peak locations resulting from using two alternative coverage counting schemes in
857 HepG2. Comparison of **(B)** overall fold enrichment level, **(C)** p-value, and **(D)** size of
858 resulting peaks. **(E)** Example highlighting the difference between fragment-based and
859 read-based coverage counting schemes and their resulting peak calls from HepG2
860 STARR-seq data. Asterisks represents statistical significance using the Mann-Whitney-
861 Wilcoxon test two-sided with Bonferroni correction; (*) P <= 0.05, (**) P <= 0.01, (***) P
862 <= 0.001, (****) P <= 0.0001.
863
864 **Figure 2** Confounding factors in the STARR-seq assay. STARR-seq output and input
865 coverages are significantly correlated with **(A)** input coverage, **(B)** GC content, **(C)**

mappability, and **(D)** RNA structure folding. PCC: Pearson Correlation Coefficient. Plots were from a sampling of 5,000 random genomic bins.

Figure 3 STARR-seq output coverage is fitted against simulated coverage using three distribution models; negative binomial, binomial, and Poisson. **(A)** Density histogram of simulated distribution against STARR-seq output coverage. **(B)** Q-Q plot of simulated distribution against STARR-seq output coverage. The red solid line represents where the observed count equals the expected count.

Figure 4 Overview of STARRPeaker peak-calling scheme. **(A)** In contrast to using read depth (grey), fragment depth (red) offers more precise and sharper STARR-seq output coverage. Fragment inserts are directly inferred from properly paired-reads. **(B)** Workflow of STARRPeaker describing how coverage is calculated for each genomic bin and modelled using a negative binomial regression model. The analysis pipeline can largely be divided into four steps: (1) Binning the genome; (2) calculating coverage and computing covariate matrix; (3) fitting the STARR-seq data to the NB regression model; and (4) peak calling, multiple hypothesis testing correction, and adjustment of the center of peaks.

Figure 5 Enrichment of epigenetic signals around peaks in HepG2. All peaks were centered at the summit, uniformly thresholded using P-value < 0.001, and 10,000 peaks were randomly selected. Aggregated read depth at 2,000 bp upstream and downstream were plotted for **(A)** DNase I hypersensitive sites (DHS), **(B)** H3K27ac, **(C)** H3K4me1,

889 and **(D)** aggregated TF ChIP-seq profile. For DNase-seq, enrichment indicates unique
890 read depth. For histone ChIP-seq, enrichment indicates fold change over control. For
891 TF ChIP-seq aggregate, enrichment indicates the number of TFs binding.
892
893 **Figure 6** Comparison of peaks using an external dataset for **(A)** HepG2 or **(B)** K562 cell
894 lines. Peaks identified from STARRPeaker as well as BasicSTARRseq and MACS2
895 were compared against a published dataset. For a fair comparison, all peaks were
896 centered at the summit, uniformly thresholded using P-value < 0.001, and 20,000 peaks
897 were randomly drawn from peaks identified by each peak caller using the recommended
898 settings. The fraction of overlap was computed for each replicate. We considered it an
899 overlap when at least 50% of peaks intersected each other.

Figure 1

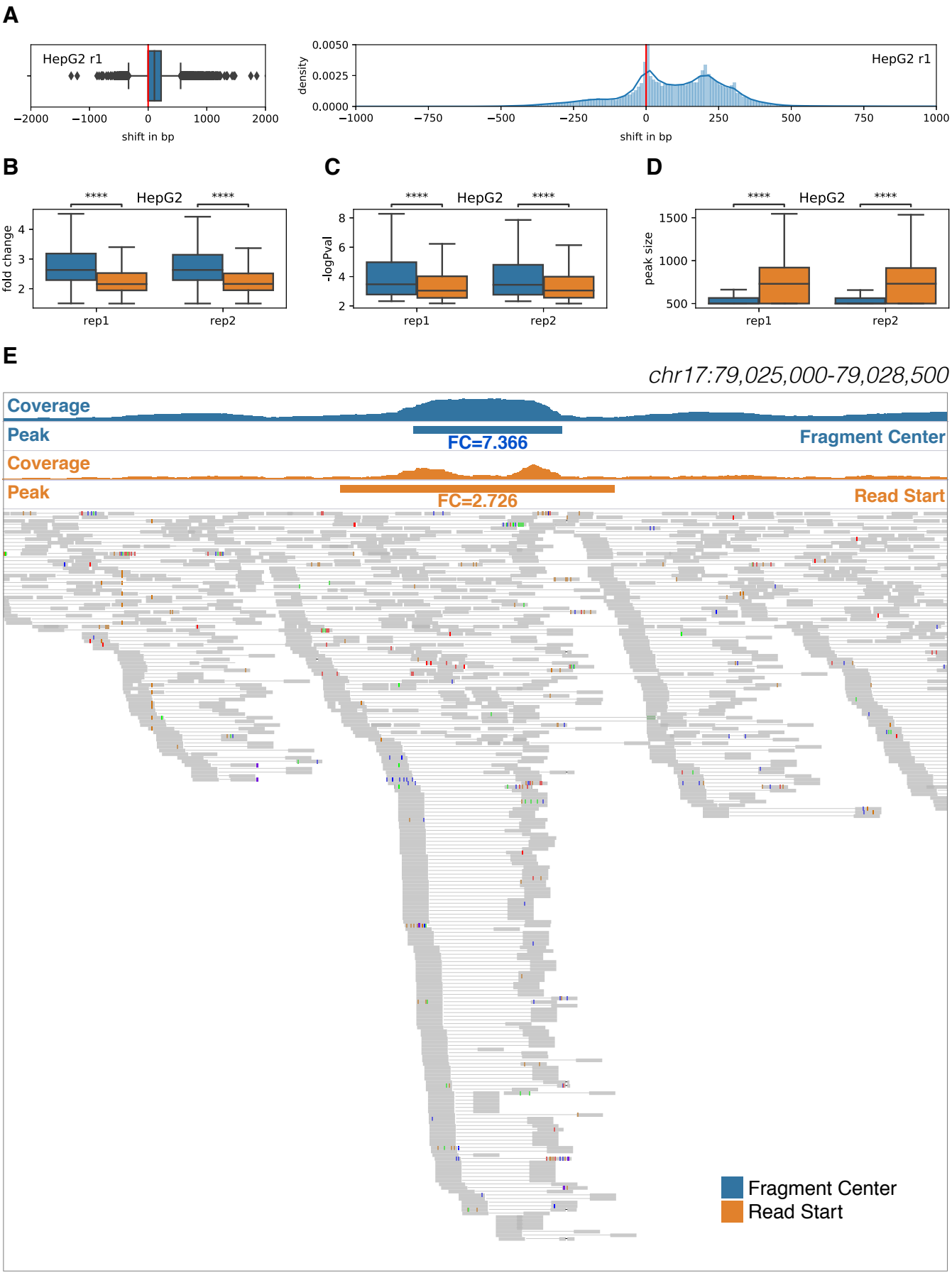


Figure 2

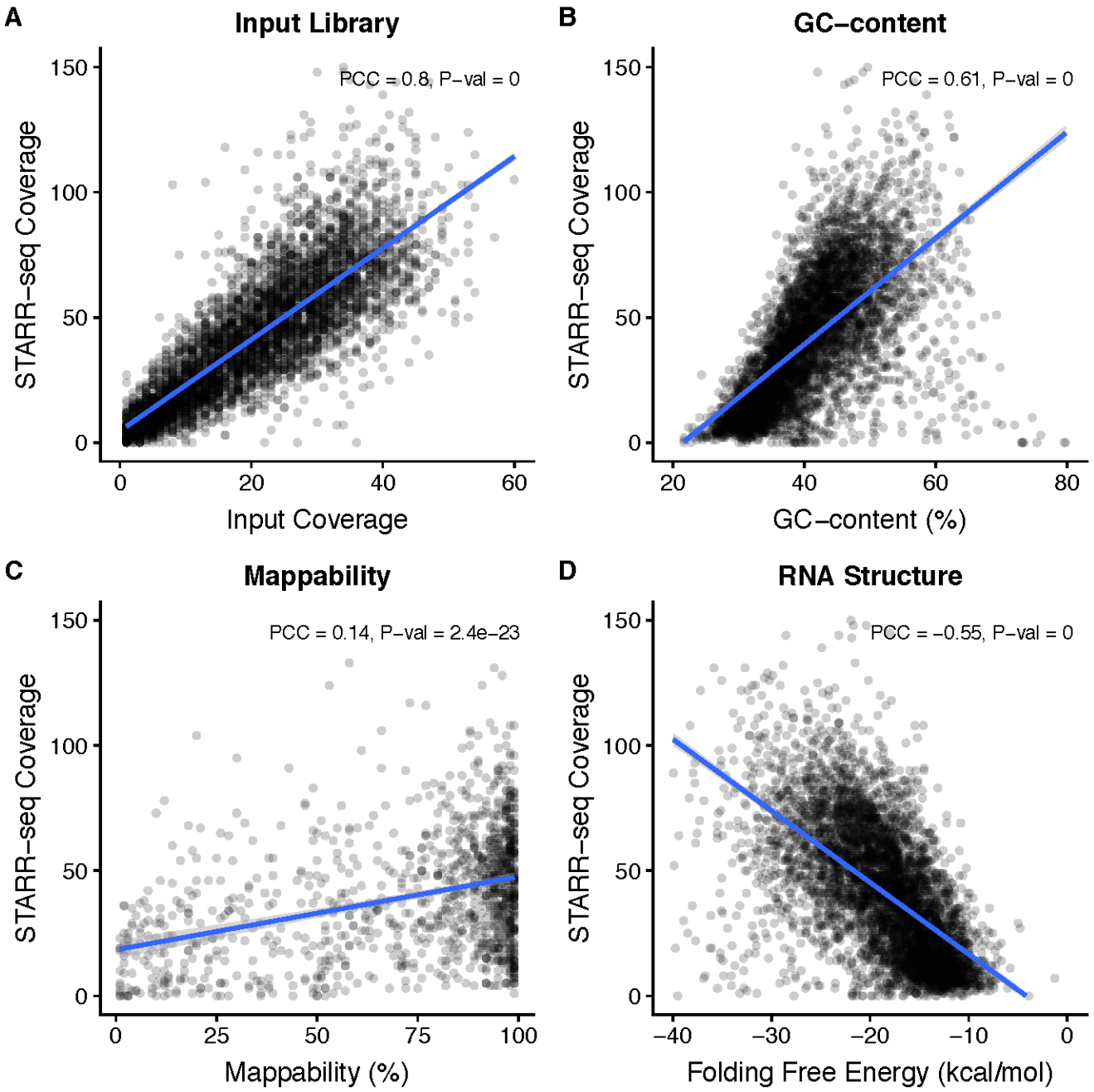


Figure 3

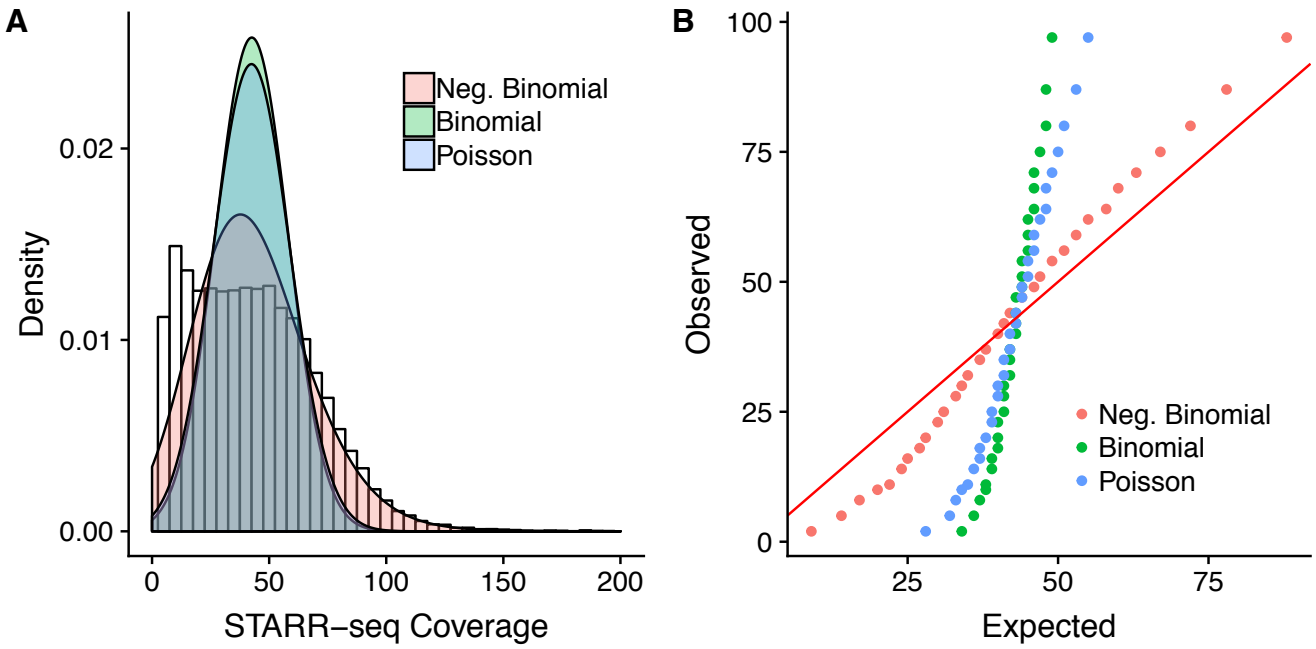


Figure 4

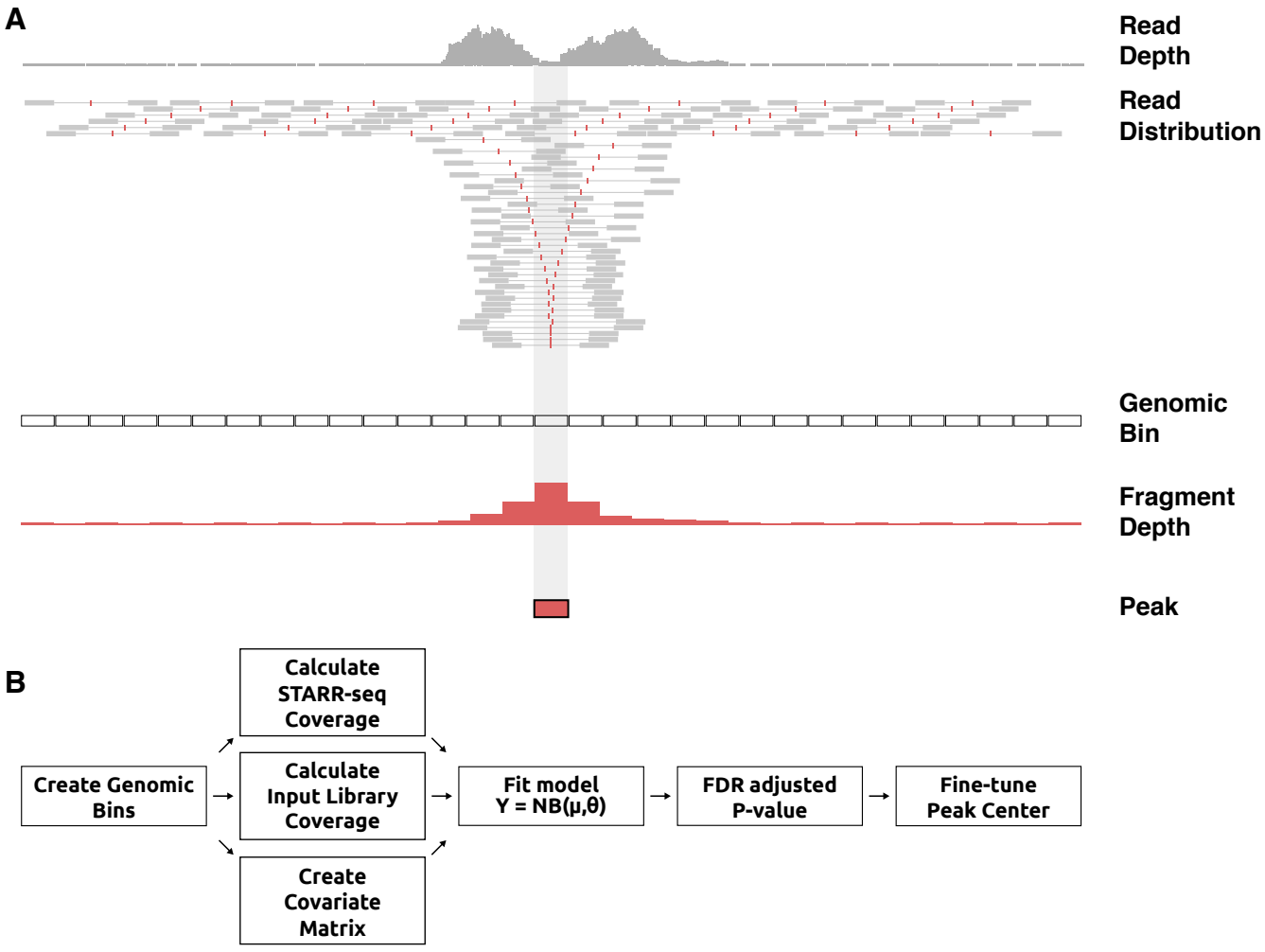


Figure 5

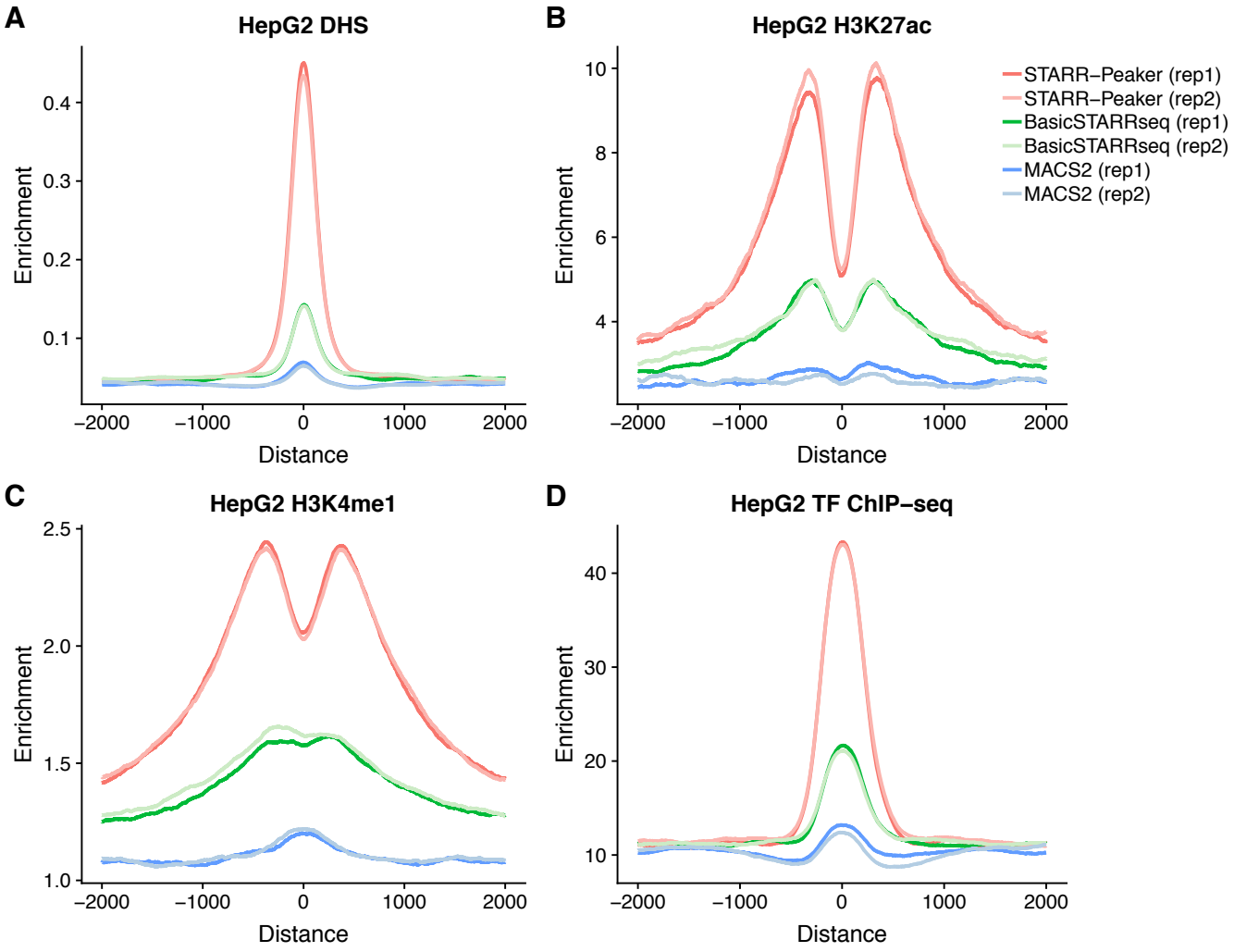
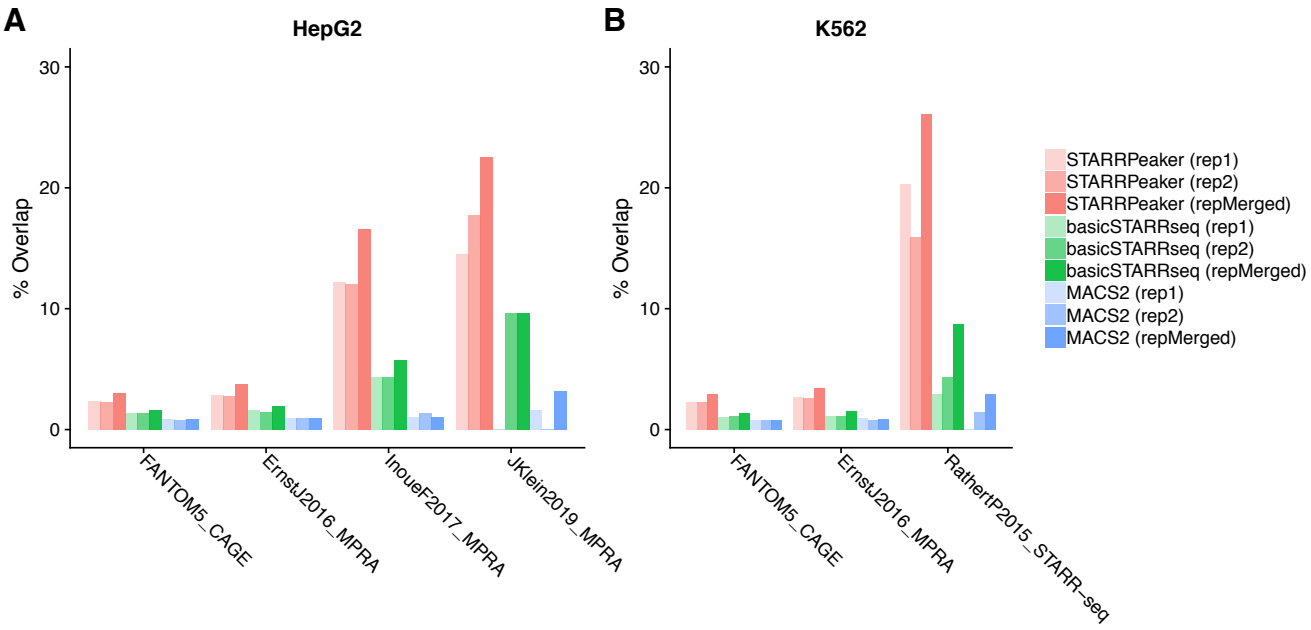
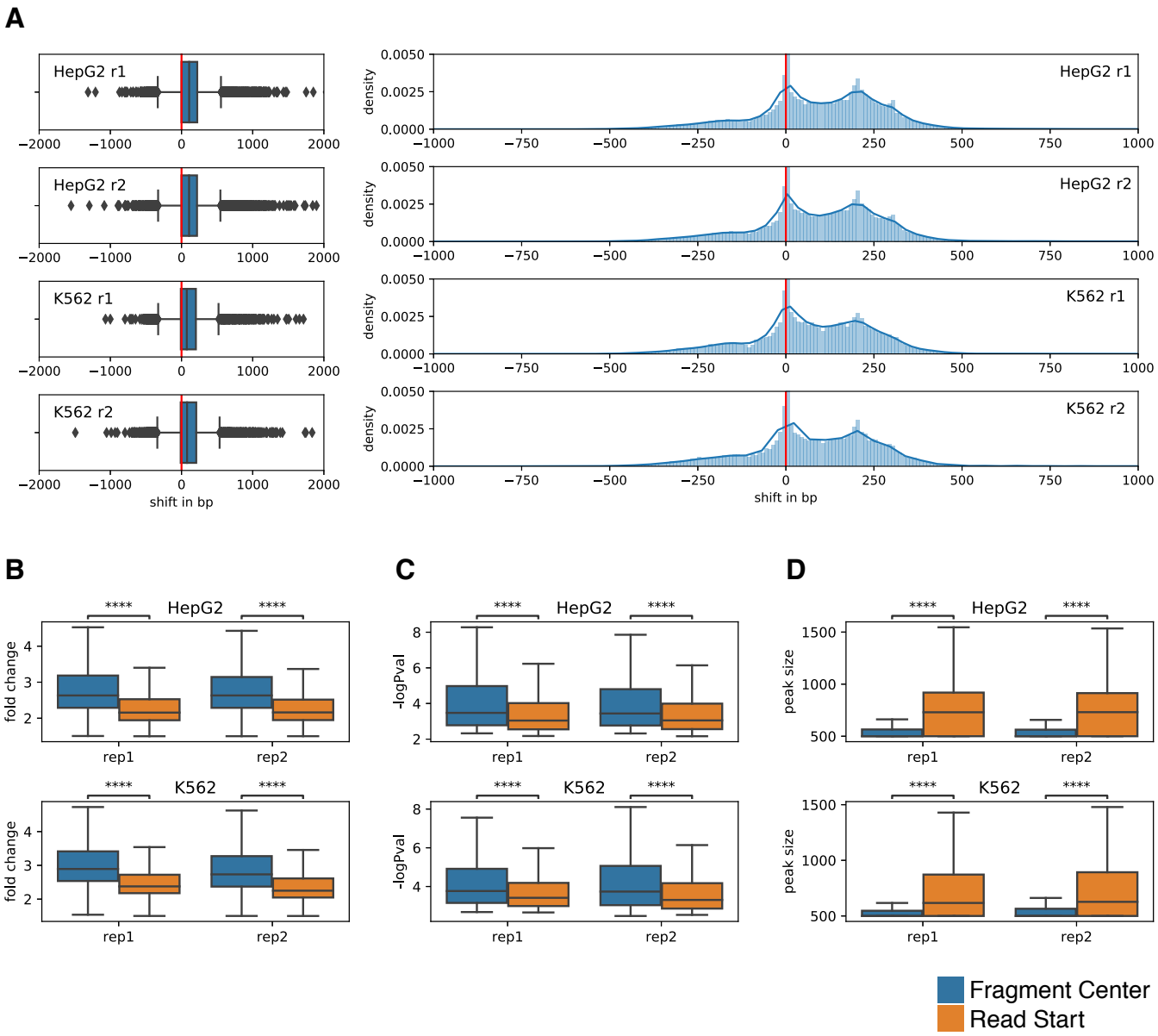


Figure 6

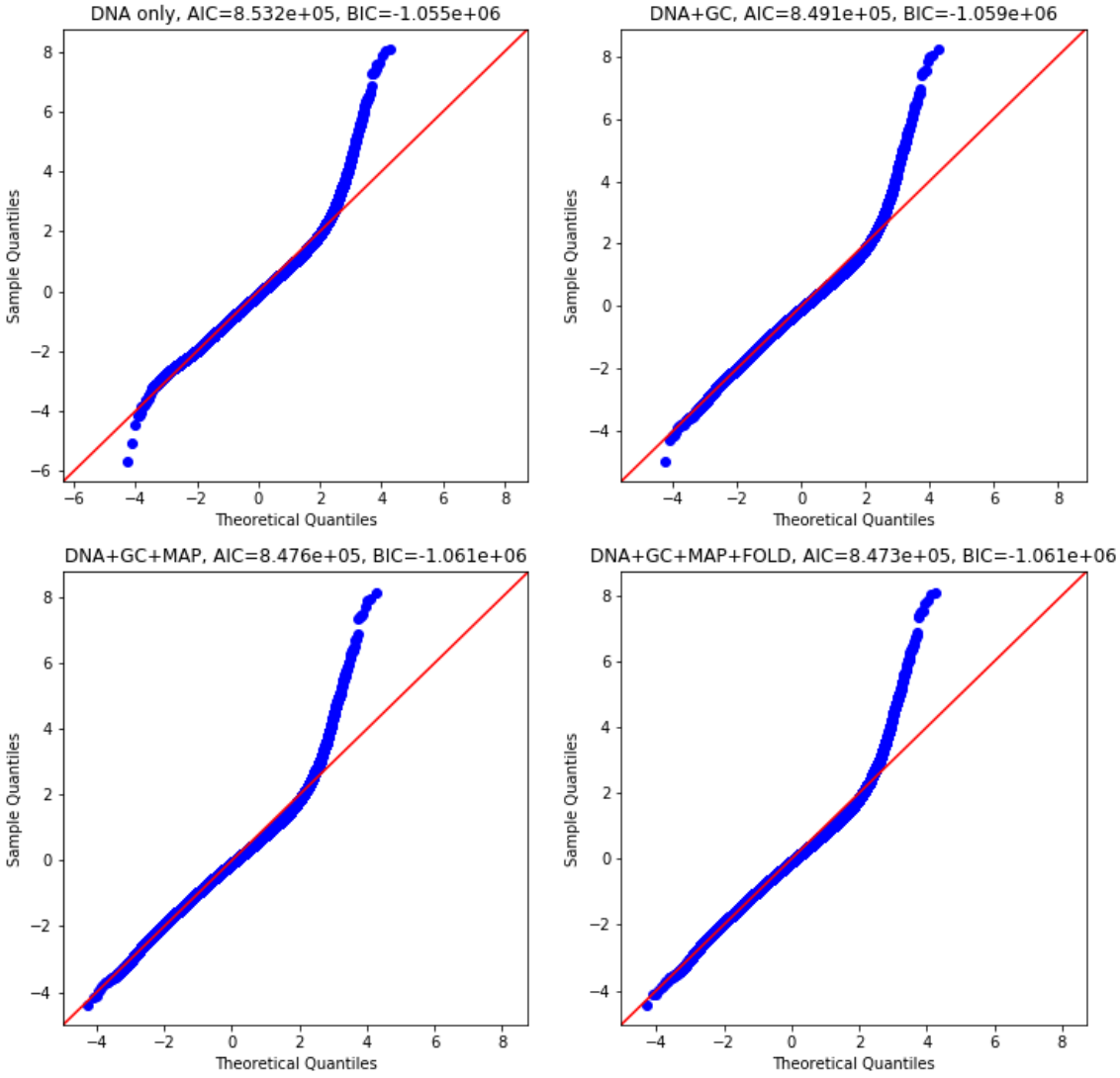


Supplementary Figure 1

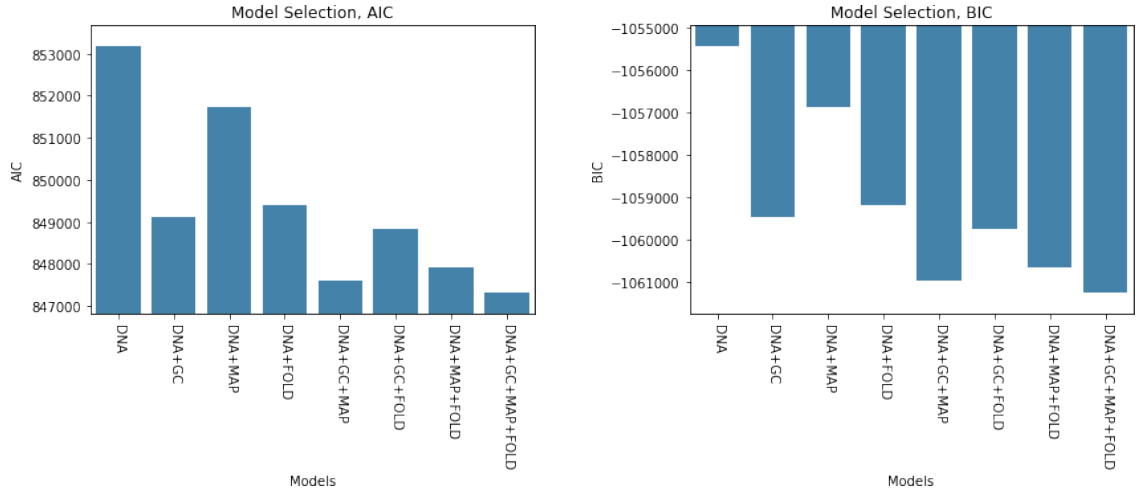


Supplementary Figure 2

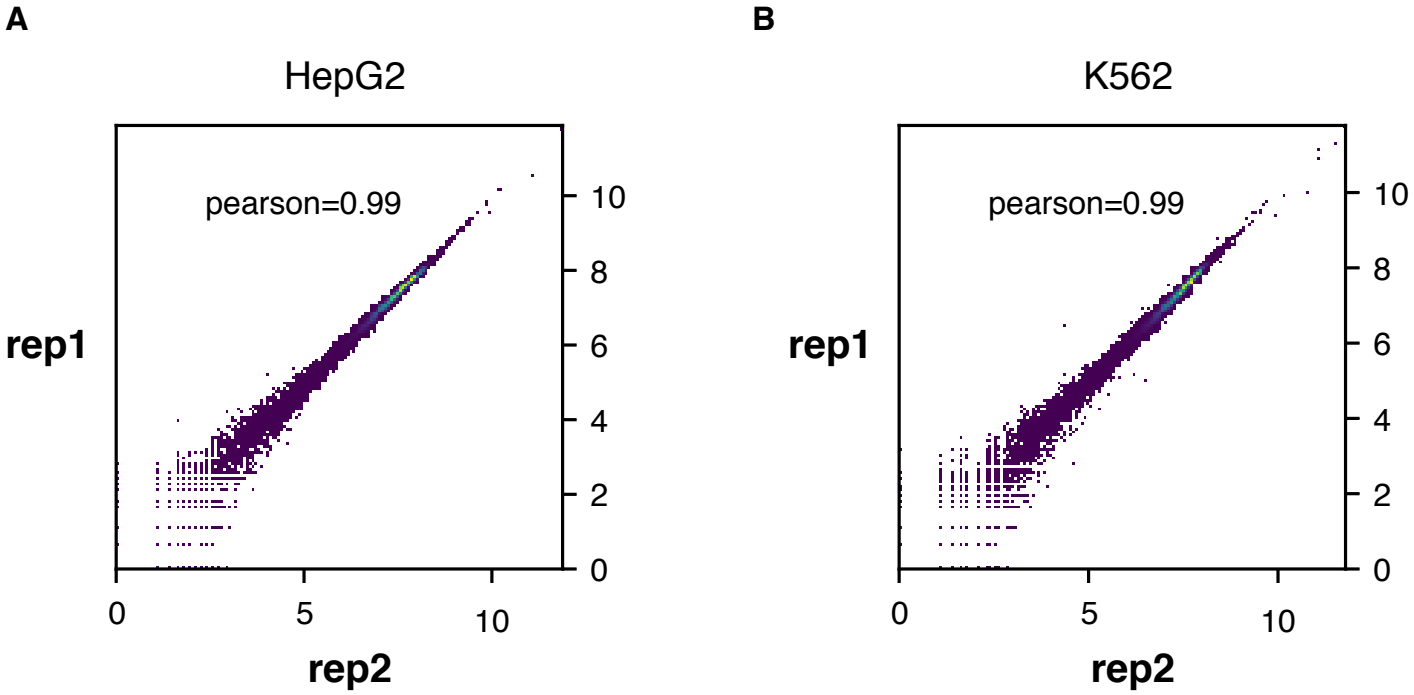
A



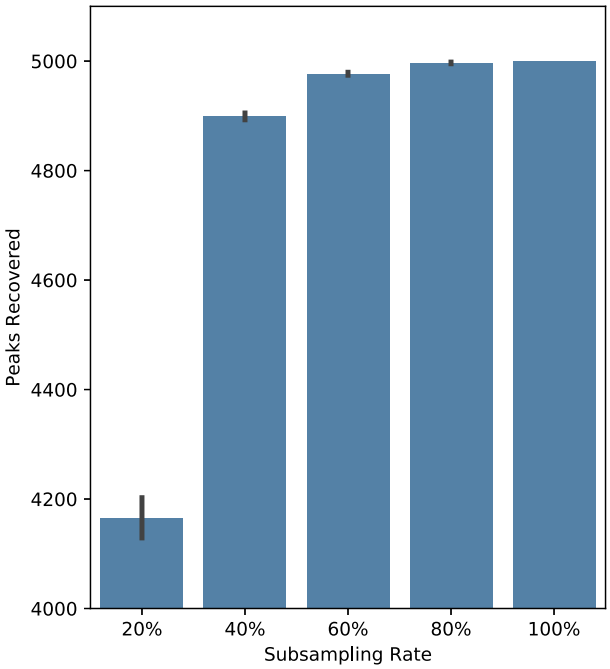
B



Supplementary Figure 3

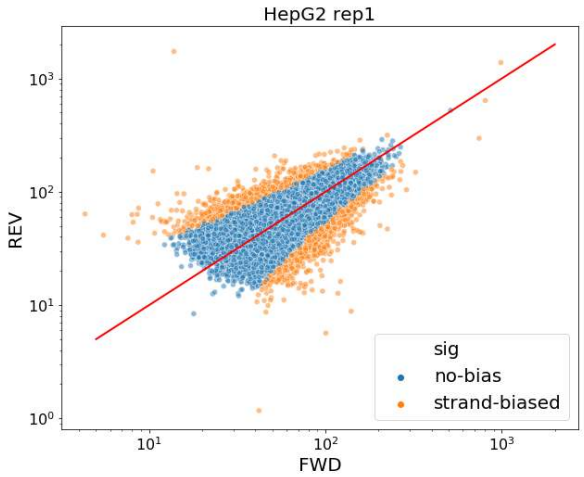


Supplementary Figure 4

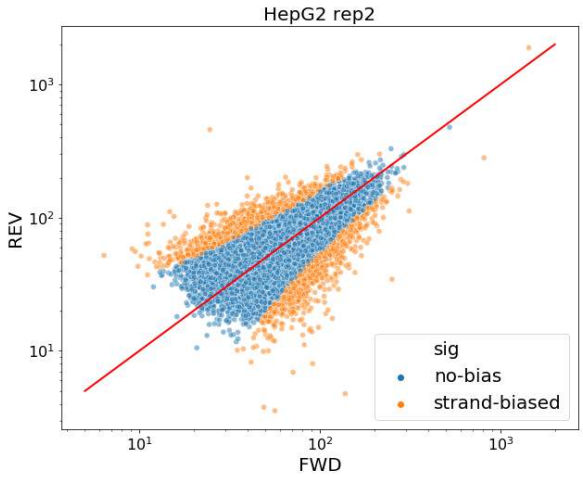


Supplementary Figure 5

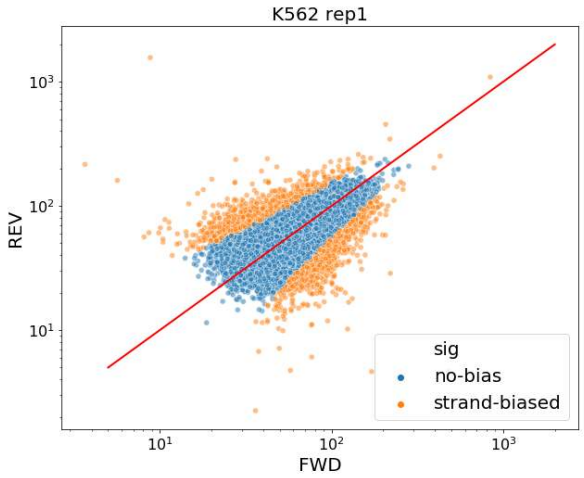
A



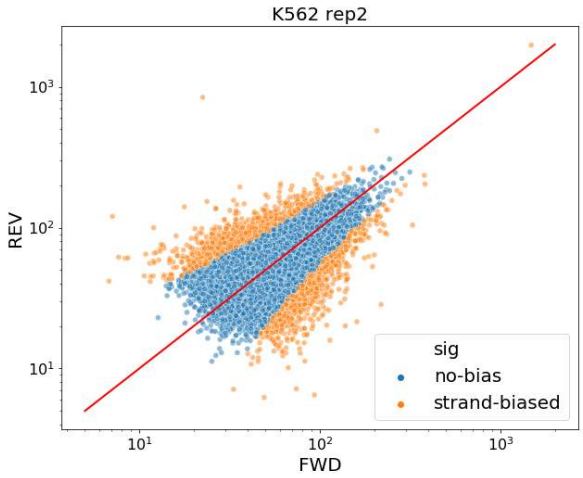
B



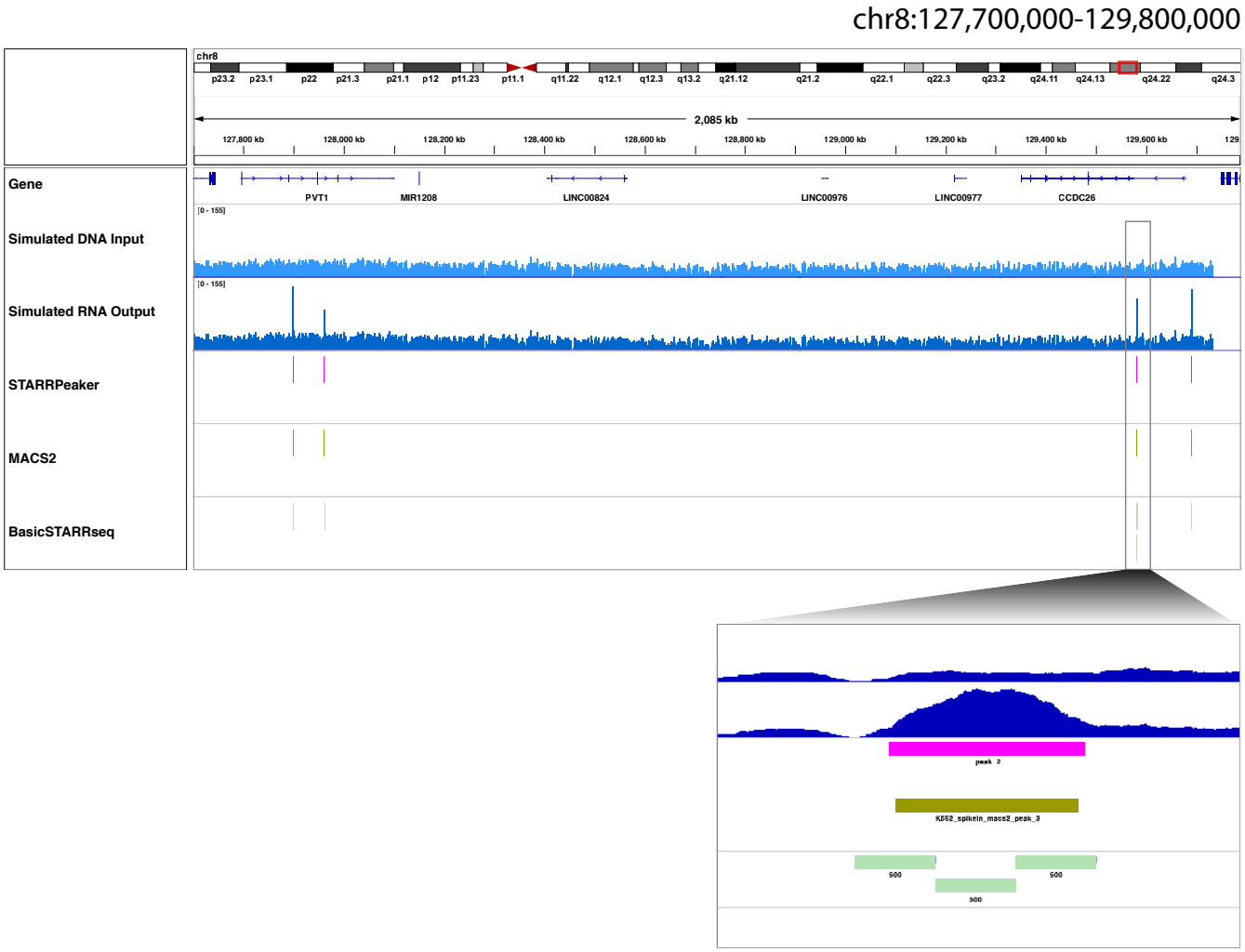
C



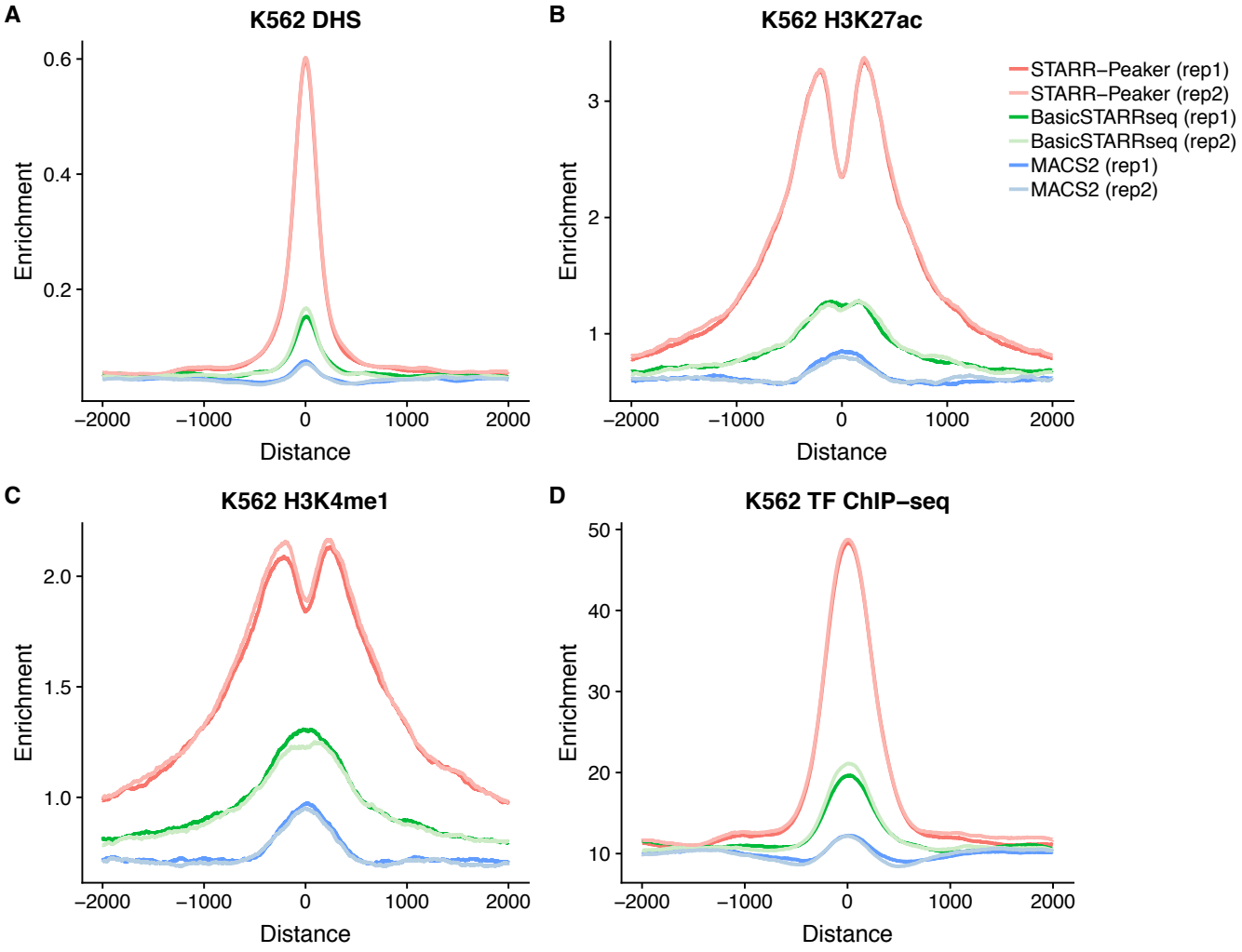
D



Supplementary Figure 6

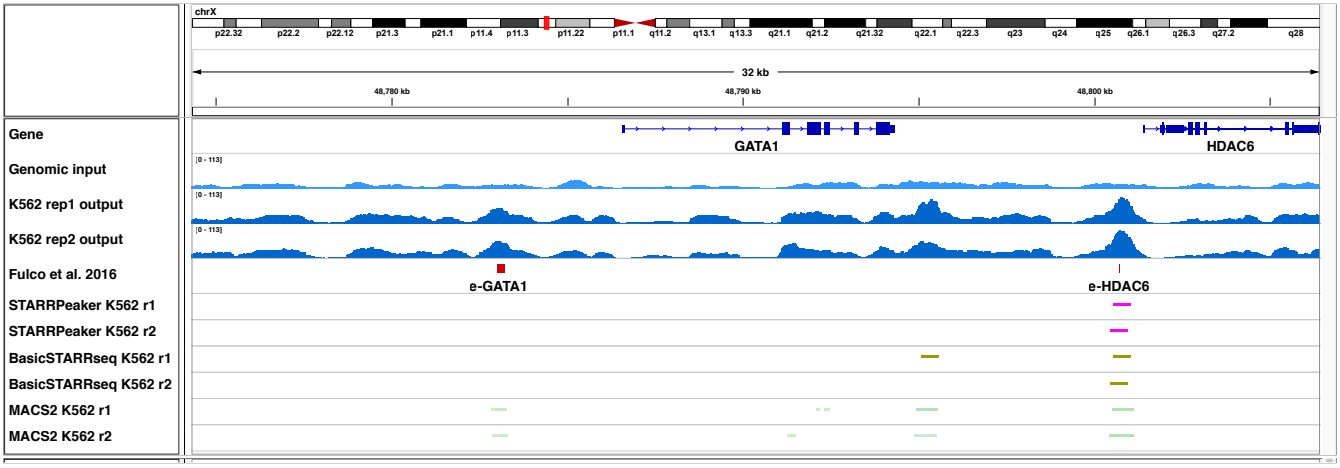


Supplementary Figure 7

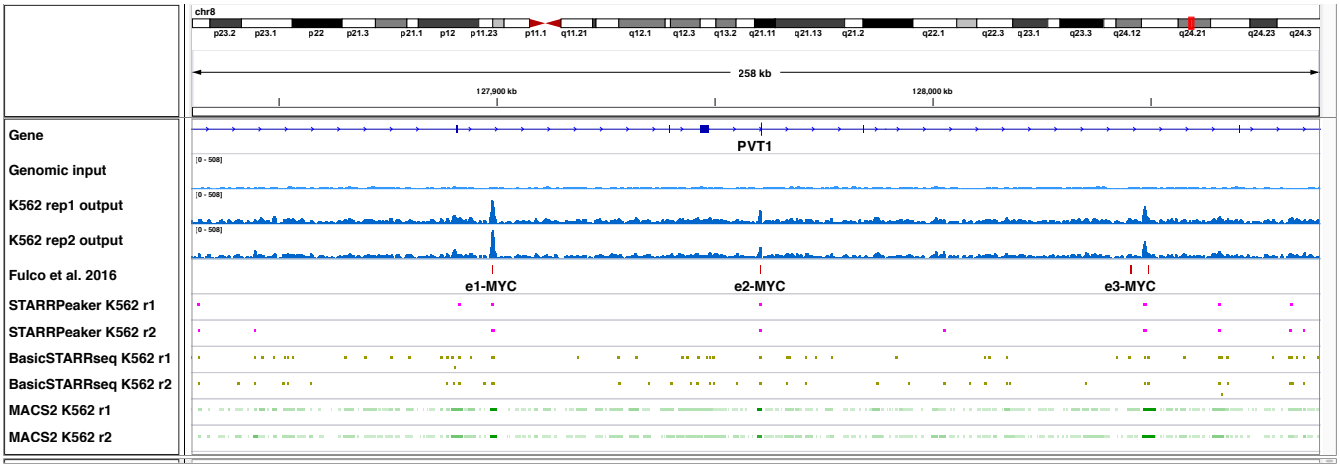


Supplementary Figure 8

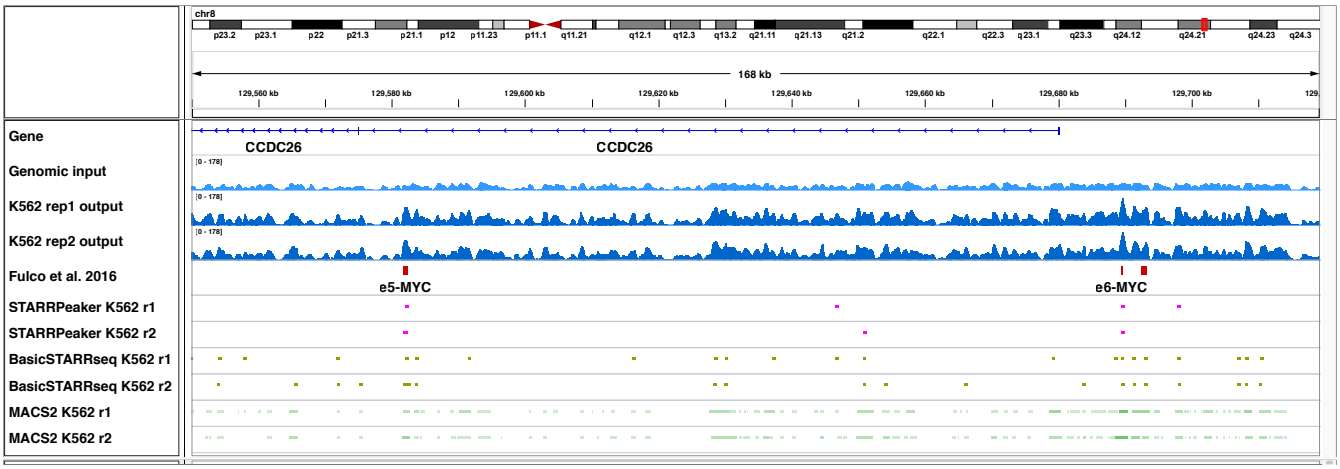
A GATA1 locus chrX:48774179-48806695



B MYC locus chr8:127830000-128090000

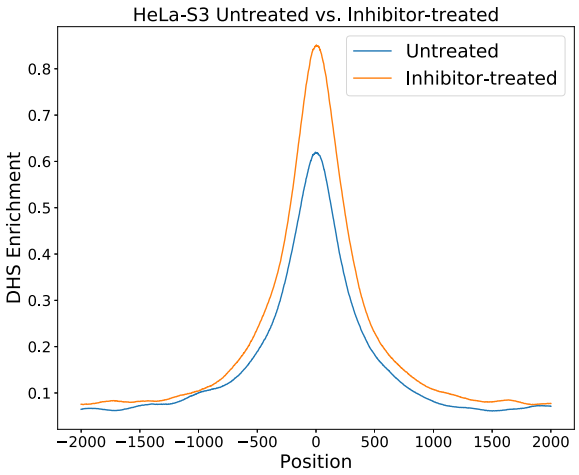


C MYC locus chr8:129550000-129720000

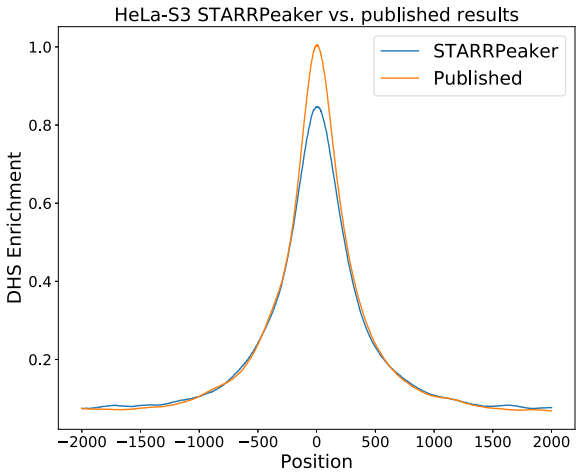


Supplementary Figure 9

A



B



Supplementary Figure 10