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

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

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

2 Institute for Genomics and System Biology, University of Chicago, Chicago, IL, 60637, USA

3 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

4 School of Life Sciences, Westlake University, Hangzhou, 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 author

E-mail: pi@gersteinlab.org

Abstract

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 level. 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. Most STARR-seq studies have relied on chromatin immunoprecipitation sequencing (ChIP-seq) processing pipeline to identify peaks. However, there are key differences in STARR-seq versus ChIP-seq data: STARR-seq uses transcribed RNA to measure enhancer activity, making determining the basal transcription rate important. Furthermore, STARR-seq coverage is non-uniform, overdispersed, and often confounded by sequencing biases such as GC content and mappability. Moreover, here, we observed a clear correlation between RNA thermodynamic stability and STARR-seq readout, suggesting that STARR-seq might be sensitive to RNA secondary structure and stability. Considering these findings, we developed STARRPeaker: a negative binomial regression framework for uniformly processing STARR-seq data. We applied STARRPeaker to two whole human genome STARR-seq experiments; HepG2 and K562. Our method identifies highly reproducible and epigenetically active enhancers across replicates. Moreover, STARRPeaker outperforms other peak callers in terms of identifying known enhancers. Thus, our framework optimized for processing STARR-seq data accurately characterizes cell-type-specific enhancers, while addressing potential confounders.

Keywords: STARR-seq, peak caller, enhancer, non-coding

Introduction

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

Unlike protein-coding genes, enhancers do not yet have a well-characterized consensus sequence. Therefore, identifying enhancers in an unbiased fashion is challenging. 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 the fair assessment of enhancer function.

Previously, putative regulatory elements were computationally predicted, indirectly, by profiling DNA accessibility (using DNase-seq, FAIRE-seq, and 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. The activities of enhancers are measured by comparing the amount of RNA produced from the input DNA library. 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 for functional elements. 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.

Processing of STARR-seq is somewhat similar to chromatin immunoprecipitation sequencing (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 default parameters optimized for processing transcription factor 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 positives 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 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.

Materials and Methods

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 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 sizes of the fragments inserted in STARR-seq libraries were approximately 500 bp, we expected that the read coverage using the start position of read may shift the estimate of the summit of signal and dilute the enrichment. Some peak callers have used read densities of forward and reverse strand 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 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 structural variation[25]. We also filtered out read pairs that had a fragment insert size less than l_{max} and greater than l_{min}. By default, we filtered out fragment insert sizes less than 100 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. We compared the coverage calculated using the start of read against the center of fragment insert and observed both a shift in the location of enrichment summit and a difference in enrichment level (**Figure 1**).

Controlling for potential systemic bias in sequencing and STARR-seq library preparation

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 the basal transcriptional level stays relatively constant across the genome and the transcriptional rate is a reflection of the regulatory activity of a test region. However, this may not always be true, and one needs to consider potential systemic biases when analyzing the result. Unlike ChIP-seq where both the experiment and input controls are 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[26]. 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[27]. 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 vary significantly across the genome and RNA degradation rates are the main determinant of cellular RNA levels[28]. Furthermore, RNA stability correlates with functionality[29,30].

There are also 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[31]. Studies have also suggested that epigenetic mechanisms and CpG methylation may influence fragmentation[32]. Furthermore, the isolated polyadenylated RNAs are reverse transcribed and PCR is amplified before sequenced, and this process can further confound the sequenced candidate fragments.

To unbiasedly test for the regulatory activity, a model needs to control for these potential systemic biases inherent to generating STARR-seq data. As we expected, we observed that STARR-seq coverage for both input and output are confounded by potential sequencing bias (**Figure 2**). Notably, STARR-seq coverage significantly correlated with GC content (PCC 0.61; P-val 1E-299), mappability (PCC 0.45; P-val 2.9E-148), and RNA thermodynamic stability (PCC -0.55; P-val 0). Hence, to unbiasedly identify the activity peaks from STARR-seq, we developed a model that accounts for variability of tested candidate fragments.

Accurate modelling of STARR-seq coverage 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 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 normalized input coverage was less than a minimum quantile t_{min}, since these regions do not have sufficient power to detect enrichment. We simulated and fitted various discrete probability distributions to STARR-seq coverage. We observed that the STARR-seq coverage data was overdispersed and fitted the best with negative binomial distribution (**Figure 3A**). We also noticed a slight negative enrichment, indicating that some candidate fragments can silence the basal transcriptional activity. A Q-Q plot of simulated coverage further demonstrated that the negative binomial model provides the best fit for the data (**Figure 3B**).

Peak caller

To accurately model the ratio of STARR-seq sequence coverage (RNA) to input sequence 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 each genomic bin. In doing so, we use covariates to model expected counts in the form of multiple regression. Once regression coefficients are estimated from a set of data, we can evaluate the likelihood of observing the fragment count for each bin and assign p-values. Ultimately, bins with significant enrichments are selected based on an adjusted p-values threshold, and they are fine-tuned to the summit of the peak fragment enrichment.

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 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 (DNA) mapped to the i-th bin. We define X 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[34–36] (see Supplementary Methods for derivation).

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

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. The variance for the NB2 model is given as

$$\sigma^2 = \mu + \frac{\mu^2}{\theta}$$

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

$$y_i \sim NB(\mu_i,\theta)$$

$$E(y_i) = \mu_i$$

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.

$$\ln \mu_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}$$

$$\mu_i = \exp(\beta_0 x_{0i} + \beta_1 x_{1i} + \cdots + \beta_m x_{mi})$$

$$\mu_i = \exp\left(\vec{x_i}^{\intercal} \beta\right)$$

Alternatively, instead of using the input library variable t_i as one covariate, we can directly use it as an offset variable. 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 described in the Supplementary Methods.

Maximum-likelihood estimation

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

$$\mathcal{L}_{NB}(\mu|y,\theta) = \sum_{i=1}^{n} y_i \ln\left(\frac{\mu_i}{\theta + \mu_i}\right) + \theta \ln\left(\frac{\theta}{\theta + \mu_i}\right) + \ln\left(\frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1) \cdot \Gamma(\theta)}\right)$$

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

$$\mathcal{L}_{NB}(\beta|y,\theta) = \sum_{i=1}^{n} y_i \ln\left(\frac{\mathrm{e}^{\vec{x_i}^{\intercal}\beta}}{\theta + \mathrm{e}^{\vec{x_i}^{\intercal}\beta}}\right) + \theta \ln\left(\frac{\theta}{\theta + \mathrm{e}^{\vec{x_i}^{\intercal}\beta}}\right) + \ln\left(\frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1)\cdot\Gamma(\theta)}\right)$$

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.

<u>*Estimation of P-value*</u>

Finally, we calculate a P-value based on the fitted value of the i-th bin from the cumulative distribution function of negative binomial distribution, and we assign false discovery rate using Benjamini & Hochberg method[37].

$$P\text{–}value = \Pr(x \geq y_i) = 1 - CDF(x = y_i - 1)$$

$$= 1 - \sum_{i=0}^{\hat{y}_i - 1} \binom{\hat{y}_i + \theta - 1}{\hat{y}_i} \frac{\theta}{\theta + \hat{y}_i}^{\hat{y}_i} (1 - \frac{\theta}{\theta + \hat{y}_i})^{\theta}$$

Source code and data availability

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. Data used in the analysis will be made available from the Gene Expression Omnibus for public use. DNase-seq and ChIP-seq data used for the analysis is publicly available from the ENCODE portal (https://www.encodeproject.org/). The specific accession codes used for the analysis are listed in Supplementary Table S3. GC content was downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/gbdb/hg38/bbi/gc5BaseBw/), and the mappability track was created using gem-library software[38] with a k-mer size of 100 bp and the reference human genome build hg38.

Results

We applied our peak calling algorithm to two whole human genome STARR-seq experiments, K562 and HepG2, utilizing origin of replication-based (ORI) plasmids. Using this dataset, we evaluated the quality and characteristics of identified enhancers as well as the performance of the peak caller by comparing to external enhancer datasets.

Accurate identification of highly reproducible enhancers

To evaluate the quality of enhancers identified from STARRPeaker, we uniformly called peaks from the whole human genome STARR-seq dataset using methods previously used to identify enhancers from STARR-seq data, namely BasicSTARRseq and MACS2, using recommended settings. We first compared 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 STARR-seq versus BasicSTARRseq or MACS2 (**Figure 5**). We also aggregated the transcription factor binding sites assayed by ChIP-seq around peaks, and we observed significant enrichment of transcription factor binding events compared to peaks identified by other methods. Furthermore, we compared STARRPeaker peaks and others to previously characterized enhancers by CAGE[39], MPRA[17,40], and STARR-seq[19] in HepG2 or K562 cell line (**Figure 6**). We observed a higher fraction of STARRPeaker peaks overlap with external datasets.

Discussion

We developed a statistically rigorous analysis pipeline for STARR-seq data in a software package named STARRPeaker. STARRPeaker has several key improvements over previous peak identification methods including (1) accurate quantification of STARR-seq coverage based on inferred fragment size from paired-end reads; (2) use of a negative binomial distribution to account for overdispersion in bin counts; and (3) modeling of STARR-seq coverage as a function of input and potential confounding variables in STARR-seq signal. We applied our method to two whole human genome ORI-STARR-seq datasets and demonstrated that it can unbiasedly identify a set of STARR-seq-positive regions better than previous methods. The STARR-seq peaks were enriched with epigenetic marks relevant to enhancers and overlapped better with previously known enhancers than previous methods.

To completely understand how noncoding regulatory elements can modulate transcriptional programs in human, STARR-seq active regions must be further characterized and validated within the cellular context. Currently, CRISPR-based screens are limited to a small number of selected targets. Our method can aid in prioritize candidate regions in unbiased fashion to maximize the functional characterization efforts.

Funding

We acknowledge support from the NIH and from the AL Williams Professorship funds.

Acknowledgements

We thank Jinrui Xu and Joel Rozowsky for thoughtful discussion about ChIP-seq processing, Michael Rutenberg Schoenberg and Zhen Chen for thoughtful discussion about RNA folding biology, and all other members of the Gerstein and White laboratories for advice and critical feedback on the manuscript.

Author Contributions

D.L., M.S., K.W., and M.G. conceived the project. D.L. and M.G. drafted the manuscript. D.L. developed the STARRPeaker software package. M.S., J.M., M.W., D.F., Y.K., and L.M. performed experimental work. M.W. performed experimental validation. D.L., J.Z., and J.L. performed the downstream analysis. M.G. and K.W. provided funding and supervised the project.

References

1. Muerdter, F., Boryń, Ł. M. & Arnold, C. D. STARR-seq — Principles and applications. *Genomics* **106**, 145–150 (2015).
2. Yáñez-Cuna, J. O., Kvon, E. Z. & Stark, A. Deciphering the transcriptional cis-regulatory code. *Trends Genet.* **29**, 11–22 (2013).
3. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
4. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
5. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797–803 (2005).
6. Melo, C. A. *et al.* eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Mol. Cell* **49**, 524–535 (2013).
7. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–13 (2012).
8. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).
9. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).
10. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

11. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).

12. Sethi, A. *et al.* A cross-organism framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *bioRxiv* 385237 (2018). doi:10.1101/385237

13. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

14. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).

15. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–7 (2013).

16. Liu, Y. *et al.* Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol.* **18**, 219 (2017).

17. Klein, J. C. *et al.* A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. *bioRxiv* 576405 (2019). doi:10.1101/576405

18. Johnson, G. D. *et al.* Human genome-wide measurement of drug-responsive regulatory activity. *Nat. Commun.* **9**, 5317 (2018).

19. Rathert, P. *et al.* Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. *Nature* **525**, 543–547 (2015).

20. Koohy, H., Down, T. A., Spivakov, M. & Hubbard, T. A comparison of peak callers used for DNase-Seq data. *PLoS One* **9**, e96303 (2014).

21. Uren, P. J. *et al.* Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**, 3013–20 (2012).

22. Strbenac, D., Armstrong, N. J. & Yang, J. Y. H. Detection and classification of peaks in 5' cap RNA sequencing data. *BMC Genomics* **14 Suppl 5**, S9 (2013).

23. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

24. Kharchenko, P. V, Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).

25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

26. Lai, D., Proctor, J. R. & Meyer, I. M. On the importance of cotranscriptional RNA structure formation. *RNA* **19**, 1461–1473 (2013).

27. Ringnér, M. & Krogh, M. Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.* **1**, e72 (2005).

28. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–42 (2011).

29. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–72 (2003).

30. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–56 (2012).

31. Poptsova, M. S. *et al.* Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* **4**, 4532 (2014).

32. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with

DNase I. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 6376–81 (2013).

33. Papoulis, A. & Athanasios. Probability, random variables and stochastic processes. *New York McGraw-Hill, 1984, 2nd ed.* (1984).

34. Hilbe, J. M. *Negative Binomial Regression*. (Cambridge University Press, 2011). doi:10.1017/CBO9780511973420

35. Cameron, A. C. A. & Trivedi, P. K. *Regression Analysis of Count Data*. (Cambridge University Press, 2013). doi:10.1017/CBO9781139013567

36. Hilbe, J. M. *Modeling Count Data*. (Cambridge University Press, 2014). doi:10.1017/CBO9781139236065

37. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

38. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLoS One* **7**, e30377 (2012).

39. Kawaji, H., Kasukawa, T., Forrest, A. & Carninci, P. The FANTOM 5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci. Data* 2016–2018 (2017). doi:10.1038/sdata.2017.113

40. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

Supplementary 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 replicate 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 hg38 using BWA-mem (v0.7.17). Alignments were filtered against unmapped, secondary alignments, mapping quality score less than 30, and PCR duplicates using SAMtools (v1.5) and Picard (v2.9.0). All of 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.

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

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

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

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

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

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

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

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

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

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.

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 (minQuantile = 0.9, peakWidth = 500, maxPval = 0.001, deduplicate = TRUE, 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.

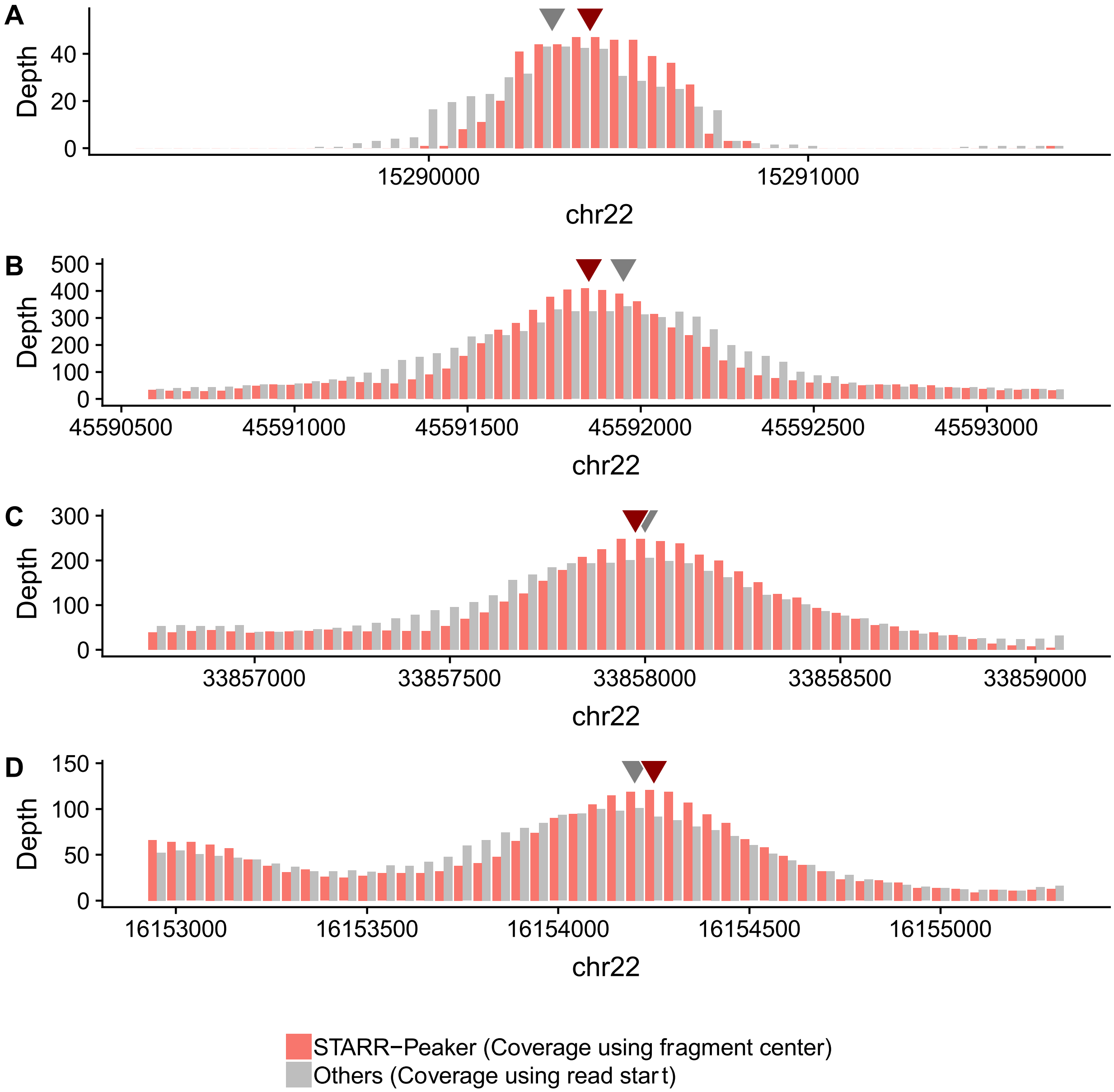
Supplementary Tables

Table S1 contains significant peaks called by STARRPeaker.

Table S2 contains various statistics from comparing STARRPeaker peaks to peaks called by BasicSTARRseq and MACS2.

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

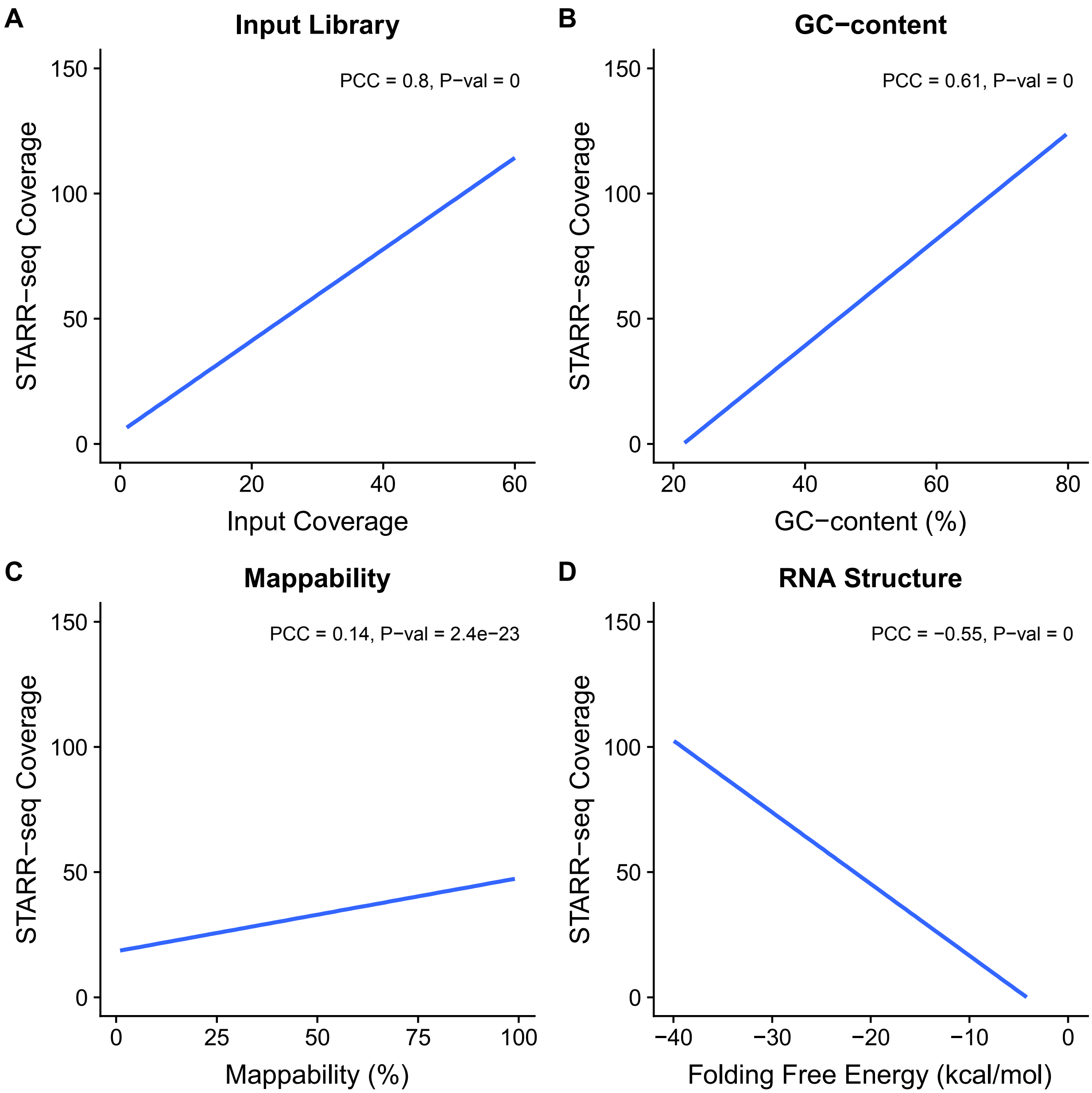
528 **Figures**



529

530 ***Figure 1** Comparison of STARR-seq coverage calculated using fragment center to*
531 *using read start position. (A)-(D) shows examples drawn from K562 STARR-seq data.*
532 *Triangle indicates the summit of coverage. Read depth was normalized, since 2 paired*
533 *reads correspond to 1 fragment.*

534



535

536 ***Figure 2*** *Confounding factors in the STARR-seq assay. STARR-seq output and input*
537 *coverages are significantly correlated with (A) input coverage (B) GC-content (C)*
538 *mappability, and (D) RNA structure folding. PCC: Pearson Correlation Coefficient. Plots*
539 *were from a sampling of 5,000 random genomic bins.*

540

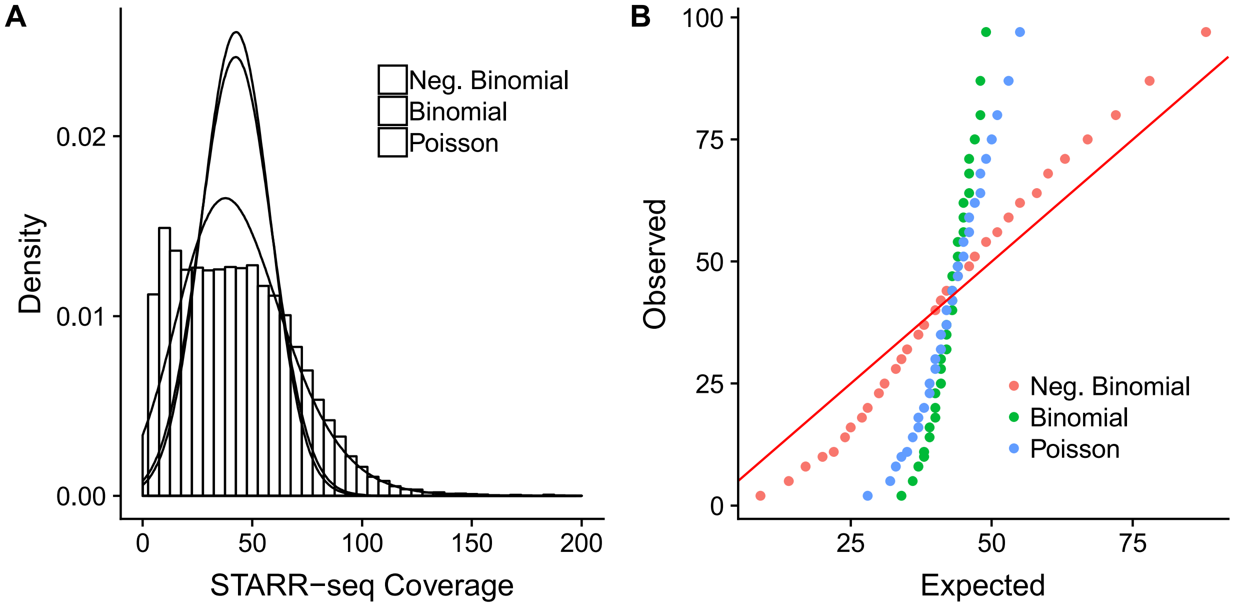


Figure 3 *STARR-seq coverage is fitted against simulated coverage using three distribution models; negative binomial, binomial, and Poisson. (A) Density histogram of simulated distribution against STARR-seq coverage. (B) Q-Q plot of simulated distribution against STARR-seq coverage. The red solid line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

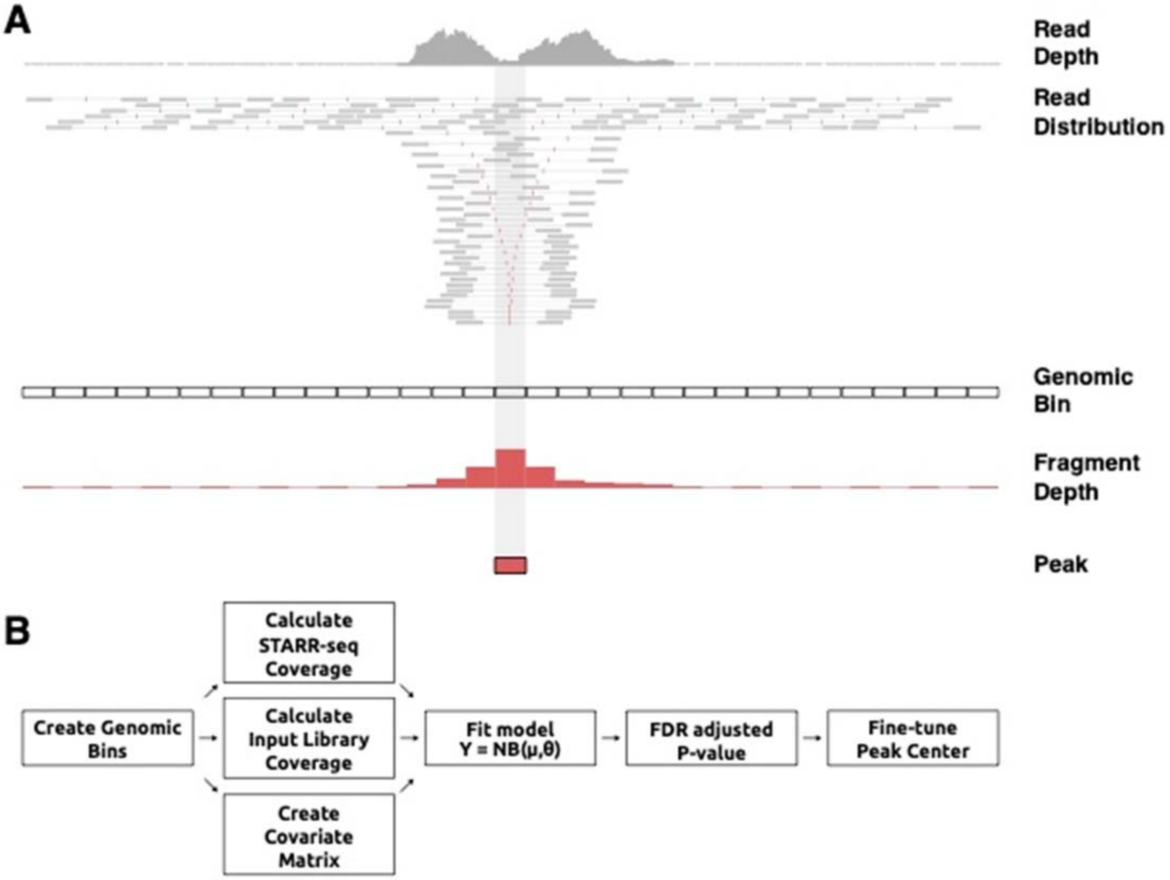


Figure 4 *Overview of STARRPeaker peak-calling scheme. (A) In contrast to using read depth (grey), fragment depth (red) offers more precise and sharper STARR-seq coverage. Fragment inserts are directly inferred from properly paired-reads. (B) Workflow of STARRPeaker describing how coverage is calculated for each genomic bin and modelled using 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 (4) Peak calling, multiple hypothesis testing correction, and adjustment of the center of peaks*

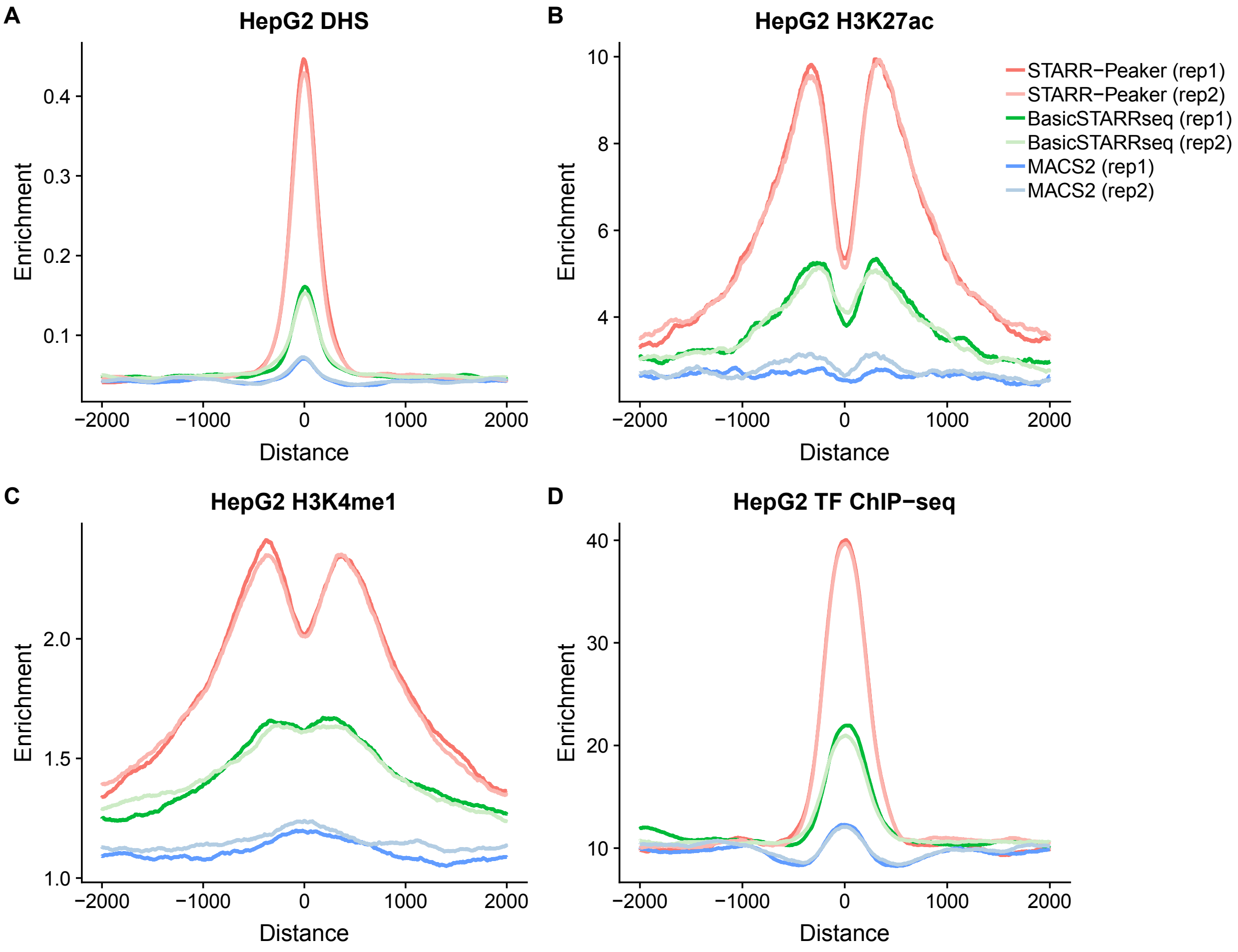


Figure 5 *Enrichment of epigenetic signals around peaks. All peaks were centered at the summit, uniformly thresholded using P-value < 0.001, and 10,000 peaks were randomly selected. Aggregated read depth at 2000 bp upstream and downstream were plotted for (A) DNase-seq (B) H3K27ac (C) H3K4me1 (D) Aggregated TF ChIP-seq profile. For TF ChIP-seq, high enrichment indicates TF binding hotspots*

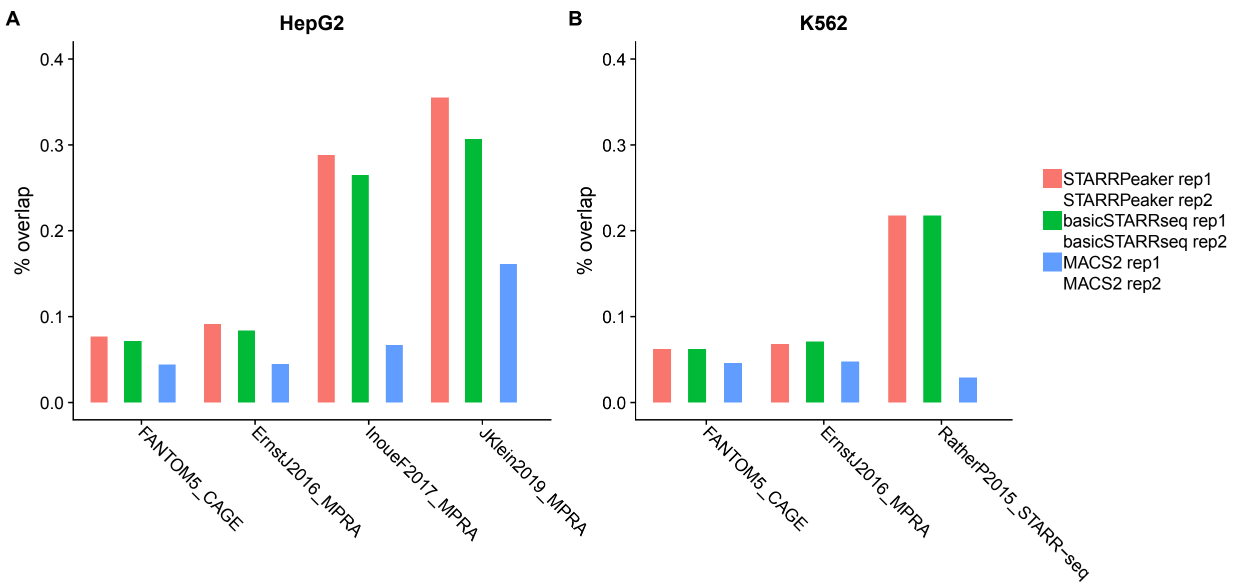


__Figure 6__ Comparison of peaks using external dataset. Peaks identified from STARRPeaker as well as BasicSTARRseq and MACS2 were compared against published dataset. For a fair comparison, 100,000 peaks were randomly drawn from peaks identified by each peak caller using the recommended settings, and the fraction of overlap was computed for each replicate. We considered it as an overlap when at least 50% of peaks intersected each other.