Systematic comparison of experimental assays and analytical pipelines for identification of active enhancers genome-wide

Li Yao[1,2], Jin Liang[2], Abdullah Ozer[3], Alden King-Yung Leung[1,2], ENCODE Consortium, John T. Lis[1,3*], and Haiyuan Yu[1,2*].

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853.

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853.

[3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853.

Correspondence to: jonhlis@cornell.edu (J.T.L) and haiyuan.yu@cornell.edu (H.Y.).

Regulation of transcription is a synergetic process that requires both trans-regulatory factors, like transcription factors, and cis-regulatory elements, like promoters and enhancers. In contrast to promoters, which initiate transcription in their proximal regions to produce stable RNA products, enhancers regulate transcription of their target gene(s) in a distal manner. Certain epigenomic signatures (enrichment of H3K4me1 and H3K27ac, high chromatin accessibility, and CBP/p300 binding) are considered to be defining features of active enhancer loci[3,4]. However, studies also revealed that enhancers could themselves produce relatively short-lived divergent transcripts, called enhancer RNAs (eRNAs)[5,6]. More recent studies further showed that distal divergent transcription events are more reliable marks for active enhancers than epigenomic signatures[1,2]. Recently we have proposed[7,8] and later experimentally verified[2] the basic unit of active enhancers that are defined by the transcription start sites (TSSs) of the divergent eRNA transcription, and delimited by, the promoter-proximal Pol II pause sites flanking these TSSs. Therefore, to identify active enhancers genome-wide in any given sample, it is critical to detect eRNAs and their TSSs with high sensitivity and specificity.

eRNAs are usually in extremely low abundance in cells due to their short half-lives. Therefore, conventional RNA-seq experiments capture eRNAs with very low efficiency overall[5]. Recently, two categories of genome-wide RNA sequencing assays have been developed, focusing either on TSSs or on the actively-transcribing polymerase positions (Fig. 1a). We named the 8 assays (GRO[9]/PRO-cap[10], CoPRO[8], Start-seq[11], CAGE[12], RAMPAGE[13], NET-CAGE[14], csRNA-seq[15], and STRIPE-seq[16]) from the former category as 5′ assays, because these assays enrich for active 5′ TSSs of promoters and enhancers (Fig. 1a). We also named the 5 assays (GRO-seq[17], PRO-seq[10], mNET-seq[18], Bru-seq[19], and BruUV-seq[20]) from the latter category as 3′ assays, because

comparison of their performance with datasets generated by the aforementioned experimental assays.

In this study, we systematically examined 13 experimental assays in terms of their sensitivity and specificity for capturing eRNAs. We also developed a novel computational tool, Peak Identifier for Nascent-Transcript Sequencing (PINTS), which is designed to identify enhancer candidates from all of these assays. Moreover, by comparing PINTS with 8 other widely-used computational tools, we found that PINTS gave the highest overall performance pertaining to robustness, sensitivity, and specificity, especially when analyzing data from 5′ RNA sequencing assays. Finally, we constructed a comprehensive enhancer candidate compendium for 120 cells and tissues using the robust and unified definition of active enhancers based on detected eRNA TSSs genome-wide[1,2,7], and developed an online web server (https://pints.yulab.org/) to navigate, prioritize, and analyze enhancers based on a wide range of genomic and epigenomic annotations. We expect our enhancer compendium will be a valuable resource to the research community for effective selection of candidate enhancers for further functional characterization in future studies.

Results

5′ assays enriching short and/or capped RNAs demonstrate higher sensitivity in eRNA detection with GRO-cap being the most sensitive assay

To perform a quantitative comparison of eRNA detection sensitivity, we first normalized all libraries by down-sampling them to the same sequencing depth as the library with the lowest depth (18.9 million mappable reads, Supplementary Table 1). We then conducted a pairwise

ability to enrich unstable transcripts, which is of particular importance for detecting eRNAs (Fig. 1d and Supplementary Fig. 1b). We evaluated the effects of technical artefacts, including strand specificity and internal priming, and our results suggest all libraries have great strand specificity (average: 0.9797, SD: 0.0209) and low internal priming rates (Supplementary Notes).

Sequencing reads from gene bodies and small RNAs in 3′ assays contribute to lower sensitivity in eRNA detection

For the two families of assays that we compared in this study, we noticed that in general, 5′ assays are more sensitive in detecting eRNAs than 3′ assays, even for assays that use very similar enrichment strategies (Fig. 1b). For instance, while both GRO-cap and PRO-seq employ similar nuclear run-on procedures, there is a 41.22% difference between their divergent coverage of the true enhancer set (Fig. 1c). When inspecting genome-wide distribution of reads (Fig. 2a and Supplementary Fig. 2a), we noticed that 3′ assays have significantly higher proportions of reads coming from gene body regions (mean of 3′ assays: 65.62%, mean of 5′ assays: 13.01%, *p*-value from Mann-Whitney *U* test: 0.0029), which is not surprising as they are designed to reveal all actively transcribing RNA polymerases, whereas 5′ assays are specifically designed for identification of TSSs. Because eRNA transcription is on average much lower than that of genes[7], such a high portion of gene body reads in 3′ assays dilute the signal from eRNAs and significantly lower their sensitivity in detecting active enhancers. As shown in Fig. 2b, 3′ assays detect a significant number of reads in the FAM89A gene body, whereas 5′ assays only have reads in the promoter regions of the FAM89A gene. As a result, almost all 3′ assays (except PRO-seq) have no discernable signal at a distal enhancer locus near the FAM89A gene that was validated by CRISPRi[35](Fig. 2b). Another potential problem for 3′ assays is that reads that are

5′ assays have better specificity in detecting active enhancers, with GRO-cap having the best performance

While genomic regions with detectable transcriptional events account for 75% of the human genome[43], many of these events are considered to be spurious transcriptional noise[44,45] because of their extremely low transcript yields compared to mRNAs and of the intrinsic promiscuity of RNA Pol II under certain circumstances[46]. Therefore, it is critical to detect and differentiate the reads that originated from spurious transcription in these assays. To that end, we collected non-enhancer loci from eight Massively Parallel Reporter Assay (MPRA)[47–51] and Self-Transcribing Active Regulatory Region Sequencing (STARR-seq)[2,52,53] studies and further removed elements overlapping with predicted Enhancer-Like Sequence (ELS) or Promoter-Like Sequence (PLS) from candidate cis-regulatory elements (cCRE) annotations[54] to generate a set of 7,097 loci (referred to as the "non-enhancer set", Supplementary Fig. 2f, Supplementary Table 3). We observed that signal intensities in the true enhancer set are often higher than that in the non-enhancer set (Fig. 2d and Supplementary Fig. 2g), with GRO-cap having the highest signal-to-noise ratio (64-fold enrichment, Fig. 2d and Supplementary Fig. 2g).

We also calculated false discovery rate (FDR) for each assay based on the overlap of their reads with both the true enhancer and non-enhancer sets. We found that 5′ assays generally have lower FDRs than those of 3′ assays (5′ assay mean: 0.2698 *vs.* 3′ assay mean: 0.5253, *p*-value from Mann-Whitney *U* test: 5.089e-5, Fig. 2e and Supplementary Fig. 2h), with GRO-cap having the lowest FDR (mean: 0.1250, SD: 0.0046).

background, resulting in diminished statistical significance for all TSSs in that locus and leading to false negatives in TSS detection. To address these issues, we developed Peak Identifier for Nascent-Transcript Sequencing (PINTS), which uses zero-inflated Poisson models to evaluate local read densities and employs interquartile range (IQR)-based refinement to ameliorate false negatives by conditionally masking candidate TSSs in the local background (Fig. 3). PINTS was inspired by MACS2[57] with modifications specifically implemented for identification of eRNA TSSs from genome-wide RNA sequencing assays (especially 5′ assays). After evaluating the significance of each TSS, PINTS defines TREs as divergent TSS pairs that are within 300bp from each other, as suggested by previous studies[2] (Supplementary Fig. 3a, Methods). We identify candidate enhancers as the distal TREs that are farther than 500 bps away from known protein-coding gene TSSs[2].

Most peak callers can identify active enhancers from high-throughput datasets, but their resolution and computational requirements vary significantly

Candidate enhancer loci identified by peak caller algorithms should share the same features as true enhancers with characteristic epigenomic marks (DHS, H3K27ac, H3K4me3, and H3K4me1) and transcription factor (CBP/p300, GATA1, and CTCF)-binding sites. Indeed, we found that candidate enhancers identified by most tools recapitulate these features of true enhancers (Fig. 4a, b and Supplementary Fig. 3a for GRO-cap and Supplementary Fig. 3b~l for all other assays). However, the distribution pattern of epigenomic marks and TF-binding sites of candidate enhancers identified by MACS2[57], a widely used peak caller for analyzing ChIP-seq data, is remarkably distinct from those of the true enhancers, suggesting the default shifting

hg38. Because there is no variation in the RNA sequencing datasets themselves and the two reference genome builds are very similar (hg38 has 0.09% more ungapped non-centromeric sequences than hg19, only 0.17% of ungapped hg19 sequences are not in hg38[58], Supplementary Fig. 4a), we expect that the differences in peak calls using these two different genome builds should be minimal across all datasets. Surprisingly, we found that by simply changing the reference genome builds, half of them have Jaccard index smaller than 0.9 for at least one assay (Fig. 4f). Furthermore, we noticed that Jaccard indices were even lower when we tried to evaluate robustness across real technical and biological replicates (average: 0.5068, SD: 0.2462), especially for TSScall and FivePrime, where their robustness is only 0.4013 (SD: 0.1038) and 0.3836 (SD: 0.2028), respectively (Supplementary Fig. 4b). PINTS consistently have great robustness in both cases (average: 0.9761, SD: 0.0081 between hg19 and hg38 genome builds, and average: 0.7279, SD: 0.0515 across replicates).

To evaluate sensitivity and specificity, two other key metrics of performance, we merged the true enhancer set with the promoter regions from GENCODE v24[40] as the positive set, and non-enhancer loci as the negative set (Method). As shown in Supplementary Fig. 4c, the negative set has distinct patterns of TF-binding motifs and chromatin marks compared to either the true enhancer or the promoter set. We then evaluated each tool's performances for all 5′ assay datasets (Fig. 4g). The results show that PINTS achieves the best balance between sensitivity and specificity (PINTS mean AUC: 0.7796, SD: 0.0821; mean AUC for the second-best tool dREG: 0.6518, SD: 0.1088). For all of these computational tools, we summarize their key requirements, main characteristics, applicability to different RNA sequencing assays in Supplementary Fig. 5.

architecture characteristics. All of these candidate enhancer calls are made publicly available through our web server (https://pints.yulab.org) described in detail below. We will regularly update our enhancer compendium as new datasets, especially those in new cell lines or samples, and assays, become available.

In human K562 cells where datasets are available from all 5′ assays, our results show that GRO-cap has by far the most number of distal TRE elements (19,006 identified by PINTS with 9,531 unique enhancer calls – not identified by any other assay; the second-best dataset, csRNA-seq, only has 14,375 enhancer calls with 5,048 unique (Fig. 5b). This is not surprising given that GRO-cap has the best sensitivity in detecting eRNA transcription (Fig. 1c and 1d) and the GRO-cap dataset has the second highest read depth (Fig. 5b). We selected three CRISPRi-validated enhancer-promoter pairs[35] to visualize these differences and showed the variety in signal abundances between all assay datasets (Fig. 5c-e). For example, the enhancer which regulates the JUND gene (Fig. 5c) has decent accessibility and is supported by epigenomic marks, including H3K27ac and H3K4me1. As expected, all four 5′ assays can identify this enhancer. The expression levels of enhancers are not necessarily to be proportional to the levels of epigenomic marks, and for eRNAs whose expression levels are lower (e.g., the enhancer that regulates FTH1 in Fig. 5d), assays that are more effective in capturing unstable transcripts are more likely to recover them. Finally, for the enhancer regulating TMA16, signals from histone marks are quite minimal, but GRO-cap still captures clear signals of eRNA transcription at this locus and readily enables identification of this enhancer (Fig. 5e).

these samples, we automatically refine our annotations by only reporting binding sites of expressed transcription factors, and associating each enhancer with epigenomic features specific to the corresponding sample.

Users can explore all annotations of their selected enhancers via our integrated genome browser; alternatively, they can easily export all annotations to a local machine in plain text format, which greatly facilitates any user-designed downstream analyses.

Discussion

eRNAs are increasingly being recognized as a critical marker for active enhancers genome-wide[1,2]; however, the optimal strategy (both experimental assays and their analytical pipelines) to detect eRNAs and thus identify enhancer loci has not been unveiled. In this study, we systematically compared 13 *in vivo* genome-wide RNA sequencing assays in K562 cells and showed that 5′ assays are in general more sensitive than 3′ assays to detect eRNAs, because signals will not be diluted by active transcription in gene bodies. One additional, and critical advantage of the 5′ assays is that they reveal the precise location of eRNA TSSs, allowing for high-resolution detection and dissection of enhancer loci genome-wide as demonstrated in our recent work[2]. Overall, our results show that GRO/PRO-cap has the best overall performance in detecting active enhancers in terms of both sensitivity and specificity.

We noticed that when using current computational tools to identify TREs from various RNA sequencing datasets, very minor changes in sample processing could lead to more than 20% of changes in the final results, which brings the robustness of the peak calls into question. To

provide basic guidelines in selecting the proper experimental assays and the correct computational tools for future studies.

Furthermore, we provide a detailed, comprehensive human enhancer compendium for 33 cell lines, 7 *in vitro* differentiated cells, 35 primary cells, and 45 tissues (120 in total). We used a unified definition[2,7] of enhancers based on the detected divergent pairs of eRNA TSSs (i.e., peak calls from various genome-wide RNA sequencing assays, especially 5′ assays). Such a robust, unified and comprehensive catalogue of enhancers across 120 cells and tissues is expected to shine light on the mechanism of gene regulation and architectural details of enhancers in general. Precise definition of enhancer element boundaries afforded by 5′ assays like PRO/GRO-cap would alleviate potential concerns regarding whether full-length enhancer elements were selected and tested in follow-up functional studies, and thus improve coverage of elements by eliminating incomplete or ill-defined candidates. Such a well-defined catalogue of enhancers also provides an invaluable resource for follow-up studies to better understand the similarities and key differences in gene regulation across various tissues and conditions, and to identify key enhancers whose malfunctions can lead to specific disorders.

Classifying the transcription units as stable and unstable units with TT-seq

Transcript annotations derived from TT-seq (GSE75792[39]) were downloaded from the GEO database. Transcription units with NA values were discarded. The 95th quantile of estimated decay rates for mRNAs was used as the cutoff between the unstable (above the cutoff) and stable (below the cutoff) transcription units.

Characterizing the genome-wide distribution of reads

The entire genome was classified into four categories based on the annotations in GENCODE (ver 24)[40]: exonic and intronic regions were defined as in GENCODE, except that any region with overlapping intronic and exonic annotation was considered as exonic; the 500bp regions flanking annotated transcription start sites of protein-coding transcripts were annotated as promoters; all other regions were considered as intergenic. Sequencing reads of various assays were assigned to the categories of promoters, introns, exons, or intergenic regions (in the exact order) if they were aligned to the corresponding annotated regions in the genome.

Identifying sequencing reads from splicing intermediates

The exact or approximate positions of transcript termini were inferred from the read ends and the abundance of their corresponding transcripts was normalized as RPM for this analysis. A list of annotated splice junctions and their 200bp flanking regions in the human genome was compiled based on GENCODE v24[40]. For each assay, we iterated through this list and recorded normalized read counts at each position. In Supplementary Fig. 2e, both the average of signals and the 95% confidence interval (estimated by bootstrap) of the averages were reported.

scoring pairing strategy[28], the nearby seeds will be merged together as peak candidates if the density after merging meets the following condition:

$$D_{merged} \geq \alpha \times min(\{D_{seed1}, D_{seed2}\})$$

The default value for α is 1, but PINTS's resolution can be further fine-tuned by incorporating reference annotations. For example, when the transcript annotation is available, PINTS will try to avoid the overlap of peak candidates with more than one transcript.

Next, to address the significantly increased sparsity of signals when only the read ends are taken into account, the Expectation-Maximization algorithm is used to fit zero-inflated Poisson (ZIP) models to both peak candidates and their neighborhood regions (λ for read density, π for the proportion of zeros that are not coming from a Poisson process), the probability mass function of these models has the following form:

$$\Pr(X = x) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, & x = 0 \\ (1-\pi)e^{-\lambda}\dfrac{\lambda^x}{x!}, & x > 0 \end{cases}$$

Assume an unobservable latent random variable z_i, for a window X of I observations, the complete log-likelihood is proportional to:

$$\ln L \propto \sum_{i=1}^{I} [z_i \ln(\pi) + (1-z_i)\ln(1-\pi) + (1-z_i)(-\lambda + x_i \ln \lambda)]$$

In E-step at the $(r+1)$th iteration, z_i is estimated by its conditional expectation:

$$\hat{z}_i^{(r+1)} = \begin{cases} \dfrac{\hat{\pi}^{(r)}}{\hat{\pi}^{(r)} + e^{-\hat{\lambda}^{(r)}}(1-\hat{\pi}^{(r)})}, & x_i = 0 \\ 0, & x_i > 0 \end{cases}$$

In M-step, given $\hat{Z}_i^{(r+1)}$ the estimations of π and λ are updated as follows:

machines with Intel Xeon Gold 6152 CPU @ 2.10 GHz with 88 cores, 1006 GB of RAM running CentOS 7.6.1810.

Evaluating the systematic biases of different peak calling methods

For each assay, divergent elements were identified using all applicable peak callers, including PINTS. To accommodate the size difference in these elements as well as elements in the true enhancer set, a 1,000-bp region centered around the midpoint of each element was used to evaluate the performance of different methods.

Evaluating the upper bound of peak caller robustness

Sequencing reads were aligned to another popular reference genome sequence, hg19, and divergent elements were identified accordingly with different peak callers. Peak calls generated from both genome releases were cross lifted using UCSC's liftover and the average between the two Jaccard indices was considered as the upper bound robustness (UBR):

$$\mathrm{UBR} = \frac{1}{2} \times \left(\frac{|Peaks_{38} \cap Peak_{37\to38}|}{|Peaks_{38} \cup Peak_{37\to38}|} + \frac{|Peaks_{37} \cap Peak_{38\to37}|}{|Peaks_{37} \cup Peak_{38\to37}|}\right)$$

ROC

For each assay, any element from the true and non-enhancer set was filtered out if there were no sequencing reads aligned to both strands of the element. The positive set was composed of an equal number of randomly sampled promoters (1kb regions flanking TSSs in GENCODE v24) from expressed genes and the filtered true enhancers. The negative set was composed as the filtered non-enhancers. ROCs were generated by calculating the number of divergent elements overlapping with the positive and negative sets under different cutoffs of scores: *p*-values of

Author contributions: Conceptualization: L.Y., J.T.L., and H.Y.; Methodology: L.Y.; Software: L.Y.; Formal analysis: L.Y.; Investigation: J.L.; Data curation: L.Y., J.L., and A.K.Y.L.; Writing – Original Draft: L.Y., and J. L.; Writing – Review & Editing: J.L., A.O., J.T.L., and H.Y.; Visualization: L.Y., J.L., A.O., and H.Y.; Supervision: J.T.L., and H.Y.

Competing interests: Authors declare no competing interests.

Data and materials availability: The source code of PINTS is publicly available at https://github.com/hyulab/PINTS, scripts used to generate results that are reported in this study can be retrieved from https://github.com/hyulab/PINTS_analysis. Raw data for all figures in this study is freely available at Zenodo (10.5281/zenodo.4527567).

Supplementary Materials

Figures S1-S5

Tables S1-S4

References (S1-S4)

References

1. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
2. Tippens, N. D. *et al.* Transcription imparts architecture, function and logic to enhancer units. *Nat. Genet.* **52**, 1067–1075 (2020).
3. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
4. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
5. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
6. Descostes, N. *et al.* Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. *Elife* **3**, e02105 (2014).
7. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
8. 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).
9. Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T. & Meyer, B. J. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *Elife* **2**, e00808 (2013).
10. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
11. Henriques, T. *et al.* Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev.* **32**, 26–41 (2018).
12. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).

transcription using dREG. *Genome Res.* **29**, 293–303 (2019).

26. Chu, T. *et al.* Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat. Genet.* **50**, 1553–1564 (2018).
27. Adiconis, X. *et al.* Comprehensive comparative analysis of 5'-end RNA-sequencing methods. *Nat. Methods* **15**, 505–511 (2018).
28. Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).
29. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
30. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
31. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
32. Wakabayashi, A. *et al.* Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4434–4439 (2016).
33. Klann, T. S. *et al.* CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
34. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* **66**, 285–299.e5 (2017).
35. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390.e19 (2019).
36. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
37. Xie, S., Armendariz, D., Zhou, P., Duan, J. & Hon, G. C. Global Analysis of Enhancer Targets Reveals Convergent Enhancer-Driven Regulatory Modules. *Cell Rep.* **29**, 2570–2578.e5 (2019).

doi:10.1038/nbt.4285.

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

53. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).

54. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

55. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

56. 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).

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

58. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

59. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

60. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).

61. Vo Ngoc, L., Huang, C. Y., Cassidy, C. J., Medrano, C. & Kadonaga, J. T. Identification of the human DPR core promoter element using machine learning. *Nature* **585**, 459–463 (2020).

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

63. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).

64. Yao, L., Wang, H., Song, Y. & Sui, G. BioQueue: a novel pipeline framework to accelerate

Figure Legends

Fig. 1. Comparison of currently available assays for detecting eRNAs. a, Schematics of enhancer and promoter/gene transcription by RNA Pol II (left panel) and characteristic profiles of 5′ and 3′ assays (right panel, light-blue shaded area). Black lines represent genomic DNA; nascent RNAs are purple curved lines with 5′ and 3′-end colored blue and red, respectively, with light-blue spheres as caps and yellow ovals indicate RNA Pol II. Arrows indicate the direction of sequencing reads, 5′ assay in blue, and 3′ assay in red. Representative read density profiles are colored accordingly in blue and red for 5′ and 3′ assays, respectively. TSS: transcription start site; CPS: cleavage polyadenylation site; TTS: transcription termination site. **b**, Enrichment strategies used by different 5′ and 3′ RNA assays. TEX: terminator exonuclease. * indicates that 3′ RNA ends can only be estimated approximately by these assays. A detailed description is available in Supplementary Notes. **c**, The capability of different assays to capture validated enhancers (true enhancer set). All libraries were downsampled to the same sequencing depth. "Unidirectional" and "divergent" indicate the detection of eRNAs originated from either one or both strands of the enhancer loci, respectively. **d**, Differences in read coverage among stable and unstable transcripts. GRO-cap has the highest coverage of both stable and unstable transcripts, and the least preference toward stable transcripts. Read counts were $\log(n+1)$ transformed; preferences (effect sizes) were evaluated as Cohen's *d*.

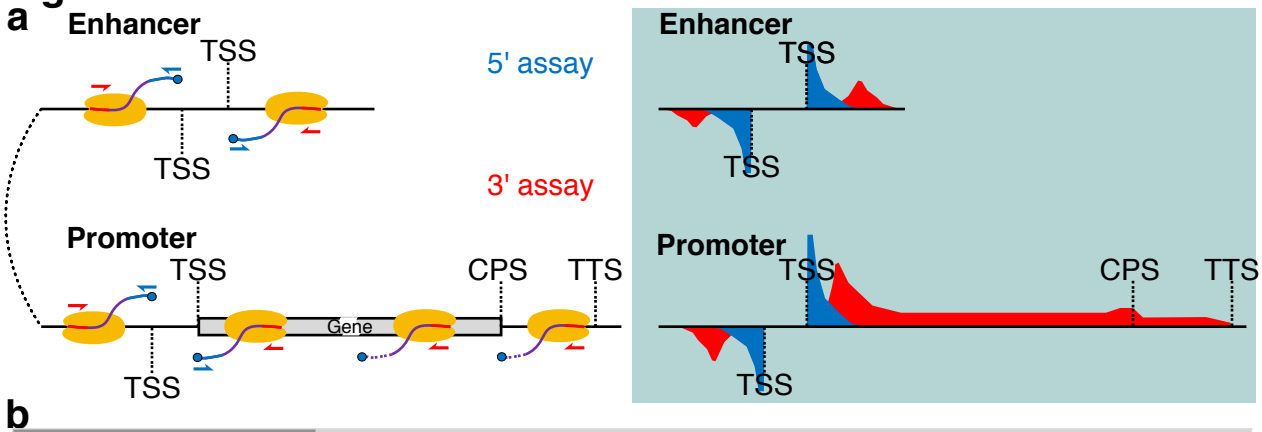
Fig. 2. Characterization of factors affecting assay sensitivity and evaluation of assay specificity in eRNA detection. **a**, Genome-wide distribution of sequencing reads originated from intergenic regions, introns, exons, and promoters detected by different assays. **b**, A genome browser snapshot of a gene (FAM89A) and its enhancer (highlighted in yellow), demonstrating

Fig. 4. PINTS achieves the best balance among resolution, robustness, sensitivity, specificity, and computational resources required. **a** and **b**, Profiles of GATA1 binding sites and H3K27ac pattern in true enhancer regions and distal TREs identified by different peak callers. **c**, Distribution of element sizes identified by different tools. **d**, CPU time consumed by peak callers to identify elements from various 5′ assay libraries. Average CPU time labeled inside each bar is in the unit of hours (N=6). The x-axis is in $\log(n+1)$ scale. **e**, Maximum memory usage during peak calling from 5′ assay libraries. Average maximum usage labeled inside or on top of each bar is in the unit of gigabytes (N=6). For **d** and **e**, error bars represent SD. **f**, Robustness of peak calls made by different tools. Libraries were mapped to both hg19 and hg38, and robustness was measured as the Jaccard index between calls from hg19 and hg38 (lifted over). Cells colored in gray indicates either that the tool cannot be applied to the corresponding assays or that one or more required datasets are not available. **g**, Aggregated ROC curves for each peak caller on all 5′ assay datasets. The solid lines represent the mean values; the corresponding shaded areas show the 95% confidence interval of the means (via bootstrap). For tools where ROCs cannot be calculated, solid dots represent their performance with default parameters, error bars show SD (see Methods).

Fig. 5. A comprehensive human enhancer compendium. **a**, Summaries of all distal elements identified from different assays with PINTS in 7 cell lines (top) and 131 datasets generated by different assays across 120 biosamples included in our enhancer compendium (bottom). Differentiated cells stand for *in vitro* differentiated cells. **b**, The number of distal elements identified by PINTS from different assays in K562. The dark colors indicate the proportion of shared elements identified from at least one other assay; light colors indicate elements unique to

Fig. 1

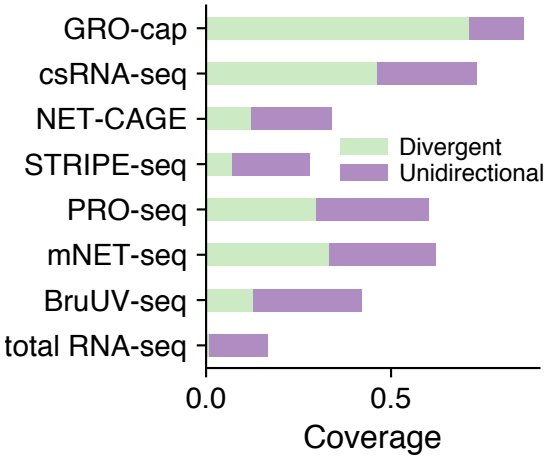
a



b

Assay		Enrichment strategies			
mNET-seq[18]	3'*	Nuclei/chromatin isolation	MNase		Size selection
PRO-seq[10] GRO-seq[17]	3'		Biotin-NTPs Or Br-UTP Nuclear run-on	Biotin BrU	
GRO[9]/PRO-cap[10] CoPRO[8]	5'				Cap selection
Start-seq[11]	5'		Size selection		
csRNA-seq[15]	5'				
NET-CAGE[14]	5'	α-amanitin	Cap trapper		
CAGE[12]	5'				
STRIPE-seq[16]	5'			TEX treatment	GGG CCC Template Switching
Bru-seq[19]	3'*	BrU PD UVC Live cell	BrU pulse-chase	BrU	
BruUV-seq[21]	3'*				

c



d

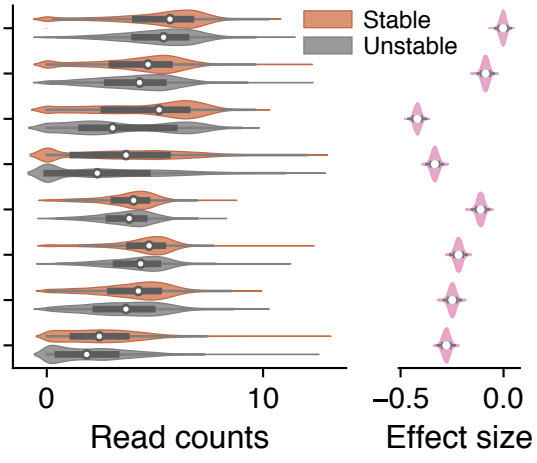


Fig. 2

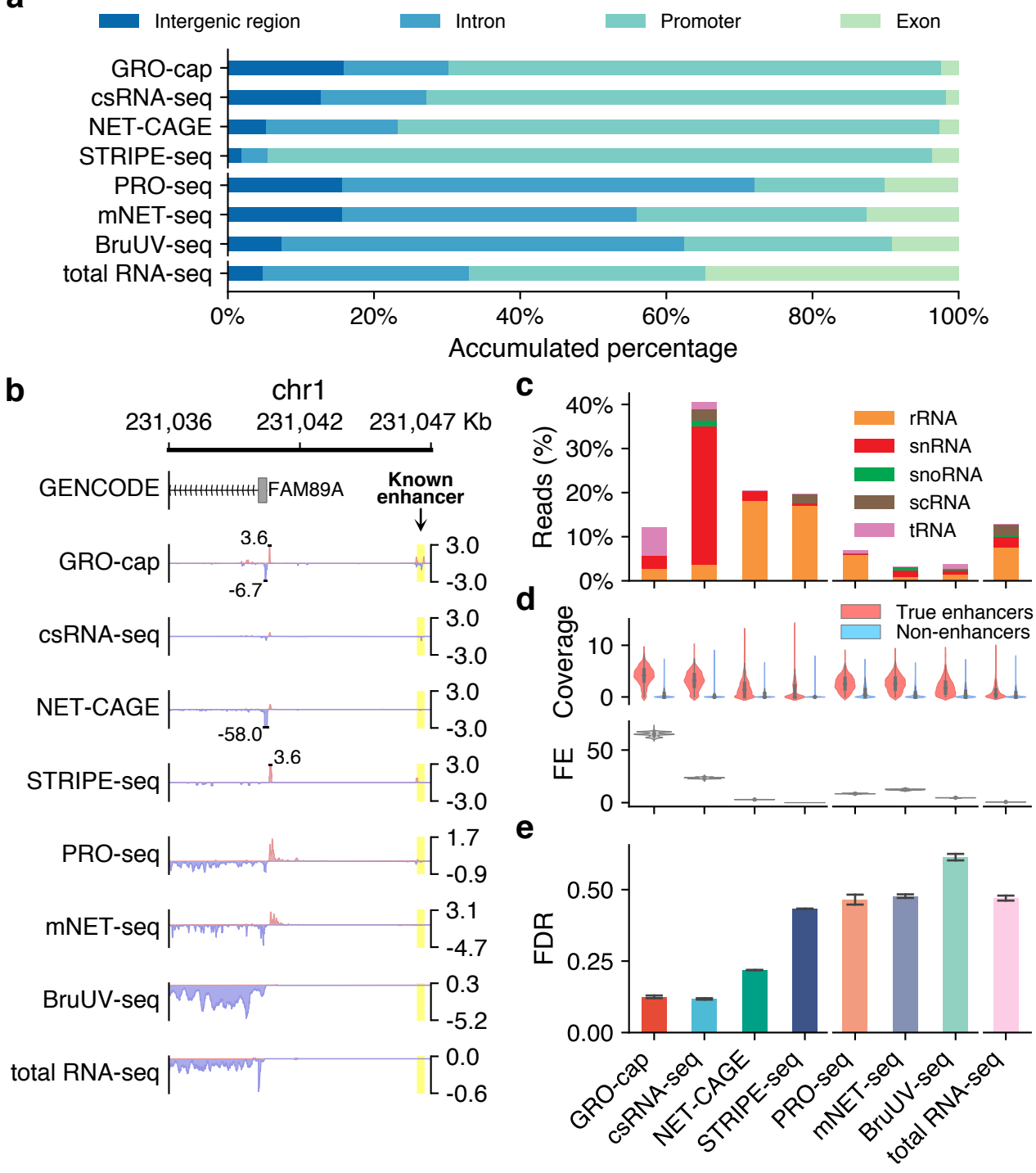


Fig. 3

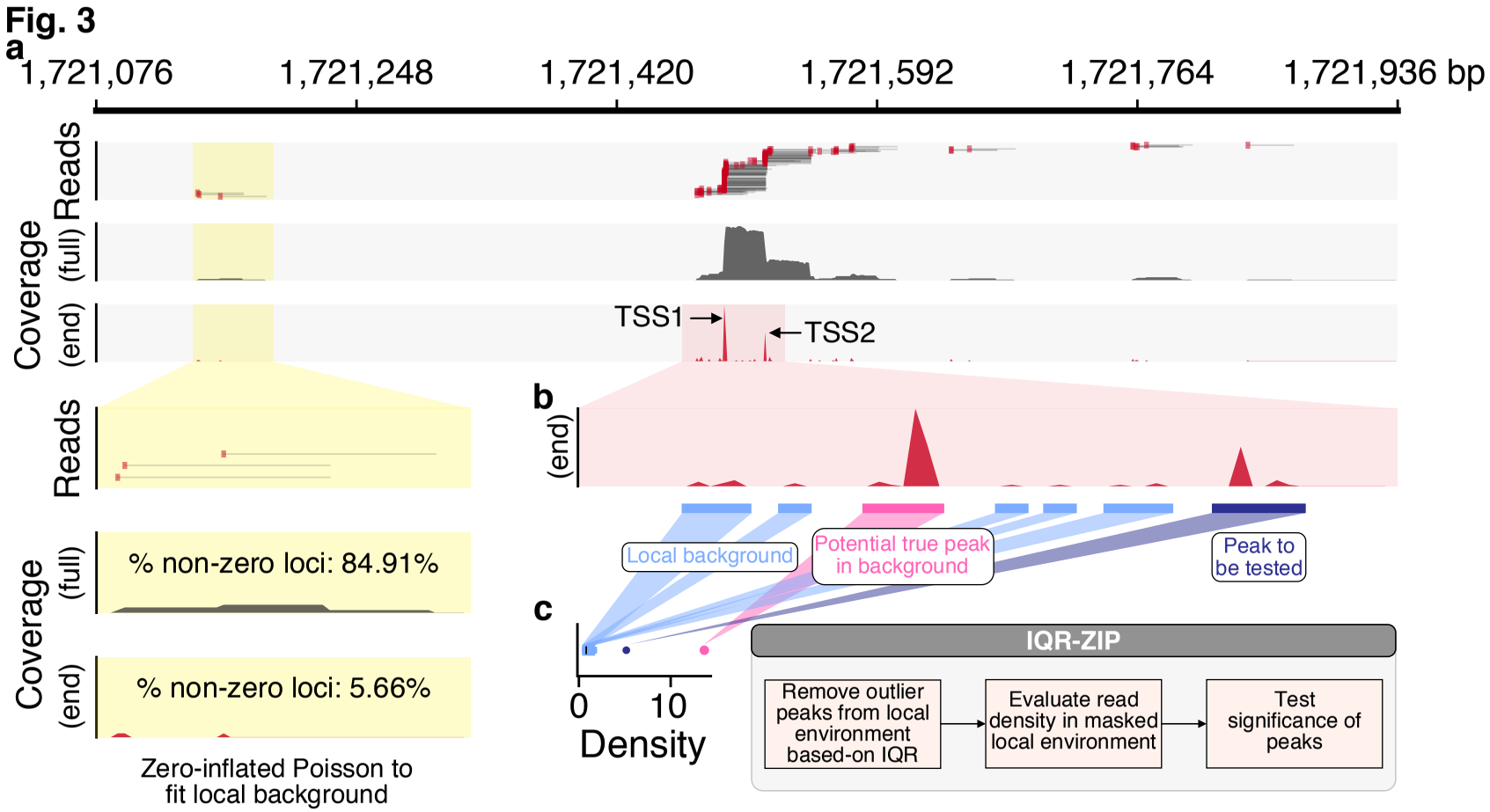


Fig. 4

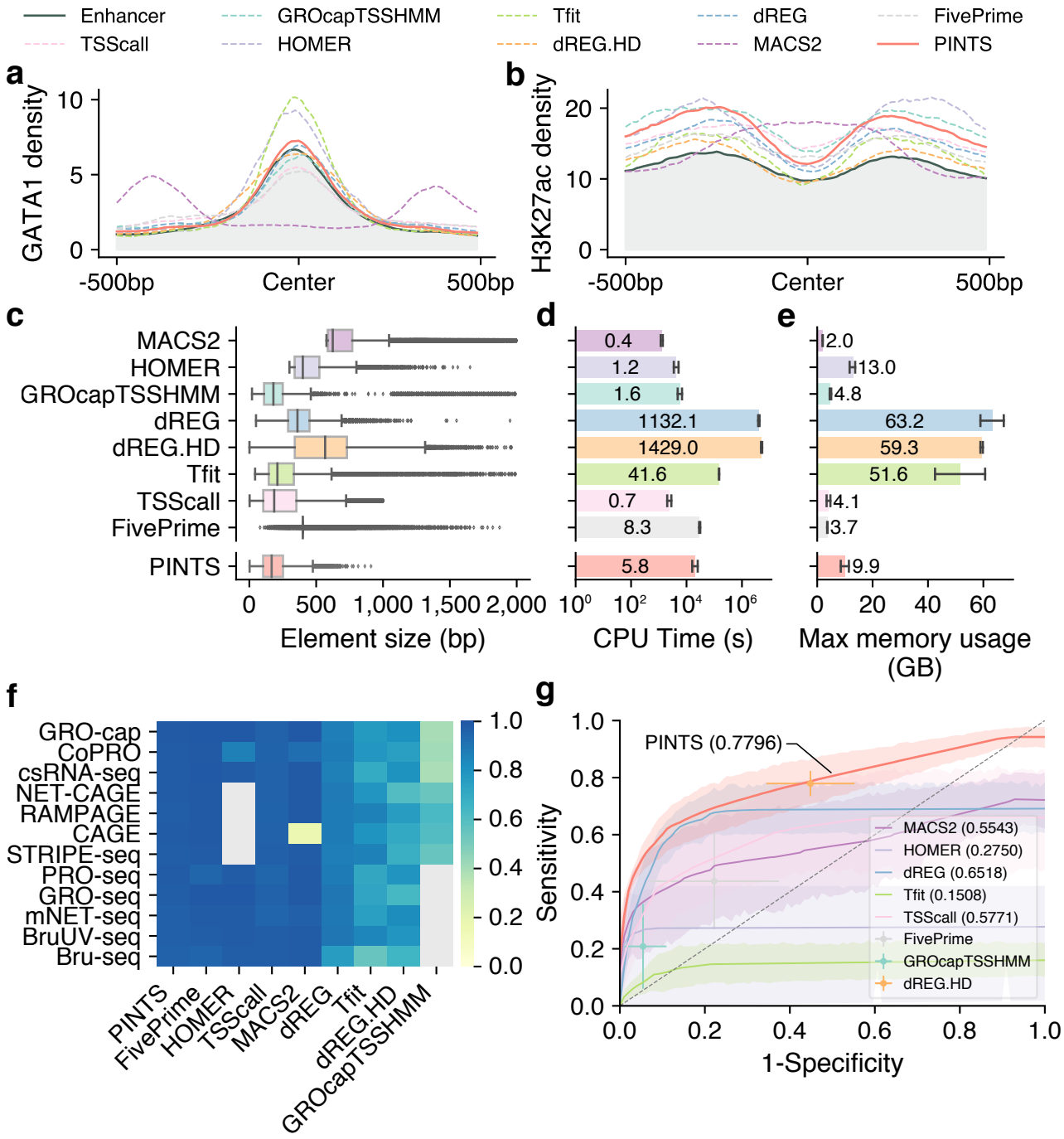


Fig. 5

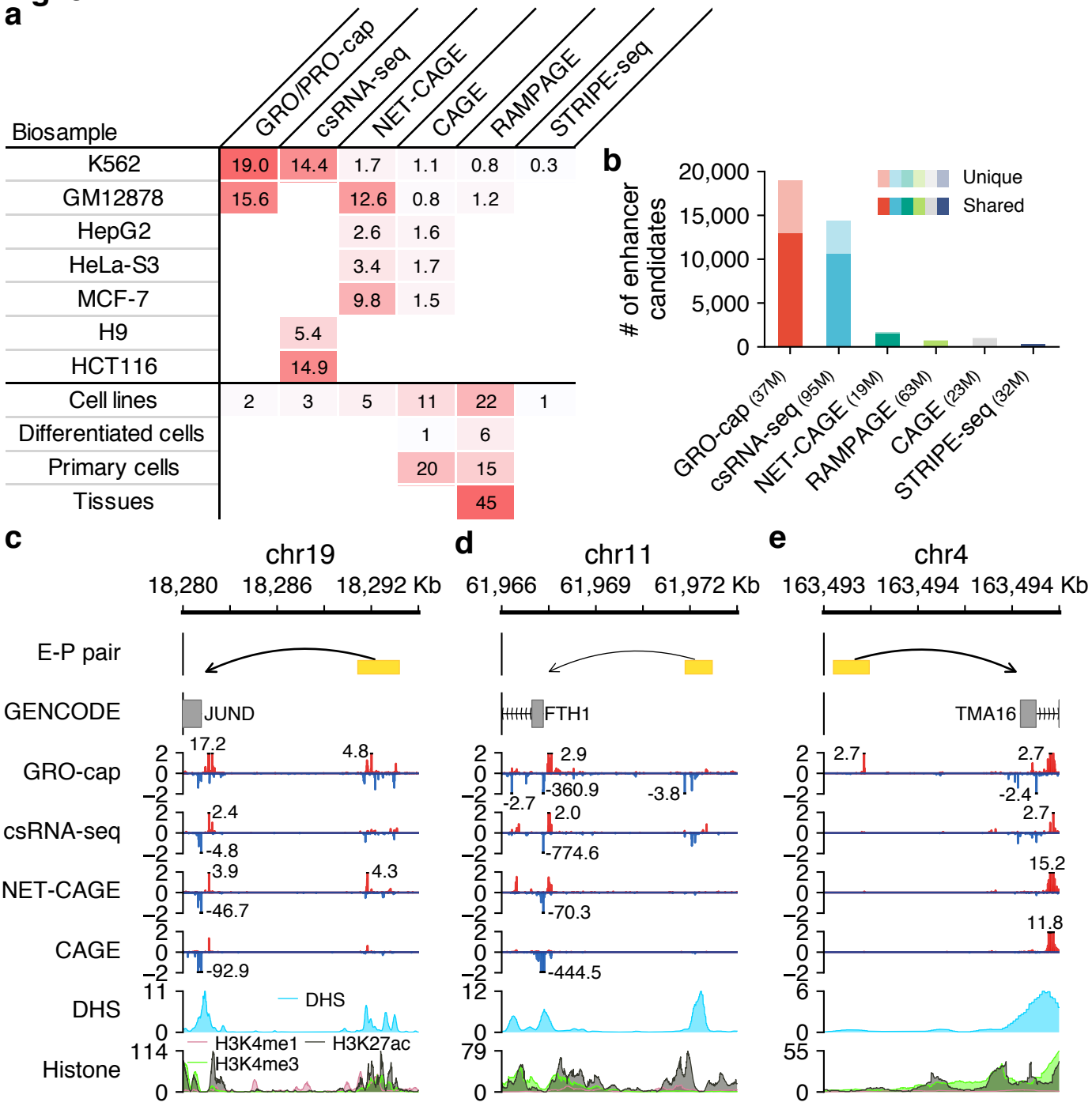


Fig. 6

1. Catalog database

TRE	Supporting assays	Supporting peak caller	Biosample	Coordinate
1	GRO-cap, csRNA-seq, NET-CAGE	PINTS, dREG, dREG.HD, HOMER	K562	chr6: 34223990-34224302
2	GRO-cap, csRNA-seq, NET-CAGE, CAGE	PINTS, dREG, dREG.HD, FivePrime	K562	chr19: 48812000-48812300
:				
N	GRO-cap	PINTS, HOMER	GM12878	chrX: 12977500-12977789

2. User interactive interface

User input genomic regions
-- OR --
Select assays and peak callers

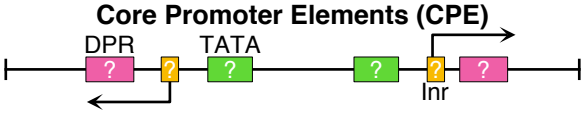
User upload peak calls generated from their data

3. Analytical engines

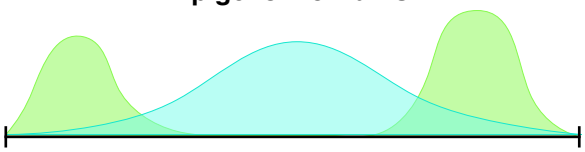
Distribution of Mutations

Transcription Factor Binding Sites

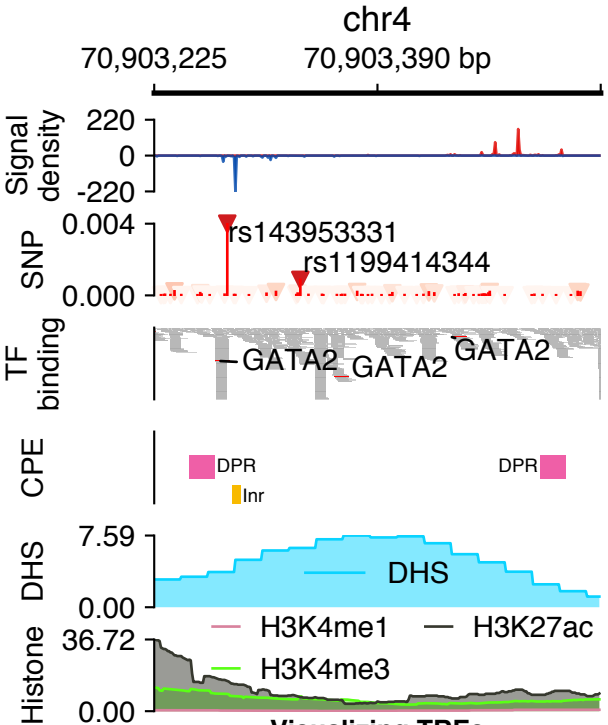
Core Promoter Elements (CPE)



Epigenomic Marks



Prioritizing TREs from multiple aspects



Visualizing TREs