nature portfolio

https://doi.org/10.1038/s41587-022-01211-7

Supplementary information

A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers

Supplementary Notes

Enrichment strategies of different assays included in this study

These strategies include isolation of nuclei where the transcriptionally-engaged RNA polymerases (yellow ovals in Fig. 1b) remain associated with genomic DNA (black lines in Fig. 1b) and the nascent transcripts (purple curved lines in Fig. 1b), which can also be achieved by inhibition of RNA Pol II with α-amanitin or by halting RNA Pol II on genomic DNA using UV-C induced crosslinking following BromoU (BrU) incorporation in the nascent transcripts. mNET-seq utilizes MNase digestion of genomic DNA and transcripts, followed by immuno-capture of RNA Pol II and their associated RNAs. Those RNAs are then size-selected before sequencing. PRO-seq and PRO/GRO-cap incorporate biotinylated nucleotides or Br-UTP into the nascent transcripts via nuclear run-on reaction, which enables a highly selective antibody- or streptavidin-based enrichment of those nascent transcripts afterward. PRO/GRO-cap further employs an additional enzymatic treatment to remove the uncapped RNA molecules. Start-seq and csRNA-seq enrich short RNA molecules by size selection. Those short RNA molecules are then subjected to cap selection. NET-CAGE and CAGE methods utilize cap trapper and affinity-based selection of capped RNAs. STRIPE-seq involves enzymatic (terminator exonuclease, TEX) removal of uncapped RNAs and library preparation with template switching. Finally, Bru-seq and BruUV-seq incorporate BrU into the nascent RNAs via a pulse-chase strategy, which allows for an antibody-based selection of the nascent transcripts.

Evaluation of the impact of technical artifacts in nascent transcript sequencing assays

When using these assays to identify and analyze genome-wide divergently transcribed enhancers, it is critical to estimate the proportion of non-strand-specific reads (i.e., noise). However, extra cautions must be taken when applying this strategy to assays sensitive to unstable transcripts because promoter-upstream transcripts (PROMPTs)[1] of some isoforms of the same gene can overlap with the exons of others which generally initiate in the antisense direction. To address this problem, we found that by limiting the regions of evaluation to the first exon of the transcript (colored in green in Extended Data Fig. 2a) with the transcription start site (TSS) location downstream of all other TSSs of the same gene with the following additional refinement: 1) exons overlapping with other genes were removed; 2) the annotations of TSSs were consistent between GENCODE and RefSeq; 3) exonic regions overlapping with Enhancer-Like Sequences predicted by cCRE (agnostic version)[2] were excluded. We then evaluated the strand specificity of three canonical stranded and unstranded RNA-seq libraries according to the refinement stated above. As expected, the stranded libraries have strand specificities around 1 (mean: 0.993, SD: 0.001), while the unstranded libraries have strand specificities around 0.5 (mean: 0.502, SD: 0.001). We then evaluated the strand specificity of all 13 assays with the same approach. The results show that all of them have great strand specificities (average: 0.984, SD: 0.019, Extended

Data Fig. 2b and 2c), suggesting that inversion or misassignment due to experimental artifacts are negligible in all of these assays.

We also surveyed the abundance of signals from internal priming in reverse transcription procedure across assays and found that all assays have internal priming rates lower than 1% (Extended Data Fig. 2d and 2e). For this part, flush end sites[3] were first identified by scanning for genomic positions with signal intensity higher than 10 Reads Per Million (RPM) and a fold change greater than 5 compared to their immediate adjacent positions, excluding positions overlapping with any known splice junctions. Flush end sites which match the last three nucleotides of the RT primer (RT3-mer) were considered as potential internal priming sites. Internal priming rates were then estimated as the proportion of reads from mispriming sites. To assess confidence about these estimations, we further compared the log odds ratio of observed RT3-mer at flushing end sites versus in the whole genome as follows:

$$LOR = \ln\left(\frac{p(\text{RT3-mer}|\text{flush end})p(\text{other 3-mer}|\text{genome})}{p(\text{other 3-mer}|\text{flush end})p(\text{RT3-mer}|\text{genome})}\right)$$

Examination of the effect of cap enrichment on eRNA capturing efficiency

To demonstrate the effect of cap enrichment strategies, e.g., enzymatic cap selection, cap trapper, on the efficiency of capturing eRNAs, we used a previous run-on dataset[4], where two types of paired-end libraries were generated (among other libraries): 1) with cap selection (i.e., the regular PRO-cap protocol, referred to as the Capped group); 2) without cap selection (i.e., contains both capped and uncapped RNAs, referred to as the RppH group). Both libraries were downsampled three times so that they all have a matched number of mapped reads. A table of read counts was generated that each downsampled library could cover among the “CRISPR-identified Enhancer Set”. The table was further aggregated by calculating the average of read counts among downsampled libraries per group. A direct comparison of log-transformed read coverages among these reference loci was performed. In addition, a series of cutoffs of read counts were then used to determine whether an enhancer locus could be classified as covered, and the percentage of enhancers covered by the RppH group and can also be covered in the Capped group were reported.

Estimation of run-on capturing efficiency on paused Pol II

To demonstrate the effectiveness of run-on procedures on paused Pol II, we retrieved POLR2A ChIP-exo libraries from a previous study (accession number: GSE108323)[5] and calculated the pausing indexes on GENCODE-annotated transcripts with HOMER analyzeRepeats.pl. Based on their pausing indexes, transcripts expressed in K562 ($\text{TPM} > 5$, $n = 18{,}107$) were classified as lowly ($< 30\text{th}$ percentile), intermediately ($> 30\text{th}$ percentile but $< 85\text{th}$ percentile), and highly paused (>85th percentile). For transcripts in each pausing group, we calculated the log-transformed read counts in the promoter regions from the PRO-seq and POLR2A ChIP-exo libraries. We reported the Pearson correlation coefficient between the two types of read counts.

Potential obstacles prohibiting the usage of ChIP-seq callers for TRE identification

MACS2 was developed specifically for analyzing ChIP-seq data[6]. Because ChIP-seq libraries are generally sequenced from the 5′ ends, at a bona fide TF-binding site, an enrichment of ChIP-seq tags is usually detected as a pair of peaks flanking the binding site, one upstream peak on the positive strand and one downstream peak on the negative strand[7]. Under its default configuration, MACS2 is designed to take advantage of this pattern: upon identification of a peak pair, it models the shifting size d based on the tag distribution in the pair, shifts the pair of peaks inward by $d/2$, and uses the shifted tag distribution to define the binding site. Therefore, MACS2 achieves satisfactory performance when used for analyzing ChIP-seq data. However, drastically different from ChIP-seq tags, sequencing reads of eRNAs originated from active enhancers are found as pairs of divergent peaks: one upstream peak on the negative strand and one downstream peak on the positive strand. Thus when using MACS2 under its default configuration to analyze the TSS-assays for eRNA detection, the algorithm is likely to pair wrong peaks for calling bidirectional elements.

When options --nomodal and --extsize 1 are used for analyzing TSS-assays, MACS2 will only call individual sequencing read peaks instead of peak pairs without any peak shifting. As a result, MACS2 will no longer identify eRNAs based on mismatched peak pairs. However, the peak-calling output of MACS2 under such settings lacks strand information which is indispensable for eRNA detection with TSS-assays. Another problem lies in the inflated signal-to-noise ratio, resulting from using the two options in combination, and this makes MACS2 prone to identify transcriptional noises as significant peaks (Extended Data Fig. 6b). Such misidentification is mainly due to a significantly lower estimation of read density in the local environment by MACS2 under such settings.

We captured an example of such misidentification near the TSS of *GATA1*, where a *GATA1* promoter (GENCODE[8], highlighted in red) and a putative enhancer[2,9] (highlighted in orange) have been identified (Extended Data Fig. 6b). When used to analyze the GRO-cap reads aligned to this region, MACS2 failed to distinguish the promoter and the enhancer even without implementing the shifting model. Such failure resulted from the incorrect identification and merging of two separate peaks as overlapping peaks when only a fraction of either peak happens to fall within the same predetermined genomic bin.

Inference of full-length eRNA transcript units from NT-libraries

To infer the full-length eRNAs from libraries generated by NT-assays, we hired findPeaks (-style groseq) from HOMER[10], genoSTAN[11], and groHMM[12] to predict transcript units. The quality of these predicted transcription units was evaluated by calculating the overlap between predicted TUs (denoted as pred.) and annotated TUs (denoted as annot.) from GENCODE in three aspects:

1) canonical Jaccard index:

$$\frac{|\text{Pred.} \cap \text{Annot.}|}{|\text{Pred.} \cup \text{Annot.}|}$$

2) the fraction of predicted TUs that overlaps with existing annotations:

$$\frac{|\text{Pred.}|}{|\text{Pred.} \cup \text{Annot.}|}$$

3) the fraction of annotations that overlaps with predictions:

$$\frac{|\text{Annot.}|}{|\text{Pred.} \cup \text{Annot.}|}$$

The current design of TSS-assays is not suitable for rigorously deciphering the full length of all eRNAs. NT-assays can theoretically be used to impute full-length eRNAs with the help of computational tools. However, when we benchmarked several previously used tools for predicting transcription units (TUs)[10–12] with the method mentioned above, we found an apparent disagreement between the predicted TUs and the curated transcript annotations from GENCODE (Extended Data Fig. 5a), suggesting that new computational tools might be needed.

Characterization of unique peaks identified by different tools

Profiles of histone marks for unique peaks were generated with a similar process as mentioned in *Evaluating the systematic biases of different peak calling methods*. To calculate the enrichment of motifs for enhancer-related and cell-type-specific transcription factors, we extended all peaks by 200 bp to ensure the window was large enough. PINTS unique peaks (with epigenomic annotation enabled) were used as the primary regions, and unique peaks from the other tool were used as the control regions. The occurrence of motifs in both the primary and control regions was scanned by AME[13] with motif information from JASPAR 2020[14]. The occurrences were converted to log odds ratio, and the z test was performed for calculating the p-values.

Supplementary References

1. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters.

Science **322**, 1851–1854 (2008).

2. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
3. Shivram, H. & Iyer, V. R. Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies. *RNA* **24**, 1266–1274 (2018).
4. Tome, J. M., Tippens, N. D. & Lis, J. T. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.* **50**, 1533–1541 (2018).
5. Mchaourab, Z. F., Perreault, A. A. & Venters, B. J. ChIP-seq and ChIP-exo profiling of Pol II, H2A.Z, and H3K4me3 in human K562 cells. *Scientific Data* vol. 5 (2018).
6. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
7. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* vol. 10 669–680 (2009).
8. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
9. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* vol. 51 1664–1669 (2019).
10. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390–394 (2011).
11. Zacher, B. *et al.* Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. *PLoS One* **12**, e0169249 (2017).
12. Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* **16**, 222 (2015).
13. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* vol. 11 (2010).

14. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).