

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

Li Yao<sup>1,2</sup>, Jin Liang<sup>2</sup>, Abdullah Ozer<sup>3</sup>, Alden King-Yung Leung<sup>1,2</sup>, ENCODE Consortium, John T. Lis<sup>1,3\*</sup>, and Haiyuan Yu<sup>1,2\*</sup>.

5 <sup>1</sup>Department of Computational Biology, Cornell University, Ithaca, NY 14853.

<sup>2</sup>Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853.

<sup>3</sup>Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853.

\*Correspondence to: [jonhlis@cornell.edu](mailto:jonhlis@cornell.edu) (J.T.L) and [haiyuan.yu@cornell.edu](mailto:haiyuan.yu@cornell.edu) (H.Y.).

## Abstract

Mounting evidence supports the idea that transcriptional patterns serve as more specific identifiers of active enhancers than histone marks<sup>1,2</sup>; however, the optimal strategy to identify active enhancers both experimentally and computationally has not been determined. In this study, we compared 13 genome-wide RNA sequencing assays in K562 cells and showed that the nuclear run-on followed by cap-selection assay (namely, GRO/PRO-cap) has significant advantages in eRNA detection and active enhancer identification. We also introduced a new analytical tool, Peak Identifier for Nascent-Transcript Sequencing (PINTS), to identify active promoters and enhancers genome-wide and pinpoint the precise location of the 5' transcription start sites (TSSs) within these regulatory elements. Finally, we compiled a comprehensive enhancer candidate compendium based on the detected eRNA TSSs available in 120 cell and tissue types. To facilitate the exploration and prioritization of these enhancer candidates, we also built a user-friendly web server (<https://pints.yulab.org>) for the compendium with various additional genomic and epigenomic annotations. With the knowledge of the best available assays and pipelines, this large-scale annotation of candidate enhancers will pave the road for selection and characterization of their functions in a time-, labor-, and cost-effective manner in the future.

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<sup>3,4</sup>. However, studies also revealed that enhancers could themselves produce relatively short-lived divergent transcripts, called enhancer RNAs (eRNAs)<sup>5,6</sup>. More recent studies further showed that distal divergent transcription events are more reliable marks for active enhancers than epigenomic signatures<sup>1,2</sup>. Recently we have proposed<sup>7,8</sup> and later experimentally verified<sup>2</sup> 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<sup>5</sup>. 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<sup>9</sup>/PRO-cap<sup>10</sup>, CoPRO<sup>8</sup>, Start-seq<sup>11</sup>, CAGE<sup>12</sup>, RAMPAGE<sup>13</sup>, NET-CAGE<sup>14</sup>, csRNA-seq<sup>15</sup>, and STRIPE-seq<sup>16</sup>) 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<sup>17</sup>, PRO-seq<sup>10</sup>, mNET-seq<sup>18</sup>, Bru-seq<sup>19</sup>, and BruUV-seq<sup>20</sup>) from the latter category as 3' assays, because

they are designed to identify 3' ends of the RNA transcripts (Fig. 1a). To enrich for RNA populations of interests, these assays implement various experimental strategies, including nuclei/chromatin isolation<sup>8-11,18</sup>, nuclear run-on<sup>8-10</sup> or metabolic labeling<sup>19,20</sup> with biotin or bromo-tagged nucleotides and affinity purification, Pol II immunoprecipitation<sup>18</sup>, size selection<sup>15,16</sup>, and enzymatic elimination of non-capped RNAs<sup>8-10,16</sup>, or chemical tagging of capped RNAs<sup>12-14</sup>. We summarized key experimental steps of both 5' and 3' assays in Fig. 1b. In fact, the list of all assays compared here, plus total RNA-seq<sup>21,22</sup>, have all been utilized in some capacity to identify enhancer elements. However, considering that most of these assays are not specifically designed to capture eRNAs, caution should be taken when exploiting the resulting data for identification of active enhancers or when deciding on the appropriate assay to use. Here, using datasets produced in the same ENCODE Tier 1 cell line, K562, we performed a comprehensive evaluation of the reliability and sensitivity in detecting eRNAs (thus identifying active enhancers genome-wide) across all available RNA sequencing assays, including seven 5' assays (no K562 Start-seq dataset is available), five 3' assays, and total RNA-seq (as the outgroup).

Because these assays were initially designed for different purposes, various computational tools were developed for exploring and interpreting the raw experimental data—for example, Tfit<sup>23</sup>, dREG<sup>24,25</sup>, and dREG.HD<sup>26</sup> were developed to identify transcriptional regulatory elements (TREs) from some 3' assays, including GRO-seq and PRO-seq; FivePrime<sup>27</sup> (based on paraclu<sup>28</sup>), GROcapTSSHMM<sup>7</sup>, and HOMER<sup>15</sup> were introduced for analyzing data from CAGE, GRO-cap, and csRNA-seq, respectively. While all these tools can potentially be used to identify eRNA transcription and active enhancers, there has not been a systematic evaluation and

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<sup>1,2,7</sup>, 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 the 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 comparison of their coverage at 741 previously identified super-enhancers<sup>29</sup>. The result shows that assays using nuclear run-on followed by cap-selection (GRO-cap<sup>7</sup>/PRO-cap, CoPRO<sup>8</sup>) have a higher coverage over all other assays (Cohen's *d* of log-transformed coverage: 0.3926~1.1642).

5 Since super-enhancers can span several kilobase pairs in the genome and are believed to be composed of clusters of enhancers, we also compared the assay sensitivity in a higher resolution manner by examining their coverage in 803 (635 intergenic, 113 intronic, and 55 others) bona fide enhancers (referred to as “true enhancer set” in this study, Supplementary Table 2) validated by CRISPR/Cas9-mediated deletion and CRISPRi in K562 cells<sup>30-38</sup>. With the same sequencing  
10 depth, GRO-cap ranks first in terms of sensitivity: it covers 85.80% of true enhancers (70.98% divergent:  $\geq 5$  reads detected from both strands and 14.82% unidirectional:  $\geq 5$  reads detected only on one strand; Fig. 1c and Supplementary Fig. 1a). csRNA-seq comes in second place with 73.35% (46.33% divergent and 26.40% unidirectional) coverage of these validated true enhancers (Fig. 1c and Supplementary Fig. 1a).

15

We further evaluated the sensitivity of these assays by their ability to capture unstable transcripts. eRNAs are usually less stable compared to RNAs of other types, as seen in the decay rates<sup>39</sup> of transcripts originated from 514 enhancers from the true enhancer set, which are significantly more rapid than those of mRNAs (*p*-value: 1.022e-86, Cohen's *d*: 0.9469). We then  
20 used as the cutoff between stable and unstable transcripts the 95<sup>th</sup> quantile of decay rates of mRNAs and surveyed the distribution of read counts captured in the two categories among all assays. Consistent with our conclusion above, GRO-cap has the smallest differences in read coverage between stable and unstable transcripts [Cohen's *d*: -0.0028, 95% CI: (-0.0328,

0.0239)], indicating assays utilizing nuclear run-on followed by cap-selection have the greatest 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<sup>7</sup>, 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<sup>35</sup>(Fig. 2b). Another potential problem for 3' assays is that reads that are mapped to intergenic or intronic regions could be derived from either eRNAs, the unprocessed precursors of RNAs from other categories (e.g., pre-mRNAs), or read-through from an upstream transcription event (Supplementary Fig. 2b), which further reduces their sensitivity for detecting  
5 eRNAs.

Most of the RNA transcripts in living cells belong to the families of highly abundant RNAs, e.g., rRNAs and snRNAs. When capturing eRNAs with next-generation sequencing-based methods, a successful exclusion of RNAs of these high-abundance families from the sequencing libraries  
10 during the preparation process would greatly enhance the efficiency and sensitivity of eRNA identification. We compiled a comprehensive list of high-abundance RNAs in human cells by incorporating annotations from GENCODE<sup>40</sup>, RefSeq<sup>41</sup>, and RMSK<sup>42</sup>. Based on this list, an average of 7.124% of the mappable reads in each assay originated from rRNAs despite most of the assays employed strategies to reduce rRNA inclusion (Fig. 2c and Supplementary Fig. 2c).  
15 By simulating a rRNA-depleted BruUV-seq library, we found that a complete depletion of rRNAs could contribute to a 1.76-fold boost in detecting eRNAs (Supplementary Fig. 2d).

When the assay specifically enriches for short transcripts, like csRNA-seq, a relatively large proportion (31.49%) of the mappable reads were found to have originated from snRNAs (Fig. 2c  
20 and Supplementary Fig. 2c). Detection of such a disproportionately large fraction of snRNAs (Fig. 2c) suggests a potential contamination in the sequencing library from the splicing intermediates. To test the possibility, we calculated the signal densities at all the splice sites in the human



genome according to GENCODE annotation (v24)<sup>40</sup>. As shown in Supplementary Fig. 2d, csRNA-seq does detect more signals at the splicing junctions compared to all the other assays.

### **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<sup>43</sup>, many of these events are considered to be spurious transcriptional noise<sup>44,45</sup> because of their extremely low transcript yields compared to mRNAs and of the intrinsic promiscuity of RNA Pol II under certain circumstances<sup>46</sup>. 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)<sup>47-51</sup> and Self-Transcribing Active Regulatory Region Sequencing (STARR-seq)<sup>2,52,53</sup> studies and further removed elements overlapping with predicted Enhancer-Like Sequence (ELS) or Promoter-Like Sequence (PLS) from candidate cis-regulatory elements (cCRE) annotations<sup>54</sup> 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).

## **A novel computational tool for identifying active enhancers and promoters from 5' assays:**

### 5 **Peak Identifier for Nascent-Transcript Sequencing (PINTS)**

High-throughput sequencing-based bioassays rely on sophisticated analyses to identify meaningful information from raw data, so it is essential to choose the appropriate computational tool to do the job. Two categories of tools are currently available for signal processing. Tools in the first category predict entire transcription units, they are primarily used for 3' assays and often use a set of cutoffs for fold-changes, e.g., HOMER (GRO-seq)<sup>55</sup> or Hidden Markov Models<sup>19,56</sup>, e.g., groHMM, to determine the start and end positions of transcription units. Tools in the second category usually identify narrower regions for potential regulatory elements (mainly promoters and enhancers, often referred to as peak callers) and include GROcapTSSHMM<sup>7</sup>, dREG<sup>24</sup>, Tfit<sup>23</sup>, dREG.HD<sup>26</sup>, TSScall<sup>11</sup>, HOMER (csRNA-seq)<sup>15</sup>, and FivePrime<sup>27</sup> (based on Paraclu<sup>28</sup>). Based on previous studies, the peak of divergent TSSs, which is associated with eRNA transcription, is an effective mark for active enhancers<sup>1,2</sup>. Therefore, in this comparative study, we focused on the second category of computational tools.

To achieve a higher resolution in identification of transcriptional regulatory elements, a common practice is to only look at the ends of the transcripts captured by these assays. When such a practice is adopted, two critical issues emerge when evaluating the statistical significance of peaks (i.e., TSSs), especially for the 5' assays. First, when only taking the transcript ends into account, the fraction of zeros (no mapped reads per basepair) in local background increases,

which deflates read density in the local background and thus inflates the statistical significance of candidate TSSs and resulting in false positives. Second, multiple TSSs can localize in close proximity in the genome and therefore inflate the estimation of read density in the local 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<sup>57</sup> 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<sup>2</sup> (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<sup>2</sup>.

15

**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<sup>57</sup>, a widely used peak caller for analyzing ChIP-seq data, is remarkably distinct from those of the true enhancers, suggesting the default shifting model of MACS2 optimized for ChIP-seq data may not be suitable for identifying eRNA TSSs of active enhancers.

5

We further surveyed the size distribution of these elements as an indication of peak-calling resolutions. We found that PINTS, GROcapTSSHMM, Tfit, and TSScall achieve higher resolution (average peak size of 185-258 bp) than other tools when analyzing 5' assay data (Fig. 4c and Supplementary Fig. 3m for all other assays). Notably peaks called by dREG.HD and MACS2 range between 548-751 bp in size, whereas dREG, FivePrime, and HOMER peaks range between 381-460 bp.

10

The amount of computational resources required by a peak caller greatly affects its general applicability. Here, we compared the total amount of CPU time and peak memory usage that each peak caller requires for identifying divergent elements from 5' assays (Fig. 4d and 4e). Based on our calculation, it is feasible to run PINTS, MACS2, HOMER, GROcapTSSHMM, TSScall, and FivePrime on a typical personal computer.

15

### **PINTS achieves high overall robustness, sensitivity and specificity in detecting TREs from**

20 **5' assays**

A common strategy in functional studies of enhancers is to compare active enhancers across different conditions and diseases<sup>21,22</sup>, which requires the *in silico* enhancer-predicting tools to be robust against biological and experimental variances. To test the upper bound of robustness for

the aforementioned tools, we performed peaks calls using these tools on datasets from 12 assays by aligning each dataset to two commonly-used builds of human reference genome: hg19 and 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<sup>58</sup>, 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<sup>40</sup> 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, and applicability to different RNA sequencing assays in Supplementary Fig. 5.

### **An enhancer compendium for human cell and tissue types**

5 Having comprehensively evaluated the performance of all available whole-genome RNA sequencing assays and computational tools in identification of candidate enhancers genome-wide, we reasoned that available 5' assay datasets across different human cells and tissues could be analyzed by PINTS to construct a human enhancer candidate compendium. Previous studies have shown that, compared with histone marks, detecting enhancers by divergent eRNA TSSs  
10 has advantages in both resolution and specificity<sup>1,2</sup>. Such an eRNA-centric enhancer compendium, in addition to all the available enhancer datasets based on histone marks<sup>54,59</sup>, will be an invaluable resource to better understand gene regulation, to functionally annotate the non-coding genome, and to help prioritize non-coding variants across disease cohorts by their potential impact on enhancer activities<sup>60</sup>. Toward this goal, we applied PINTS to identify  
15 candidate enhancers using 5' assay datasets (i.e., GRO/PRO-cap, Co-PRO, csRNA-seq, NET-CAGE, RAMPAGE, CAGE, and STRIPE-seq) across 33 cell lines, 7 *in vitro* differentiated cells, 35 primary cells, and 45 tissue samples, including all available 5' assay datasets through the ENCODE portal (Fig. 5a). Such a comprehensive catalog of enhancers across a wide range of human cells and tissues analyzed by the same exact computational pipeline provides an excellent  
20 resource to perform meaningful comparative genomic analyses to study the dynamics of enhancers and gene regulation in general, which will help focus on true biological differences while minimizing technical variations. In addition, for 7 human cell lines (K562, GM12878, HepG2, HeLa-S3, MCF-7, H9, and HCT116), we applied all other available tools (FivePrime,

HOMER, TSScall, dREG, dREG.HD, and Tfit) to identify candidate enhancers. We believe this unique resource of enhancers in 7 cell lines with multiple 5' assay datasets analyzed by all available computational tools will greatly help further the studies of enhancers and their key 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<sup>35</sup> 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).

### **An online web server for exploring and analyzing enhancers and various genomic and epigenomic features genome-wide**

To make it easier for biologists to explore the tens of thousands of candidate enhancers in our compendium and to prioritize these enhancers for further studies, we constructed an online web server (<https://pints.yulab.org>) that is publicly available, where users can query any human genomic region of their interest in a given biological sample to get a comprehensive list of candidate enhancers detected by any available 5' RNA sequence assays (i.e., GRO/PRO-cap, csRNA-seq, NET-CAGE, CAGE, RAMPAGE, and STRIPE-seq) in that sample using any of the 7 peak callers (PINTS, FivePrime, dREG, dREG.HD, HOMER, Tfit, and GROcapTSSHMM). We also included detailed annotations for each candidate enhancer, including epigenetic features<sup>54</sup>, core promoter elements<sup>61</sup>, potential transcription factor binding sites<sup>62</sup>, and presence of population variants and ClinVar mutations<sup>63</sup>, as shown in Fig. 6. Users can prioritize enhancers by requiring specific supporting 5' assay data available, the presence of certain epigenomic features, core promoter elements and transcription factor binding sites, or containing different categories of population variants (common vs. rare) and ClinVar mutations (benign vs. pathogenic). For the transcription factor binding site analysis, our web server will automatically integrate the information from available RNA-seq data to only include those factors that are expressed in the selected cell line.



Furthermore, users can upload their own enhancers calls in a human cell line or type, our web server will automatically annotate all of their enhancers calls with the same pipeline as described above. We have integrated epigenomic features and RNA-seq data for 197 human-derived biosamples from the ENCODE project in our web server. For user-uploaded enhancer data in  
5 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;  
10 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-  
15 wide<sup>1,2</sup>; 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  
20 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<sup>2</sup>. 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 address this issue, we introduced a new tool, PINTS, which was inspired by MACS<sup>57</sup> but with unique implementations (e.g., using zero inflated Poisson distribution to account for sparsity of only mapping 5' of all reads) that were designed specifically for defining TSSs precisely when analyzing 5' RNA sequencing data. Our benchmarks indicate that PINTS achieves the best balance among robustness, applicability, sensitivity, and specificity, especially for 5' RNA sequencing assays capable of detecting the precise location of eRNA TSSs.

In this study, we used CRISPR/Cas9 and CRISPRi validated enhancers<sup>30-38</sup> as the positive reference set, and MPRA/STARR-seq negative segments<sup>2,47-53</sup> as non-enhancers. Although these two sets show quite different epigenomic profiles (Supplementary Fig. 2f), which indicates that our non-enhancer set is depleted of any true enhancers, there might still be a few false negatives in the non-enhancer set because not every enhancer works with every promoter, but in the published MPRA/STARR-seq datasets, only a very small number of promoters were used to test all candidate elements; furthermore, some tested elements might be truncated due to synthesis limitations (<200 bp) or random fragmentation of the genome. However, such cases are not expected to affect our relative ranking of different assays and thus will have minimal impact on our conclusions.

In summary, the comprehensive comparison and thorough analyses of 13 genome-wide RNA sequencing assays and 9 computational tools presented here help better understand the strengths and weaknesses for each assay and tool, with regard to detecting active enhancers *in vivo*. Given the fundamental importance of enhancers in gene regulation and the immense interest in accurately identifying active enhancers in a wide variety of samples across species, our results 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<sup>2,7</sup> 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 cell types 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.

## Methods

### Data preprocessing

All datasets were managed and analyzed with BioQueue<sup>64</sup> (Supplementary Table 1). Raw reads were preprocessed with fastp<sup>65</sup> according to the experimental designs reported in original papers.

- 5 Only reads longer than 14bp were kept for downstream analyses. All processed reads from RNA assays were aligned using STAR<sup>66</sup> to primary assemblies of human reference genome hg38 (GCF\_000001305.15) together with rDNA (U13369.1) with parameters as --outSAMattributes All --outSAMmultNmax 1 --outFilterMultimapNmax 50. For studies using Drosophila cells (dm6) or other specific samples as spike-ins, Drosophila reference genome (dm6) or the
- 10 corresponding reference sequences used in the original studies were incorporated into the index. To measure the robustness of peak prediction, we also mapped reads to primary assemblies of human reference genome hg19 (from UCSC, sequences from alternative loci/haplotypes were removed the same way as for hg38).

### Determining read coverage among reference regions

- 15 Sequencing reads of replicates for the same assay were merged together and downsampled to the same sequencing depth (the same number of mappable reads) for three times using picard with parameter STRATEGY = Chained. These downsampled data were then converted to bed files to calculate the fraction of overlap between the sequencing reads and the reference regions in a strand-specific manner.

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

Transcript annotations derived from TT-seq (GSE75792<sup>39</sup>) were downloaded from the GEO database. Transcription units with NA values were discarded. The 95<sup>th</sup> quantile of estimated decay rates for mRNAs was used as the cutoff between the unstable (above the cutoff) and stable  
5 (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)<sup>40</sup>: 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  
10 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

15 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<sup>40</sup>. 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%  
20 confidence interval (estimated by bootstrap) of the averages were reported.

### Compiling the true enhancer and non-enhancer sets

The experimentally quantified enhancer activity of various DNA elements were collected from previous studies (enhancers: 938 from CRIPSR<sup>32</sup> or CRIPSRi<sup>30,31,33-38</sup>; non-enhancers: 20,941 from STARR-seq<sup>2,52,53</sup> and 17,462 from MPRA<sup>47-51</sup>). Overlapping elements within the same category were merged until the resulting elements overlap with elements in the other category. Non-enhancer loci were excluded in the final set if they 1) were shorter than 250bp; 2) overlapped with PLS or ELS predicted by cCRE<sup>54</sup>, or 3) overlapped with possible promoters (1kb regions flanking TSSs in GENCODE).

### Calculating FDRs for assays

multiBamSummary was used to generate tables of read counts across the genome in 500-bp bins. For each assay, the bins were ranked by counts in them and the top n bins were considered as true signals from the assay (four cutoffs were tested: 5,000, 10,000, 20,000, and 100,000, Supplementary Fig. 2h). If a bin overlapped with a locus in the true enhancer set, then the bin was considered as a true positive (TP); if a bin overlapped with a locus in the non-enhancer set, it was considered as a false positive (FP). The false discovery rates were calculated as:

$$FDR = \frac{\sum FP}{\sum TP + \sum FP}$$

### PINTS

Briefly, read ends are separated based on their mapping directions on the reference genome (forward or reverse), and the read counts are binned into 100-bp windows. Adjacent windows with reads available are merged to avoid splitting potential TRE elements. Within each window, the algorithm first finds peak seeds using a prominence-based approach. Then with a maximum-

scoring pairing strategy<sup>28</sup>, 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  $\alpha$  is 1, and 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 ( $\lambda$  for read density,  $\pi$  for the proportion of zeros that are not coming from a Poisson process), the probability mass function of these models has the following form:

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

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:

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

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

$$\hat{\pi}^{(r+1)} = \frac{\sum_{i=1}^I \hat{z}_i^{(r+1)}}{I}$$
$$\hat{\lambda}^{(r+1)} = \frac{\sum_{i=1}^I (1 - z_i) x_i}{\sum_{i=1}^I (1 - z_i)}$$

An interquartile range (IQR)-based refinement is applied before fitting ZIP models to the neighborhood regions. In this case, if certain peak candidates  $s$  in a local environment are considered as outliers by IQR (their densities are above  $Q3 + 1.5 \times IQR$ , where  $IQR = Q3 -$   
5  $Q1$ ), these candidates will be masked. For libraries with low sequencing depth, instead of simultaneously masking all outlier peak candidates in the local background, PINTS masks one peak candidate at a time and calculates the resulting peak density in the local background. The process will be reiterated until PINTS either identifies the outlier candidates or reports the non-existence of such outliers. The estimated densities are then used to determine the statistically  
10 significant peaks, which are further categorized into divergent peak pairs (peaks on opposite strands and within 300 bp) and unidirectional peaks.

### Generating peak calls with existing tools

Peak calls for different assays were made using default parameters for other existing peak callers with the following exceptions. For MACS2, `--keep-dup all` was set so that reads mapped to the  
15 same loci would be kept. For FivePrime, parameters  $D_{min}$ ,  $P_{min}$ , and  $S_{min}$  were optimized according to the sequencing depth of corresponding libraries, and both divergent TSS calls and enhancer calls were combined as the final output. All tools were allowed to create up to 16 threads/subprocesses if they allow multithreading or parallel computing. For peak callers that do not primarily identify divergent peaks, unidirectional peaks were paired as long as they were  
20 within 300 bp and on opposite strands. Maximum memory usage and CPU time (sum of all threads) were monitored with the help from BioQueue<sup>64</sup>. All peak calls were generated on



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  
5 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  
10 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):

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

### ROC

15 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  
20 overlapping with the positive and negative sets under different cutoffs of scores: *p*-values of peaks for PINTS and MACS2, FDRs for TSScall, output SVR scores for dREG, likelihood ratio

scores for Tfit, and peak scores for HOMER (findcsRNATSS.pl and findPeak -style TSS). For dREG.HD, GROcapTSSHMM, FivePrime, and HOMER (GRO-seq), since there are either no scores or multiple scores returned in the final output, the sensitivity and specificity was evaluated and reported with their default parameters.

## 5 PINTS web server

The PINTS web server is powered by the Django web framework. dbSNP v153<sup>67</sup> was used for annotating mutations among TREs; JASPAR 2020<sup>62</sup> was employed for annotating transcription factor registries and only TFs expressed in the corresponding the biosample (based on RNA-seq data from ENCODE) were included (see Supplementary Table 4 for list of datasets used and  
10 accession information). cCRE v2<sup>54</sup> was used for epigenomic annotation. Core promoter elements were annotated using the following strategy: for each major TSS (+1), the portal annotated the elements as having an initiator or an initiator-like element when the sequence of -3~+3 matches BBCABW; or the sequence of +1~+2 matches YR, respectively. The TATA box (-32~-21) and DPR elements (+17~+35) were identified using the previously published SVR model<sup>61</sup>.

15

**Acknowledgments:** Computation was performed on a cluster administered by the Biotechnology Resource Center at Cornell University. We thank members of the Yu and Lis laboratories, and the ENCODE Consortium (specifically Ali Mortazavi, Mats Ljungman, and Jill E. Moore) for helpful discussions and guidance; Hongya Zhu for her suggestions on concept visualization. This  
20 work was supported by a grant from the National Institutes of Health (HG009393 to J.T.L. and H.Y.). L. Y. was supported by the Cornell Presidential Life Sciences Fellowship.

**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.

5

**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

10 can be retrieved from [https://github.com/hyulab/PINTS\\_analysis](https://github.com/hyulab/PINTS_analysis). All sequencing data was retrieved from public databases (NCBI GEO and ENCODE portal), lists of accessions are available in Supplementary Table 1 and 2.

## Supplementary Materials

15 Figures S1-S5

Supplementary Notes

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).

13. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
14. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed  
5 cis-regulatory elements. *Nat. Genet.* **51**, 1369–1379 (2019).
15. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* **29**, 1836–1846 (2019).
16. Policastro, R. A., Raborn, R. T., Brendel, V. P. & Zentner, G. E. Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq. *Genome Res.* **30**, 910–923 (2020).
- 10 17. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
18. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526–540 (2015).
19. Paulsen, M. T. *et al.* Coordinated regulation of synthesis and stability of RNA during the acute TNF-  
15 induced proinflammatory response. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2240–2245 (2013).
20. Magnuson, B. *et al.* Identifying transcription start sites and active enhancer elements using BruUV-seq. *Sci. Rep.* **5**, 17978 (2015).
21. Chen, H. *et al.* A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **173**, 386–399.e12 (2018).
- 20 22. Zhang, Z. *et al.* Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. *Nat. Commun.* **10**, 4562 (2019).
23. Azofeifa, J. G. & Dowell, R. D. A generative model for the behavior of RNA polymerase. *Bioinformatics* **33**, 227–234 (2017).
24. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data.  
25 *Nat. Methods* **12**, 433–438 (2015).
25. Wang, Z., Chu, T., Choate, L. A. & Danko, C. G. Identification of regulatory elements from nascent

- 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).
- 5
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).
- 10
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
- 15
- 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
- 20
- 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).
- 25
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).

38. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
39. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
40. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
- 5 41. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
42. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
- 10 43. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
44. Field, A. & Adelman, K. Evaluating Enhancer Function and Transcription. *Annu. Rev. Biochem.* **89**, 213–234 (2020).
45. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).
- 15 46. Palazzo, A. F. & Koonin, E. V. Functional Long Non-coding RNAs Evolve from Junk Transcripts. *Cell* **183**, 1151–1161 (2020).
47. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
48. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
- 20 49. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).
50. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).
- 25 51. Maricque, B. B., Chaudhari, H. G. & Cohen, B. A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat. Biotechnol.* (2018)

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).
- 10 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).
- 15 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).
- 20 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
- 25 (2020).
64. Yao, L., Wang, H., Song, Y. & Sui, G. BioQueue: a novel pipeline framework to accelerate



- bioinformatics analysis. *Bioinformatics* **33**, 3286–3288 (2017).
65. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
66. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- 5 67. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

## 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

the different patterns of signals captured by 5' (enriched in the promoter and enhancer regions) vs. 3' (enriched in the promoter and gene body regions) assays. Signals are normalized by reads per million (RPM). **c**, Proportion of mappable reads from different assays originated from various abundant RNA families. **d**, Specificity in detecting eRNAs. Top track: differences of read coverage among true enhancer and non-enhancer loci, a pseudo-count (1) was added to each locus, and the coverage was log-transformed. Bottom track: signal-to-noise ratios depicted in forms of fold enrichment (FE), the results are based on bootstrapped samples, median statistics are used to calculate the fold changes. **e**, False discovery rates (FDRs) estimated by the overlap between the top 10,000 genomic bins and the reference (true and non- enhancer) loci. Pseudo-replicates were used (N=3); error bars represent SD.

**Fig. 3. Peak Identifier for Nascent-Transcript Sequencing (PINTS): a novel computational tool for identifying active enhancers and promoters from 5' assays.** **a**, Improvement of TSS identification resolution by focusing only on read ends and using zero-inflated Poisson (ZIP) models to fit local background to address the significantly increased sparsity of signals. The thin grey lines indicate sequencing reads with the 5' ends highlighted in red. **b**, The existence of other potential true peaks (pink) elevates the estimation of read density in the local background. **c**, Schematic plot shows how IQR-ZIP works. The blue box shows the read density distribution of the local background; the purple dot shows the density of the peak to be tested; the pink dot shows the density of a potential true peak close to the peak to be tested, whose read density is a clear outlier and thus excluded from local background estimation.

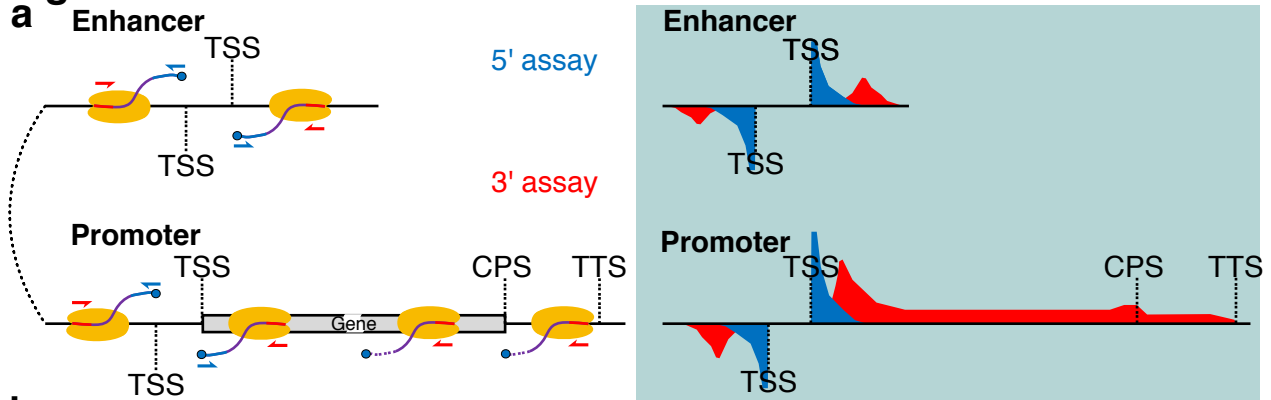
**Fig. 4. PINTS achieves the best balance among resolution, robustness, sensitivity, specificity, and computational resources required.** **a** and **b**, Profiles of GATA1 binding sites and H3K27ac pattern in true enhancer regions and distal TREs identified by different peak callers. **c**, Distribution 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

the corresponding assay. The number of mappable reads for each assay is listed in the parentheses. **c**, A known enhancer locus where most of these assays captured signals on both strands. **d**, A known enhancer locus where only GRO-cap and csRNA-seq captured detectable signals. **e**, A known enhancer locus where only GRO-cap captures clear signals. Signal tracks in **c**, **d**, and **e** were normalized by their sequencing depths (RPM).

**Fig. 6. Interactive PINTS web server.** The server is composed of three modules: a catalog database containing our human enhancer compendium across 120 biosamples, a user interactive interface, and analytical engines. When the user provides genomic loci of their interest as search queries, previously annotated information related to the loci will be retrieved from the catalog database. The user can further refine the search results by a series of filters, including the supporting assays, callers, the presence of mutations, transcription factor binding sites, core promoter elements, and epigenomic marks. On the other hand, if the user chooses to upload peak calls generated from their own data, the uploads will be analyzed by the analytical engines to provide the same types of annotations as those in the catalog database.

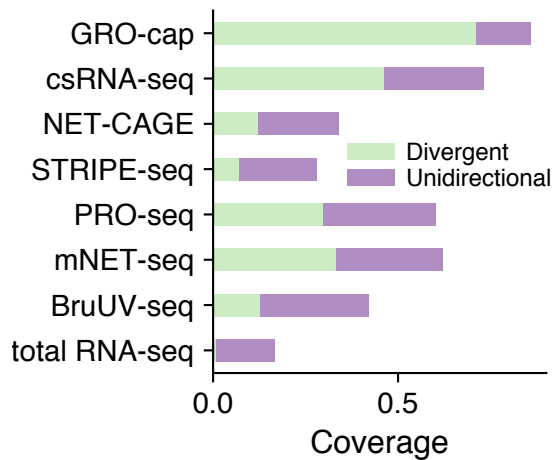
**Fig. 1**



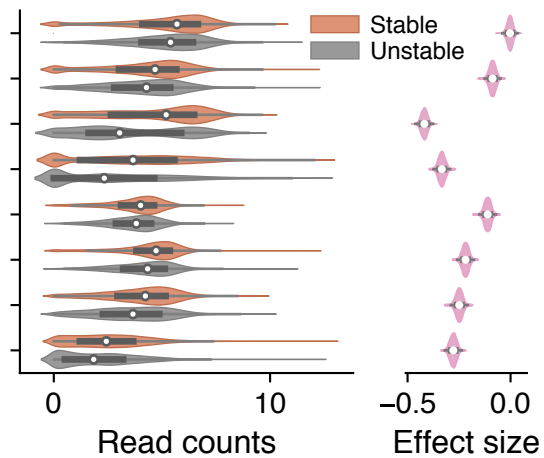
**b**

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

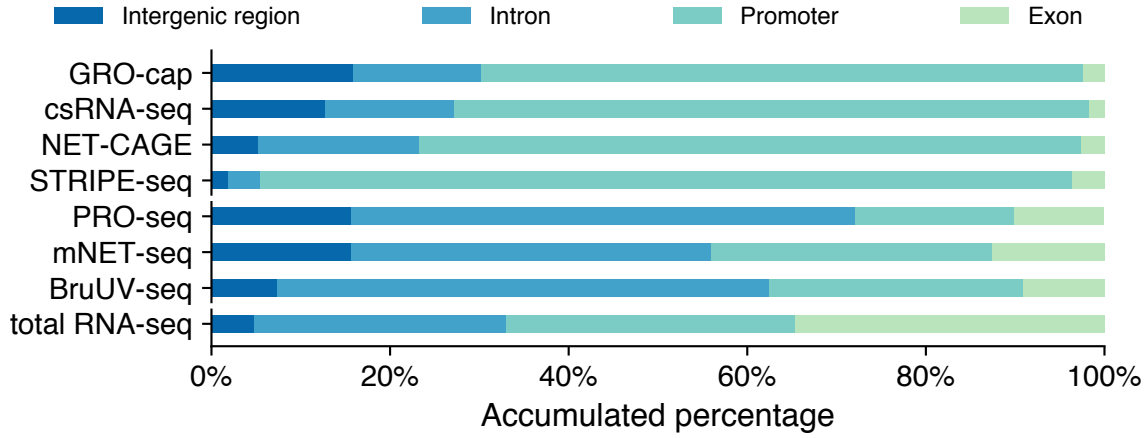
**c**



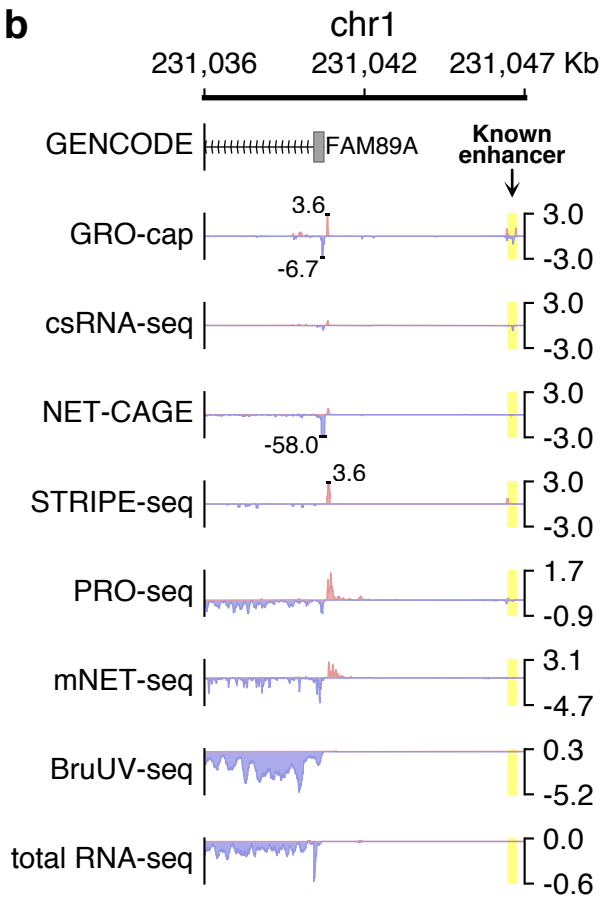
**d**



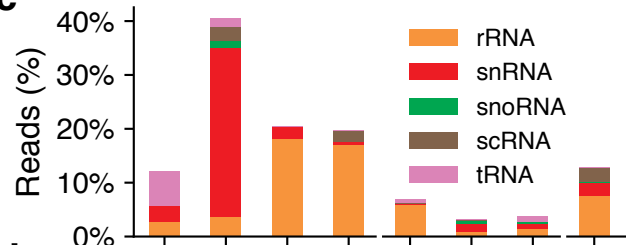
**Fig. 2**  
**a**



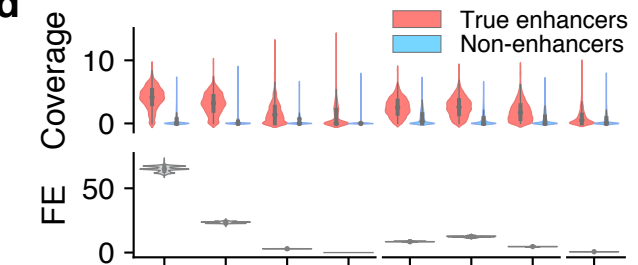
**b**



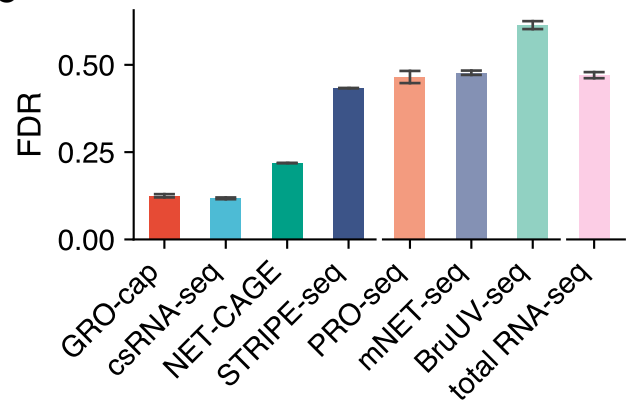
**c**



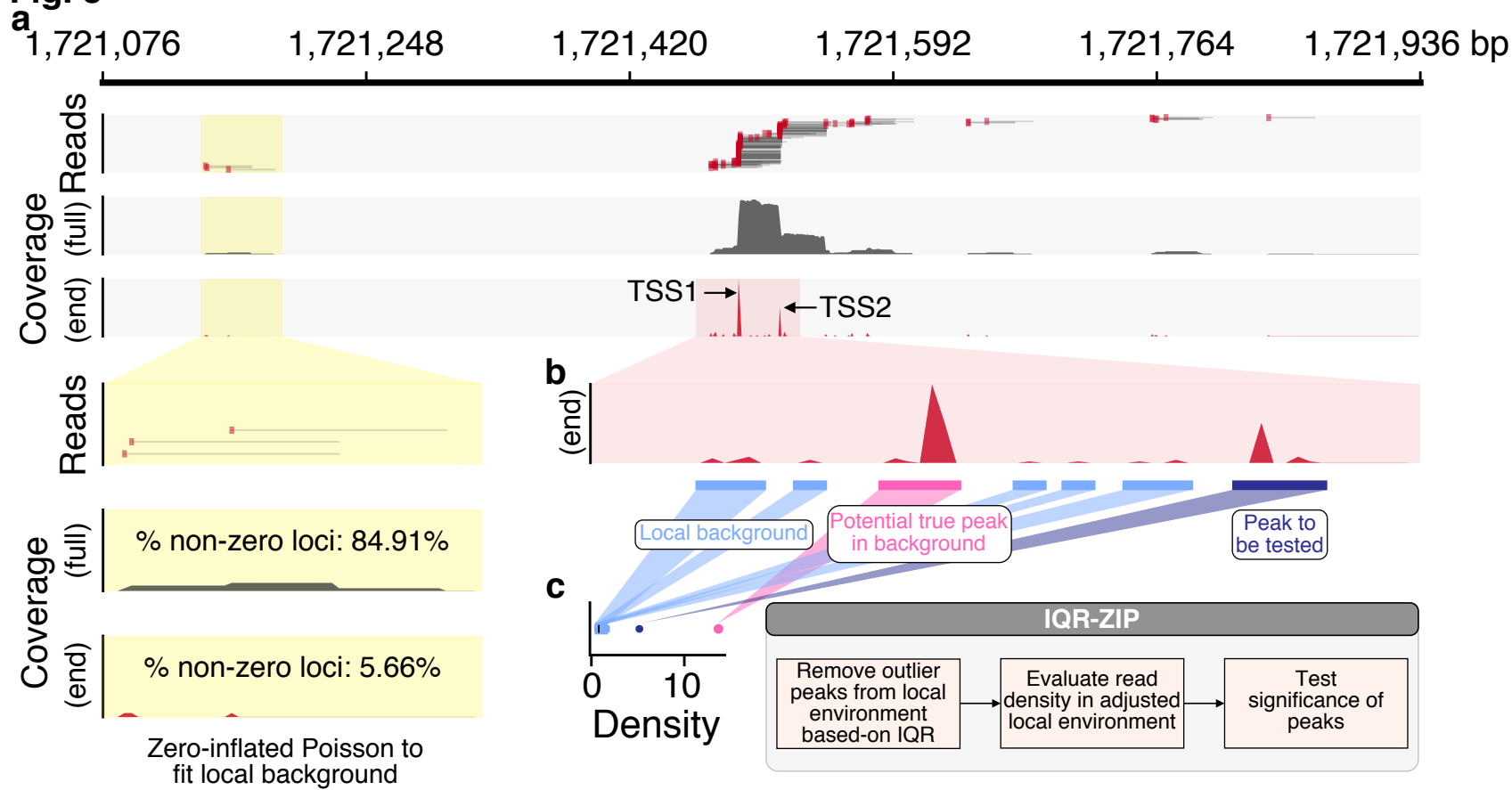
**d**



**e**

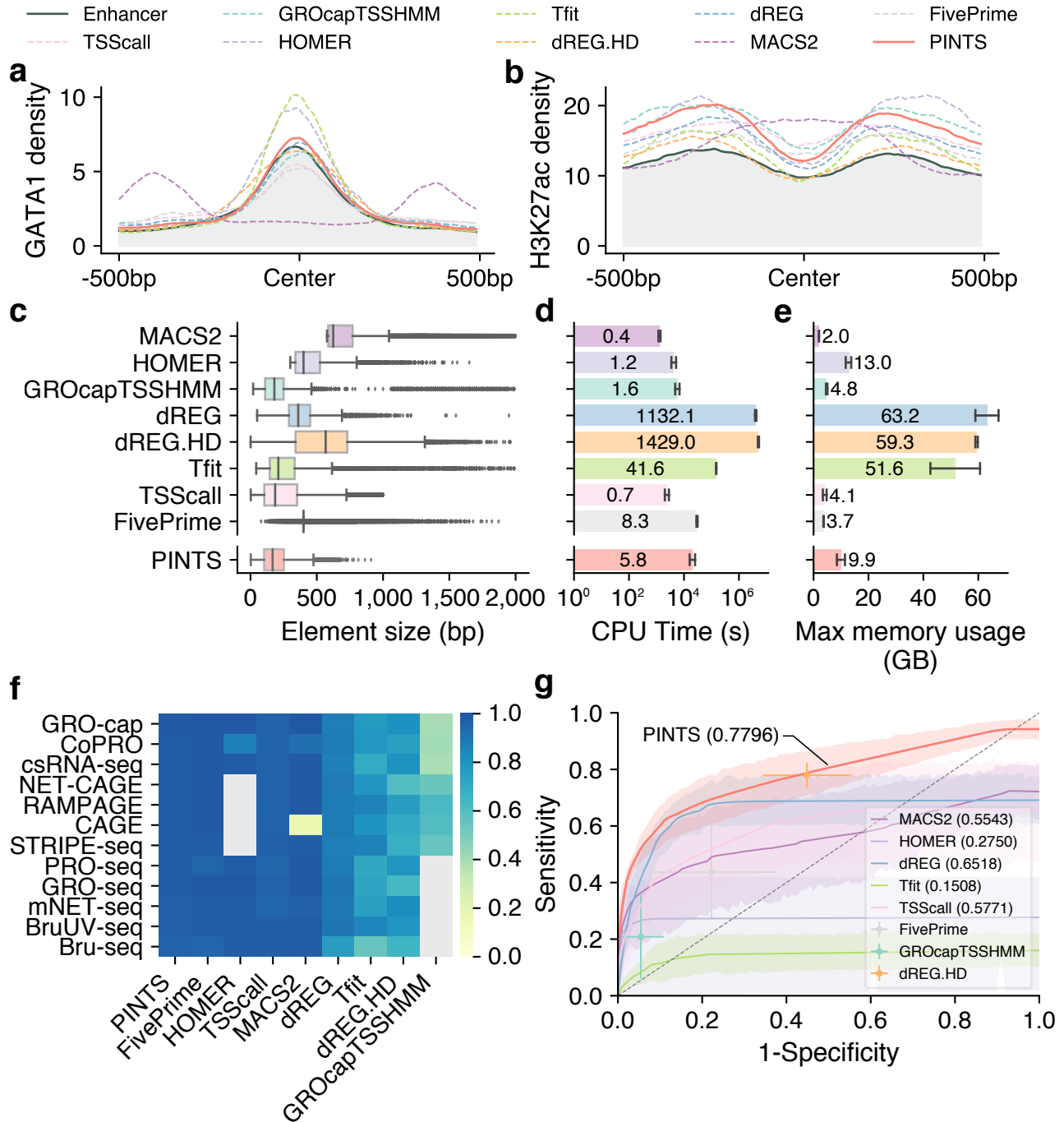


**Fig. 3**



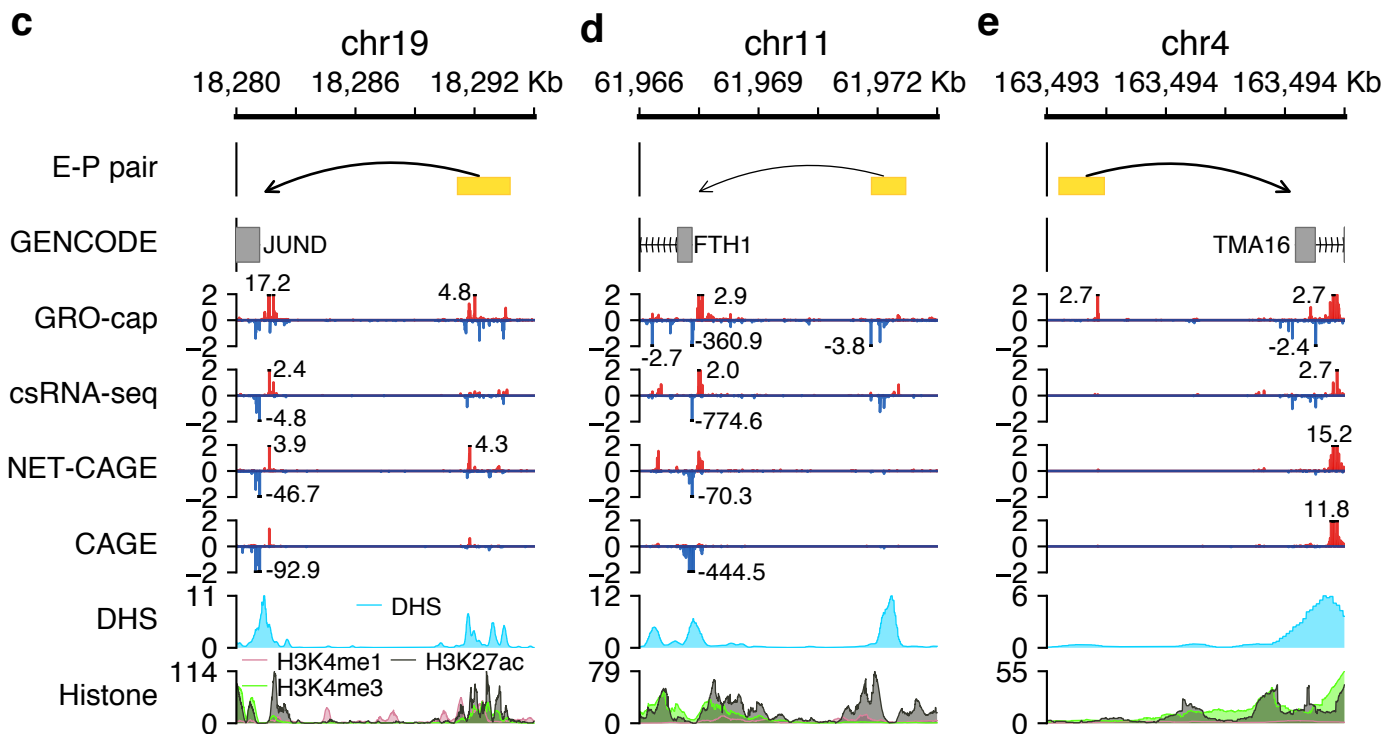
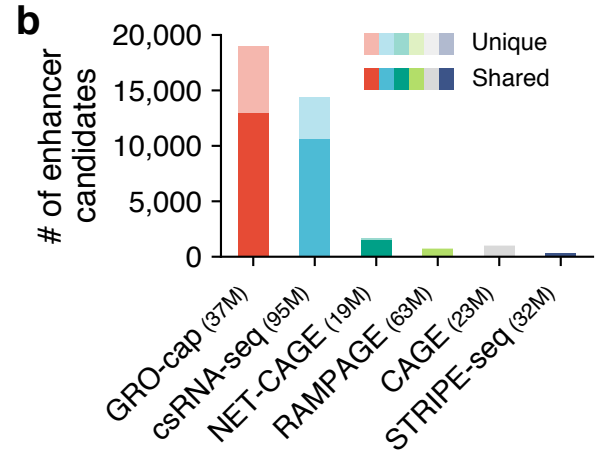


**Fig. 4**



**Fig. 5**  
**a**

Biosample	GRO/PRO-cap	csRNA-seq	NET-CAGE	CAGE	RAMPAGE	STRIPE-seq
K562	19.0	14.4	1.7	1.1	0.8	0.3
GM12878	15.6		12.6	0.8	1.2	
HepG2			2.6	1.6		
HeLa-S3			3.4	1.7		
MCF-7			9.8	1.5		
H9		5.4				
HCT116		14.9				
Cell lines	2	3	5	11	22	1
Differentiated cells				1	6	
Primary cells				20	15	
Tissues					45	



## 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



### 3. Analytical engines

