Lee et al. Genome Biology (2020) 21:298
https://doi.org/10.1186/s13059-020-02194-x

Genome Biology

METHOD Open Access

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

Donghoon Lee[1,2,3,4], Manman Shi[5,6], Jennifer Moran[5,6], Martha Wall[5,6], Jing Zhang[7], Jason Liu[3,4], Dominic Fitzgerald[5], Yasuhiro Kyono[5,6], Lijia Ma[5,8], Kevin P. White[5,6*] and Mark Gerstein[3,4,9,10*]

* Correspondence: kevin@tempus.com; mark@gersteinlab.org
[5]Institute for Genomics and System Biology, University of Chicago, Chicago, IL 60637, USA
[3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

Abstract

STARR-seq technology has employed progressively more complex genomic libraries and increased sequencing depths. An issue with the increased complexity and depth is that the coverage in STARR-seq experiments is non-uniform, overdispersed, and often confounded by sequencing biases, such as GC content. Furthermore, STARR-seq readout is confounded by RNA secondary structure and thermodynamic stability. To address these potential confounders, we developed a negative binomial regression framework for uniformly processing STARR-seq data, called STARRPeaker. Moreover, to aid our effort, we generated whole-genome STARR-seq data from the HepG2 and K562 human cell lines and applied STARRPeaker to comprehensively and unbiasedly call enhancers in them.

Background

The transcription of eukaryotic genes is precisely coordinated by an interplay between *cis*-regulatory elements. For example, enhancers and promoters serve as binding platforms for transcription factors (TFs) and allow them to interact with each other via three-dimensional looping of chromatin. Their interactions are often required to initiate transcription [1, 2]. Enhancers, which are often distant 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 of 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].

Consensus sequences (or canonical sequences) have been identified at certain protein binding sites, splice sites, and boundaries of protein-coding genes. However, there are no known consensus sequences that characterize enhancer function, making it

challenging to identify enhancers based on sequence alone in an unbiased fashion. The noncoding 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. Generally, STARR-seq can be genome-wide and unbiased, but the technique is limited to sequences that already exist in the genome. Although MPRA is more limited in scale, it enables testing of multiple perturbed synthetic sequences in the same system.

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 1000 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 (Fig. 1a, Additional file 1: Fig. S1A). 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 (ratio of STARR-seq output to input) and were less confident (based on statistical significance of the identified peaks) and broader in size (Fig. 1b–d, Additional file 1: Fig. S1B-D). Overall, the fragment-based coverage offered a more accurate representation of the STARR-seq readout compared to the read-based coverage counting scheme. We illustrated the benefits of using the center of the fragment in Fig. 1e, which compares the average size of peaks and fold enrichments using alternate coverage counting methods.

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 (Fig. 2 and Additional file 1: Fig. S2). We found that STARR-seq RNA coverage was significantly correlated with GC content (PCC 0.61; *P* value 1E−299) and mappability (PCC 0.45; *P* value 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* value 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].

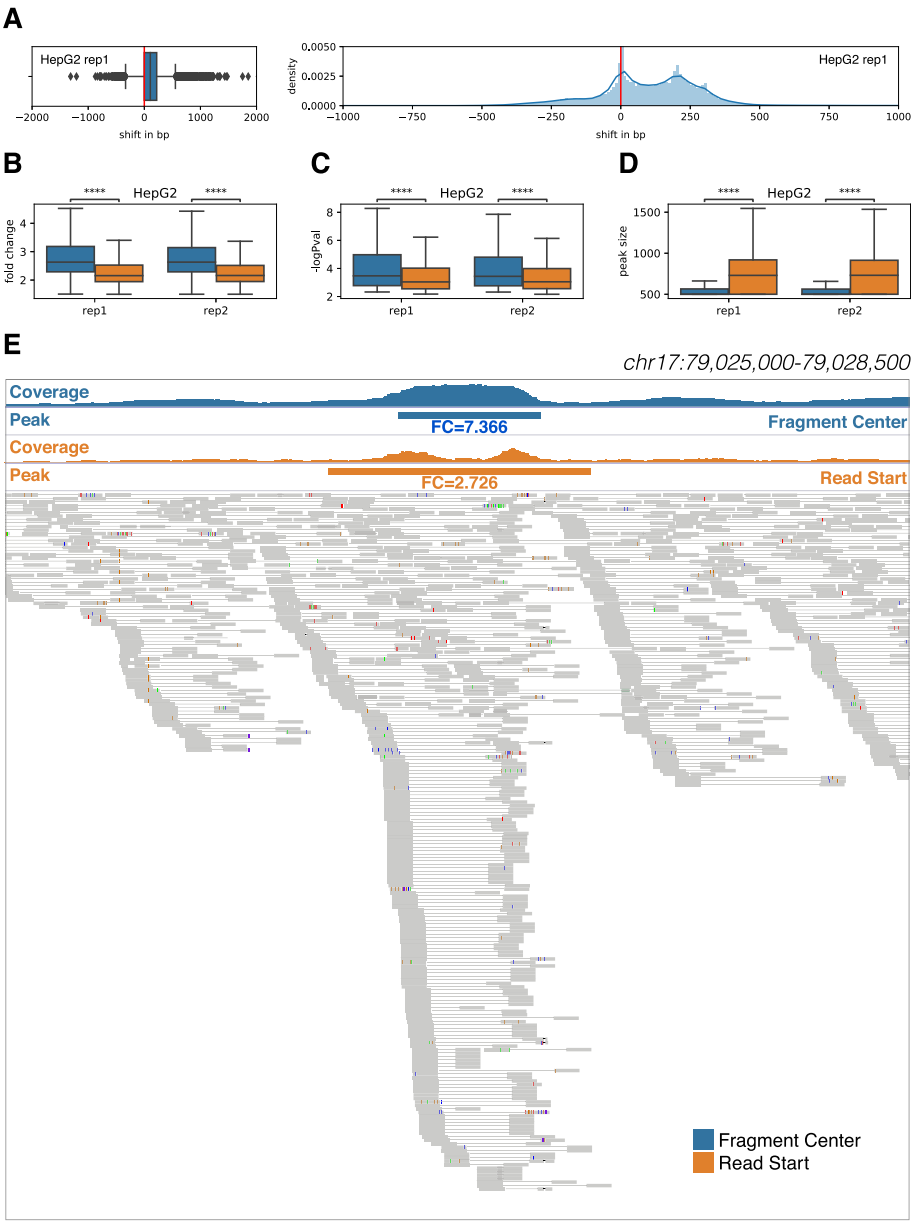

Fig. 1 Comparison of STARR-seq output coverage calculated using the center of the fragment to using the start position of the sequencing read. **a** Distribution of the shift in final peak locations resulting from using two alternative coverage counting schemes in HepG2. Comparison of **b** overall fold enrichment level, **c** *P* value, and **d** size of resulting peaks. **e** Example highlighting the difference between fragment-based and read-based coverage counting schemes and their resulting peak calls from HepG2 STARR-seq data. Asterisks represents statistical significance using the Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction; **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001, *****P* ≤ 0.0001

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

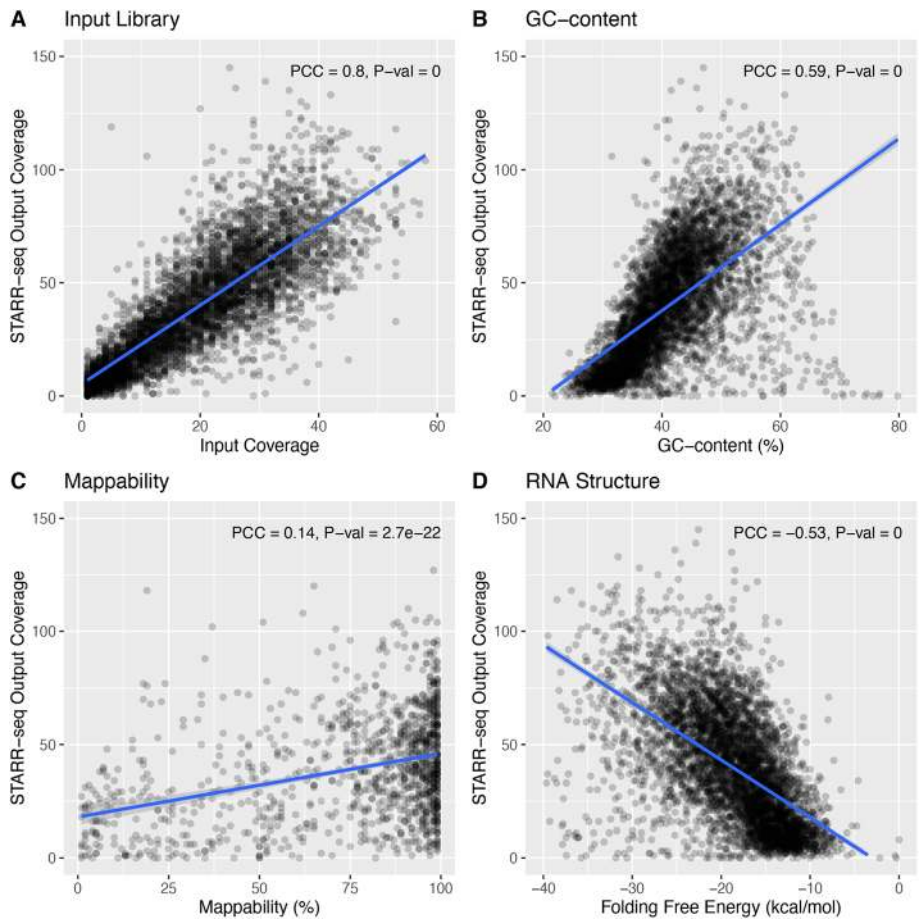

Fig. 2 Confounding factors in the STARR-seq assay. STARR-seq output and input coverages are significantly correlated with **a** input coverage, **b** GC content, **c** mappability, and **d** RNA structure folding. PCC: Pearson correlation coefficient. Plots were from a sampling of 5000 random genomic bins. The HepG2 dataset was used for the plot

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) (Additional file 1: Fig. S3).

Accurate modeling 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 fit 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 (Fig. 3a). Moreover, a Q-Q plot of expected coverage against observed coverage further demonstrated that the negative binomial model provides the best fit for the STARR-seq readout (Fig. 3b). To demonstrate the generalizability of the method and assumptions, we tested the model fit against previously published datasets that utilized different STARR-seq protocols (Additional file 1: Fig. S4). We observed consistent properties of the STARR-seq data across the different STARR-seq protocols and datasets.

In principle, we can also detect negative enrichments 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. In addition, additional experiments are necessary to demonstrate that STARR-seq can reliably detect silencers. Therefore, we focus on positive enrichments in this study, and we defer to a systemic method 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 Fig. 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 fit, 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

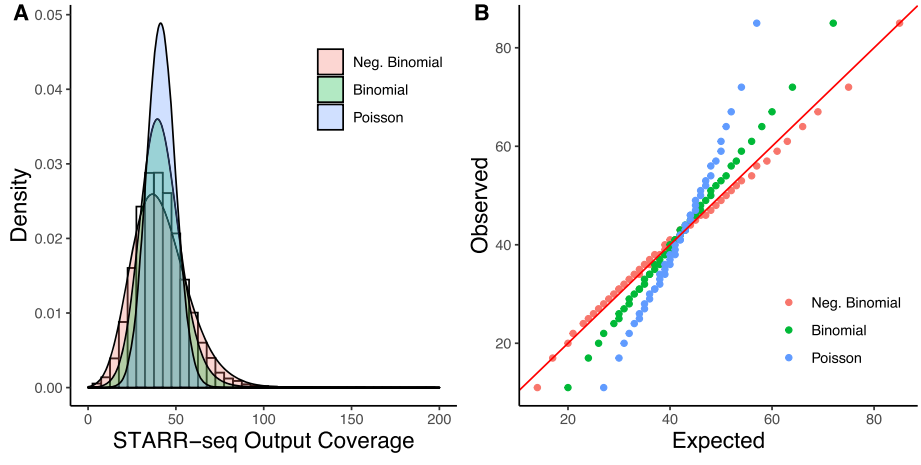

Fig. 3 STARR-seq output coverage is fit against simulated coverage using three distribution models; negative binomial, binomial, and Poisson. **a** Density histogram of simulated distribution against STARR-seq output coverage. **b** Q-Q plot of simulated distribution against STARR-seq output coverage. The red solid line represents where the observed count equals the expected count

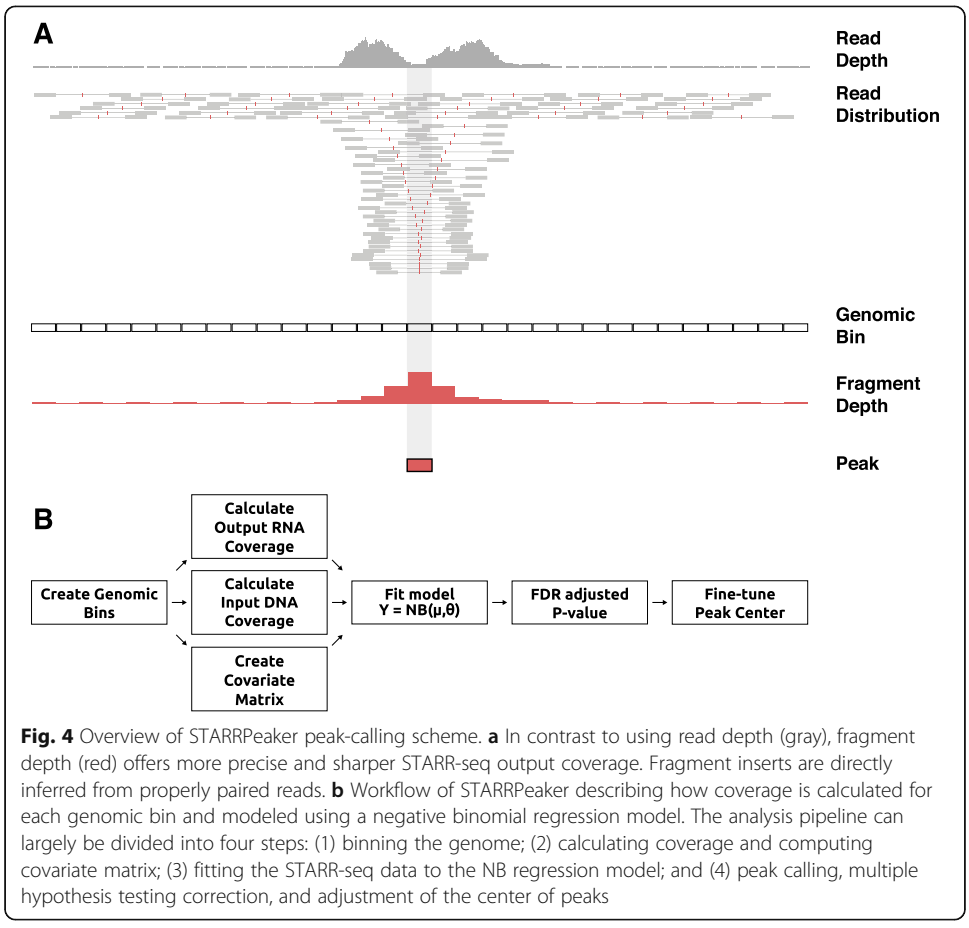

Fig. 4 Overview of STARRPeaker peak-calling scheme. **a** In contrast to using read depth (gray), fragment depth (red) offers more precise and sharper STARR-seq output coverage. Fragment inserts are directly inferred from properly paired reads. **b** Workflow of STARRPeaker describing how coverage is calculated for each genomic bin and modeled 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

summit of the direct fragment coverage. Note that the adjusted P value only refers to the unlikelihood of a candidate region being an enhancer by chance 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 $1 \le i \le n$ denotes the number of RNA fragments from a STARR-seq experiment mapped to the ith bin from the total of n genomic bins. Let t_i be the number of input library fragments (DNA) mapped to the ith bin. We define X to be the matrix of covariates, where $\overrightarrow{x_i}$ is the vector of covariates corresponding to the ith bin and x_{ij} is the jth covariate for the ith bin.

Negative binomial distribution

A negative binomial distribution, which arises from a Gamma-Poisson mixture, can be parameterized 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 parameterization 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

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

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

$$\begin{aligned} E(y_i) &= \mu_i \\ Y &\sim \mathrm{NB}(\mu, \theta) \end{aligned} \tag{3}$$

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.

$$\begin{aligned} \ln\mu_i &= \beta_0 x_{0i} + \beta_1 x_{1i} + \cdots + \beta_m x_{mi} \\ \mu_i &= \exp(\beta_0 x_{0i} + \beta_1 x_{1i} + \cdots \beta_m x_{mi}) \\ \mu_i &= \exp(\vec{x}_i^{\top}\beta) \end{aligned} \tag{4}$$

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. This mode could be beneficial in multiple scenarios. First, for a STARR-seq dataset in which the guide DNA library only contains discrete elements, direct modeling of the basal transcription rate would provide a more accurate measure of activity, especially in the absence of readouts from adjacent regions. Second, if a user ignores the effect of covariate, this mode simplifies the model and provides peaks purely based on the ratio of RNA to DNA. More details on this alternative parameterization 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.

$$\mathcal{L}_{\mathrm{NB}}(\mu|y,\theta) = \sum_{i=1}^{n} y_i \ln\left(\frac{\mu_i}{\theta+\mu_i}\right) + \theta \ln\left(\frac{\theta}{\theta+\mu_i}\right) + \ln\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\cdot\Gamma(\theta)}\right) \quad (5)$$

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

$$\mathcal{L}_{\mathrm{NB}}(\beta|y,\theta) = \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\overrightarrow{x_i}^{\top}\beta}}{\theta+e^{\overrightarrow{x_i}^{\top}\beta}}\right) + \theta \ln\left(\frac{\theta}{\theta+e^{\overrightarrow{x_i}^{\top}\beta}}\right) + \ln\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\cdot\Gamma(\theta)}\right) \quad (6)$$

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 β. 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 ith bin, y_i, under the null hypothesis.

$$P\,\mathrm{value}_i = \Pr(Y \geq y_i|H) \quad (7)$$

As defined earlier, we assume the random variable Y comes from a negative binomial distribution with the fit mean at the ith 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 $y_i - 1$.

$$\Pr(Y \geq y_i|H) = 1 - CDF(y_i - 1) = 1 - \sum_{k=0}^{y_i-1} f_Y\left(k\middle|\mu_i,\theta\right) \quad (8)$$

Substituting (1) gives

$$P\,\mathrm{value}_i = 1 - \sum_{k=0}^{y_i-1} \frac{\Gamma(k+\theta)}{\Gamma(k+1)\cdot\Gamma(\theta)} \cdot \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} \cdot \left(\frac{\mu_i}{\theta+\mu_i}\right)^{k} \quad (9)$$

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

Application of STARRPeaker

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

Initial evaluation of STARRPeaker enhancers

Our model assumes that most genomic regions have a basal level of transcription activity. To validate that our model is well calibrated under this assumption, we examined the *P* value distribution against a theoretical uniform distribution. As expected, most observed *P* values followed the null distribution, and the *P* values only deviated from expectation for very low values (Additional file 1: Fig. S5).

We processed two biological replicates from each cell type independently and assessed the correlation between each pair (see "Methods" for detail). We identified 50,389 and 52,927 candidate enhancers from HepG2 replicates 1 and 2, and by consolidating peaks from both replicates, we identified 32,929 (65.3% of rep 1; 62.2% of rep2) reproducible candidate enhancers from the HepG2 cell line. Similarly, by consolidating 30,194 and 41,810 candidate enhancers from K562 replicates 1 and 2, we identified 20,471 (67.8% of rep 1; 49.0% of rep 2) reproducible candidate enhancers from the K562 cell line (Additional file 2: Table S1). Overall, we observed high correlation of RNA fragment coverage between two replicates (PCC = 0.99 for both HepG2 and K562; see Additional file 1: Fig. S6). This result indicates the correlation is dominated by negatives. 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 (36.5%) and 11,420 (55.8%) 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 allows one to identify stable peaks from suboptimal datasets. To evaluate the robustness of the STARRPeaker method, we generated random subsets of HepG2 whole-genome STARR-seq library after aligning the reads and compared the quality of the peak calls. We subsampled randomly at various rates from 20 to 80% of the total dataset. We found that STARRPeaker was able to reliably identify approximately 90% of the candidate enhancers (consolidated) using 80% of the original sequencing library and 80% of the candidate enhancers using 40% of the original sequencing library (Additional file 1: Fig. S7A). When we focused on strong enhancer candidates, approximately 98% of the top 5000 enhancers were recovered using only 60% of the original sequencing library (Additional file 1: Fig. S7B). 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 series of binomial tests on the candidate enhancers we identified and evaluated for possible orientation-specific activities (see "Methods"). We observed a small fraction of candidate enhancers showing orientation bias [2.58% for HepG2 ($n = 850$); 4.49% for K562 ($n = 919$); FDR ≤ 0.01] in both replicates (Additional file 1: Fig. S8). Furthermore, only a few candidates showed extreme bias [$n = 3$ for HepG2; $n = 4$ for K562; > 90% fragments on one strand]. Thus, true orientation-dependent activities are unlikely in our STARR-seq data, but that the orientation may have an effect on the efficiency. These findings provide 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 alternative methods, namely BasicSTARRseq and MACS2.

First, we qualitatively assessed the peak-calling algorithms using a simulated dataset where the ground truth exists. The simulation was primarily designed for positives in the dataset, and as a result, it emphasizes the sensitivity (as opposed to specificity) of each method. We created an artificial STARR-seq dataset that contains 28 spike-in controls; a hybrid of DNA input and RNA output libraries with active elements at predefined loci (Additional file 7: Table S6; see "Methods" for details). All three methods successfully identified peaks at 28 control regions (Additional file 1: Fig. S9). However, we noticed that BasicSTARRseq peaks were fragmented and off-centered from the controls due to its limitations of fixed peak size and read-based coverage calculation. As a result, several false-positive peaks were called adjacent to the controls. We calculated the sensitivity and specificity of each method using 28 controls as the gold standard (Additional file 8: Table S7). We considered a result positive if at least 80% of the peak overlapped with the control, and we assumed the presence of approximately 1000 true negative regions. Both STARRPeaker and MACS2 had 100% sensitivity and specificity, whereas BasicSTARRseq had 75% sensitivity and 94.9% specificity.

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 (Additional file 5: Table S4). While it is uncertain how many true enhancers were present in each sample, we had to ensure that we make 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. Suppose each method is generating a set of randomized peaks, then the method with the greater number of peaks is likely to have more overlap solely by chance. To eliminate this artifact, we used a uniform P value threshold of 0.001 and subsampled the peaks down to the same number 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 (Fig. 5, Additional file 1: Fig. S10). 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] (Fig. 6, Additional file 3: Table S2). Since most of the previous enhancer assays were not at a genome-wide scale and were limited in scope to specific loci, we focused on how many previously characterized enhancers were recovered using different methods. 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

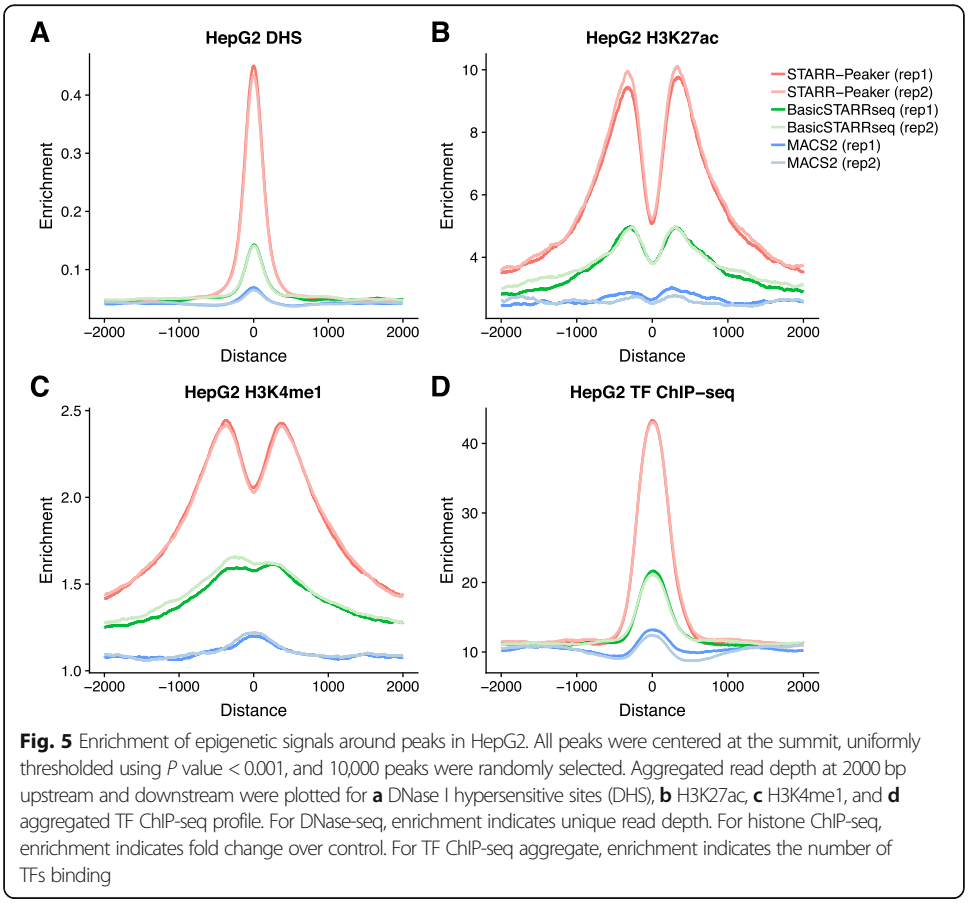

Fig. 5 Enrichment of epigenetic signals around peaks in HepG2. All peaks were centered at the summit, uniformly thresholded using *P* value < 0.001, and 10,000 peaks were randomly selected. Aggregated read depth at 2000 bp upstream and downstream were plotted for **a** DNase I hypersensitive sites (DHS), **b** H3K27ac, **c** H3K4me1, and **d** aggregated TF ChIP-seq profile. For DNase-seq, enrichment indicates unique read depth. For histone ChIP-seq, enrichment indicates fold change over control. For TF ChIP-seq aggregate, enrichment indicates the number of TFs binding

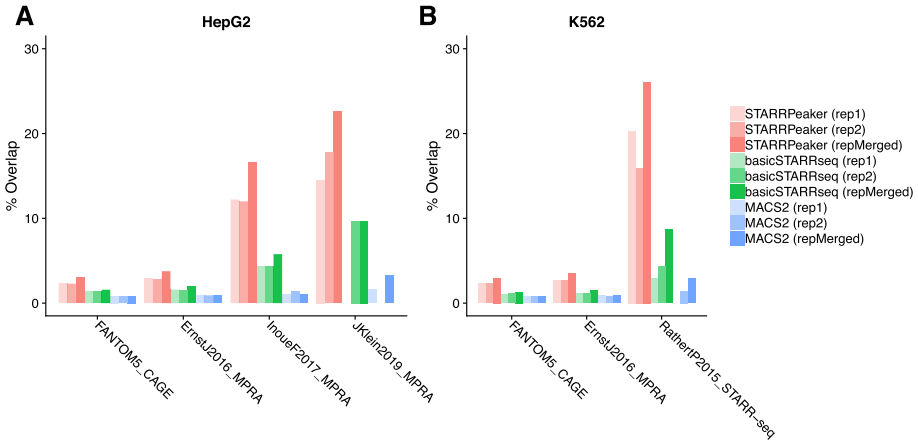

Fig. 6 Comparison of peaks using an external dataset for **a** HepG2 or **b** K562 cell lines. Peaks identified from STARRPeaker as well as BasicSTARRseq and MACS2 were compared against a published dataset. For a fair comparison, all peaks were centered at the summit, uniformly thresholded using *P* value < 0.001, and 20,000 peaks were randomly drawn from peaks identified by each peak caller using the recommended settings. The fraction of overlap was computed for each replicate. We used the total number of peaks in each dataset as the denominator. We considered it an overlap when at least 50% of peaks intersected each other

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 a CRISPRi tiling screen [43]. Using the nine enhancers as the ground truth and assuming there are approximately 1300 potential negative elements in the span of a 1.3-Mb genomic sequence, we performed sensitivity and specificity analyses of three competing methods. We think this is a substantial dataset to benchmark the performance as it provides an orthogonal line of evidence for true in vivo enhancers and highlights that the most genomic regions are dominated by negatives. As speculated, we found that STARRPeaker had the highest specificity. While both BasicSTARRseq and MACS2 identified a few more enhancers, their false-positive rate was much higher than that of STARRPeaker (Additional file 1: Fig. S11, Additional file 9: Table S8). Furthermore, upon close examination of three enhancers that were missed by STARRPeaker, we observed the coverage was much lower than other regions identified as enhancers.

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 (Additional file 1: Fig. S12A). Moreover, STARRPeaker recovered 77.5% ($n = 7451$) of published peaks, which were called using BasicSTARRseq and then further shortlisted using a stringent threshold (P value 1E−5 with corrected enrichment ≥ 4). When we compared the quality of STARRPeaker peaks ($n = 16{,}150$) to the published post hoc filtered peaks ($n = 9610$), we found that STARRPeaker peaks were highly enriched with chromatin accessibility signals despite having 6540 additional peaks from the same HeLa-S3 dataset (Additional file 1: Fig. S12B). When we applied the same post hoc filtering approach to STARRPeaker peaks, the chromatin accessibility enrichment was further improved (Additional file 1: Fig. S12C).

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 [44] 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 52,857 regions with significant enhancer activities from the five replicates they produced. Approximately half of STARRPeaker peaks overlapped with the published results ($n = 26{,}318$). While it is debatable to claim that one method is superior to the other, this result clearly demonstrates that STARRPeaker can be applied to 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 on K562 [19]. The dataset covers approximately 91% of the surrounding 3 Mb of the MYC locus. Consistent with the earlier analysis, we observed that STAR RPeaker 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 (Additional file 1: Fig. S13). 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 [45, 46]. The next step for CRISPR-based functional screens is to overcome the current limitation of a small scale by leveraging barcodes and single-cell sequencing technology [47]. 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 s) 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 s; 10 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s; 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 (Thermo Scientific), 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 12.5-L 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 h after electroporation and extracted total RNA using an RNeasy Maxi kit (Qiagen). We further isolated polyA-plus mRNA using Dynabeads® Oligo (dT) kit (Thermo Fisher Scientific), treated it with TURBO DNase (Invitrogen), and purified the reaction using an RNeasy MinElute Kit (Qiagen). We synthesized cDNA using SuperScript III (Thermo Fisher 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 (Thermo Fisher 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.

Fitting of distributions to the STARR-seq dataset

To build a statistical model that best describes the STARR-seq readout, we tested fitting of various univariate distributions to the output coverages of the STARR-seq dataset. To eliminate the influence of input coverage on output coverage, we subsampled bins with an input coverage value of 20 (approximately a median input coverage for both the HepG2 and K562 datasets) and used their output coverage values for the underlying observed distribution. We fit binomial, Poisson, and negative binomial distributions and estimated parameters using MLE. For binomial distributions, we assumed the number of the trial as the sum of the STARR-seq input and output coverage, and the probability of success as the sum of the STARR-seq output coverage

divided by the total input and output coverage, as described previously [15]. The Poisson distribution is described by a single parameter λ, whereas the negative binomial distribution is described by both μ and the shape parameter θ. For example, we found $\lambda = 41.6$, $\mu = 41.6$, and $\theta = 8.9$ for the HepG2 dataset. Based on estimated parameters, we plotted the expected distribution quantiles from 0 to 100 percentiles against the actual observed quantiles. We plotted a straight diagonal line to show how well each distribution fit with the actual observed data.

Negative binomial distribution

A negative binomial distribution, which arises from Gamma-Poisson mixture, can be parameterized for $y \geq 0$ 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 parameterization of negative binomial regression using a rate model

Alternative parameterization allows STARR-seq data to be modeled 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 ($\overrightarrow{x}$).

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

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. This alternate model is useful if constant basal transcription is expected throughout the genome or if covariates are available for directly modeling the basal transcription rate π.

Evaluation of potential orientation bias

For all enhancer peaks identified from both the HepG2 and K562 cell lines, we evaluated if there was an overrepresentation of fragments in a specific orientation. If there is no orientation bias, the STARR-seq active region should be equally represented by both forward- and reverse-stranded fragments. We performed a binomial test for the statistical significance on how unlikely it is to have a fragment distribution skewed on one strand.

$$\binom{n}{k} p^k (1-p)^{n-k}$$

where p is 0.5 (equal chance of being either forward or reverse stranded), n is number of all supporting fragments, and k is the number of forward-stranded fragments. We adjusted P values using FDR (BH) and used 0.01 as a cutoff. We ran binomial tests for both replicates. To call a region orientation-biased, we ensured that genomic DNA fragments were equally represented in both forward and reverse strands, and we checked for significant strand bias in both replicates. If strand bias is already present in a genomic DNA library, it is more likely that the bias will be due to amplification by PCR rather than being a result of orientation-specific activity.

BasicSTARRseq

We used BasicSTARRseq R package version 1.10.0 downloaded from Bioconductor (https://bioconductor.org/packages/release/bioc/html/BasicSTARRseq.html). We used the default settings 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] with the optimal parameters suggested for a human STARR-seq dataset (-f BAMPE -g hs). We also used an option to allow duplicates in read (--keep-dup all), since our STARR-seq dataset was multiplexed. We called peaks with an FDR cutoff of 0.05 (-q 0.05), as recommended by the author of the software.

Calculating folding free energy

We used the LinearFold [48] algorithm to estimate the folding energy of each genomic bin iteratively across the whole genome. Specifically, we used the Vienna RNAfold thermodynamic model [49] with parameters from Mathews et al. [50]. We implemented a parallel processing scheme to leverage multicore processors to expedite the calculation of folding free energy.

Recommended parameters for the model

We determined the model parameters from the experimental design of the STARR-seq assay. Our STARR-seq input library was based on DNA fragments that were size-selected on an agarose gel for 500 bp. Therefore, we defined the bin size to be 500 bp with a step size of 100 bp. Furthermore, we chose the minimum and maximum template size to be 200 and 1000 bp, respectively, to recover RNA generated from the STARR-seq input library. We recommend setting the minimum template size no less than 200 bp, as a shorter template can become more prone to PCR bias during sequencing. Based on the sequencing depth, we determined the minimum coverage for the whole genome to be 10. To guide users on how to choose these parameters, we provide a sensitivity analysis of how changing the parameters affects the results (Additional file 6: Table S5). Overall, more than 80% of peaks were consistent regardless of varying the parameters. We found that changing the bin size affected the average peak size and

changing the step size affected the resolution. However, we found that using a 50-bp step size approximately doubled the processing time.

Simulation of the STARR-seq dataset

We created an artificial STARR-seq dataset where the ground truth exists. We used the original STARR-seq input library as the base of the simulation. We focused on an approximately 2.6-Mbp region spanning the MYC locus (chr8:127128459-129731914). We artificially selected 28 control regions equally sized at 500 bp. Among these control regions, we included coordinates of four in vivo MYC enhancers identified from Fulco et al. (MYC-e1, chr8:127898897-127898963; MYC-e2, chr8:127960392-127960502; MYC-e5, chr8:129581781-129582461; MYC-e6, chr8:129689361-129689694) [43]. We selectively generated paired-end reads that only support control regions. Since we expect the background regions to have a basal level of transcription, we matched the read distribution of the input library to make the transcription rate equal to 1. All 28 control regions were then merged to create an artificial STARR-seq output library, and finally, the read coverage was visually inspected to ensure that it resembles one from the actual STARR-seq dataset.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-020-02194-x.

Additional file 1: Supplementary Figure S1. 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 S2.** (Shadow figure of Fig. 2) Correlation of STARR-seq output coverage with covariates (STARR-seq input coverage, GC content, mappability, and RNA structure) in various STARR-seq datasets, including (A) K562, (B) HeLa-S3 untreated WG-STARR-seq (Muerdter 2018), (C) GM12878 HiDRA ATAC-STARR-seq (Wang 2018), and (D) K562 Cap-STARR-seq (Rathert 2015). PCC: Pearson Correlation Coefficient. **Supplementary Figure S3.** Contribution of covariates and model selection. (A) Q-Q plots of residuals for 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. DNA: genomic DNA fragment coverage; GC: GC-bias; MAP: mappability, FOLD: folding free energy; AIC: Akaike information criterion; BIC: Bayesian information criterion. **Supplementary Figure S4.** (Shadow figure of Fig. 3) Fitting of various STARR-seq datasets using three distribution models: negative binomial, binomial, and Poisson. Datasets included (A) K562, (B) HeLa-S3 untreated WG-STARR-seq (Muerdter 2018), (C) GM12878 HiDRA ATAC-STARR-seq (Wang 2018), and (D) K562 Cap-STARR-seq (Rathert 2015). **Supplementary Figure S5.** Q-Q plots of the *P* value distribution. (A) HepG2 and (B) K562. The red line is a reference line where the expected *P* values match the observed ones. **Supplementary Figure S6.** Correlation of RNA fragment coverage between replicates for (A) HepG2 or (B) K562 cell lines. Pairwise coverage was calculated using 1000 bp bins across the genome. The X- and Y-axis represent the natural log of fragment counts. A pseudo count of 1 was added to the fragment counts for plotting only. Bins with abnormally high fragment counts were removed to avoid inflation of the Pearson correlation. We used the median absolute deviation method with a scaling factor of 200 to filter extremely large deviations from the median. **Supplementary Figure S7.** Comparison of peaks called from subsamples of the original STARR-seq library, highlighting the robustness of STARRPeaker using (A) the whole dataset and (B) 5000 subset of peaks. **Supplementary Figure S8.** Orientation biases analysis for (A-B) HepG2 or (C-D) K562 cell lines. The ratio between forward and reverse stranded fragments was tested for statistical significance using a binomial test. Orange dots represent peaks with significant strand bias (FDR q-value < 0.01). **Supplementary Figure S9.** Comparison of peaks identified by various methods using a simulated STARR-seq dataset containing 28 spike-in control regions. **Supplementary Figure S10.** Enrichment of epigenetic signals around peaks in K562. 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 I hypersensitive sites (DHS), (B) H3K27ac, (C) H3K4me1, and (D) aggregated TF ChIP-seq profile. For DNase-seq, enrichment indicates unique read depth. For histone ChIP-seq, enrichment indicates fold change over control. For TF ChIP-seq aggregate, enrichment indicates the number of TFs binding. **Supplementary Figure S11.** (A-C) Genome browser session comparing STARRPeaker to other peak-calling methods at validated distal enhancers from CRISPRi tiling screen. **Supplementary Figure S12.** Application of STARRPeaker on an external HeLa-S3 dataset. (A) Comparison of chromatin accessibility (DNase-seq) for STARRPeaker peaks between untreated and inhibitor-treated samples. (B) Comparison of STARRPeaker peaks to published results without post-hoc filtering. STARRPeaker found 6540 additional peaks that were equally enriched with chromatin accessibility signals. (C) Comparison of STARRPeaker peaks to published results with the same post-hoc filtering approach (P-value ≤1E-5 with corrected enrichment ≥4).

Supplementary Figure S13. Venn diagram for four-way comparison of peaks identified by various methods using a published dataset from Rathert et al. 2015.

Additional file 2: Supplementary Table S1. Peaks identified by STARRPeaker in each replicate and merged replicate from HepG2 and K562 cell lines.

Additional file 3: Supplementary Table S2. Number of overlaps between peaks from various peak callers (STARRPeaker, BasicSTARRseq, and MACS2) and previously published enhancers identified using alternative enhancer assays.

Additional file 4: Supplementary Table S3. List of data sources and accession numbers used for the analysis.

Additional file 5: Supplementary Table S4. Comparison of peaks identified from various peak callers (STARRPeaker, BasicSTARRseq, and MACS2).

Additional file 6: Supplementary Table S5. Sensitivity analysis of using alternative parameters against default parameters (bin 500 bp, step 100 bp, min 200 bp, max 1000 bp) using the HepG2 rep1 dataset.

Additional file 7: Supplementary Table S6. List of control regions in the MYC locus for simulation.

Additional file 8: Supplementary Table S7. Sensitivity and specificity analysis using simulated STARR-seq data.

Additional file 9: Supplementary Table S8. Sensitivity and specificity analysis using the Fulco 2016 CRISPRi dataset.

Additional file 10. Review history.

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.

Review history
The review history is available as Additional file 10.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' 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 works. M.W. and Y.K. performed experimental validations. D.L., J.Z., and J.L. performed the downstream analyses. M.G. and K.W. provided funding and supervised the project. The authors read and approved the final manuscript.

Authors' information
Twitter handles: @hoondy (Donghoon Lee), @MarkGerstein (Mark Gerstein).

Funding
We acknowledge support from the National Institutes of Health (U24HG009446, UM1HG009426, K01MH123896) and from the AL Williams Professorship funds. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
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 [51]. 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 [52]. 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 analysis are listed in Additional file 4: 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 [53] with a k-mer size of 100 bp and the reference human genome build hg38.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [4]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [5]Institute for Genomics and System Biology, University of Chicago, Chicago, IL 60637, USA. [6]Tempus Labs, Inc., Chicago, IL 60654, USA. [7]School of Information and

Computer Sciences, University of California, Irvine, CA 92697, USA. [8]School of Life Sciences, Westlake University, Hangzhou 310024, Zhejiang, China. [9]Department of Computer Science, Yale University, New Haven, CT 06520, USA. [10]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA.

Received: 2 August 2019 Accepted: 4 November 2020
Published online: 08 December 2020

References

1. Muerdter F, Boryń ŁM, Arnold CD. STARR-seq—principles and applications. Genomics. 2015;106:145–50. https://doi.org/10.1016/J.YGENO.2015.06.001.
2. Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. Trends Genet. 2013;29:11–22. https://doi.org/10.1016/j.tig.2012.09.007.
3. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003;12:1725–35. https://doi.org/10.1093/hmg/ddg180.
4. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981;27(2 Pt 1):299–308. https://doi.org/10.1016/0092-8674(81)90413-x.
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. 2005;132:797–803. https://doi.org/10.1242/dev.01613.
6. Melo CA, Drost J, Wijchers PJ, van de Werken H, de Wit E, Vrielink JAFO, et al. eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell. 2013;49:524–35. https://doi.org/10.1016/j.molcel.2012.11.021.
7. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13. https://doi.org/10.1038/nature11279.
8. Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Genet. 2017;49:1073–81. https://doi.org/10.1038/ng.3884.
9. Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods. 2017;14:629–35. https://doi.org/10.1038/nmeth.4264.
10. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6. https://doi.org/10.1038/nmeth.1906.
11. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9:473–6. https://doi.org/10.1038/nmeth.1937.
12. Sethi A, Gu M, Gumusgoz E, Chan L, Yan K-K, Rozowsky J, et al. A cross-organism framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation. bioRxiv. 2018:385237. https://doi.org/10.1101/385237.
13. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012;30:265–70. https://doi.org/10.1038/nbt.2136.
14. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30:271–7. https://doi.org/10.1038/nbt.2137.
15. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science (80-). 2013;339:1074–7. https://doi.org/10.1126/science.1232542.
16. Liu Y, Yu S, Dhiman VK, Brunetti T, Eckart H, White KP. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. Genome Biol. 2017;18:219. https://doi.org/10.1186/s13059-017-1345-5.
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. 2019:576405. https://doi.org/10.1101/576405.
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;9:1–9.
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. https://doi.org/10.1038/nature14898.
20. Koohy H, Down TA, Spivakov M, Hubbard T. A comparison of peak callers used for DNase-Seq data. PLoS One. 2014;9:e96303. https://doi.org/10.1371/journal.pone.0096303.
21. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, et al. Site identification in high-throughput RNA-protein interaction data. Bioinformatics. 2012;28:3013–20. https://doi.org/10.1093/bioinformatics/bts569.
22. Strbenac D, Armstrong NJ, Yang JYH. Detection and classification of peaks in 5′ cap RNA sequencing data. BMC Genomics. 2013;14(Suppl 5):S9. https://doi.org/10.1186/1471-2164-14-S5-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. 2008;9:R137. https://doi.org/10.1186/gb-2008-9-9-r137.
24. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008;26:1351–9. https://doi.org/10.1038/nbt.1508.
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. https://doi.org/10.1093/bioinformatics/btp352.
26. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep. 2014;4:4532. https://doi.org/10.1038/srep04532.
27. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci U S A. 2013;110:6376–81. https://doi.org/10.1073/pnas.1216822110.
28. Lai D, Proctor JR, Meyer IM. On the importance of cotranscriptional RNA structure formation. RNA. 2013;19:1461–73. https://doi.org/10.1261/rna.037390.112.

29. Ringnér M, Krogh M. Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. PLoS Comput Biol. 2005;1:e72. https://doi.org/10.1371/journal.pcbi.0010072.
30. Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol. 2011;29:436–42. https://doi.org/10.1038/nbt.1861.
31. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res. 2003;13:1863–72. https://doi.org/10.1101/gr.1272403.
32. Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, Wakamatsu A, et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. 2012;22:947–56. https://doi.org/10.1101/gr.130559.111.
33. Papoulis A. Probability, random variables and stochastic processes. 2nd ed. New York: McGraw-Hill; 1984. p. 1984. http://adsabs.harvard.edu/abs/1984prvs.book.....P.
34. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52:1–10. https://doi.org/10.1038/s41588-020-0578-5.
35. Hilbe JM. Negative Binomial Regression. Cambridge: Cambridge University Press; 2011. https://doi.org/10.1017/CBO9780511973420.
36. Cameron ACA, Trivedi PK. Regression analysis of count data. Cambridge: Cambridge University Press; 2013. https://doi.org/10.1017/CBO9781139013567.
37. Hilbe JM. Modeling count data. Cambridge: Cambridge University Press; 2014. https://doi.org/10.1017/CBO9781139236065.
38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
39. Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, et al. Resolving systematic errors in widely used enhancer activity assays in human cells. Nat Methods. 2018;15:141–9. https://doi.org/10.1038/nmeth.4534.
40. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14:288–95. https://doi.org/10.1038/nrg3458.
41. 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. 2017;4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5574373/.
42. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 2017;27:38–52. https://doi.org/10.1101/gr.212092.116.
43. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science (80-). 2016; http://science.sciencemag.org/content/early/2016/10/05/science.aag2445.
44. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun. 2018;9:1–15.
45. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016;167:1853–1866.e17.
46. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. Mol Cell. 2017;66:285–299.e5.
47. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176:377–390.e19.
48. Huang L, Zhang H, Deng D, Zhao K, Liu K, Hendrix DA, et al. LinearFold: linear-time approximate RNA folding by 5′-to-3′ dynamic programming and beam search. Bioinformatics. 2019;35:i295–304. https://doi.org/10.1093/bioinformatics/btz375.
49. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.
50. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A. 2004;101:7287–92.
51. Lee D. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. Github. 2020; http://github.com/gersteinlab/starrpeaker.
52. Lee D, Shi M, Moran J, Wall M, Zhang J, Liu J, et al. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. ENCODE Project. 2020; https://www.encodeproject.org/functional-characterization-experiments.
53. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome mappability. PLoS One. 2012;7:e30377. https://doi.org/10.1371/journal.pone.0030377.

Publisher's Note