Exploring the epigenome profiles of repetitive elements with the WashU Repeat Browser

Jiawei Shen, Siyuan Cheng, Deepak Purushotham, et al.

Genome Res. published online February 21, 2025
Access the most recent version at doi:10.1101/gr.279764.124

Exploring the epigenome profiles of repetitive elements with the WashU Repeat Browser

Jiawei Shen[1, 2], Siyuan Cheng[1], Deepak Purushotham[1], Xiaoyu Zhuo[1], Alan Y. Du[1], Wenjin Zhang[1], Daofeng Li[1, *], Ting Wang[1, 2, 3, *]

[1] Department of Genetics, The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, USA
[2] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, USA
[3] McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

Abstract

Repetitive elements, mostly derived from transposable elements (TEs), account for half the DNA in human and other mammalian genomes. Although epigenetic mechanisms, including DNA methylation and repressive histone modifications, have evolved to suppress TE activities, TEs have substantially shaped the regulatory landscape of the host genome by contributing regulatory sequences to it. TE-derived sequences are often highly repetitive and thus have low mappability, making it difficult to profile the genomics of TEs using short-read sequencing technology. Many specialized bioinformatics tools have been developed for TE-related analysis, but meaningfully visualizing, navigating, and interpreting such data remains challenging.

Here, we describe the WashU Repeat Browser to host genomics profiles of human and mouse TEs using data produced by the ENCODE Project and to support the navigation, interactive visualization, integration, comparison, and analysis in the context of TEs. WashU Repeat Browser is a web-based platform allowing users to browse genomic and statistical signals over repetitive elements derived from ENCODE, Roadmap, and FANTOM datasets. The Browser provides a TE-centric view including TE subfamily enrichments, TE subfamily profiling, as well as overviews of genomic signals on individual TE loci where we extend the WashU Epigenome Browser to display user-selected datasets and TE loci. These features could help to close the gaps in our understanding of the repetitive sequences and their putative regulatory functions and aid investigators in formulating new hypotheses by integrating their data with public data.

Introduction

A large portion of eukaryotic genomes is derived from transposable elements (TEs) (Kidwell and Lisch 2001; Kazazian Jr 2004; Feschotte and Pritham 2007; Kapitonov and Jurka 2008). Recent studies suggest that sequences derived from TEs could be responsible for wiring tissue or cell type-specific regulatory networks (Chuong et al. 2017; Trizzino et al. 2018) and have acquired tissue-specific epigenetic regulations (Bourque et al. 2008; Chuong et al. 2013; Wang et al. 2015). Studies have also revealed that TE DNA methylation is dynamic during development and is connected to epigenetic control of cell type-specific enhancer functions (Kunarso et al. 2010; Xie et al. 2013; Dominguez et al. 2016). TEs are known to contain transcriptional

enhancers (Bourque et al. 2008; Chuong et al. 2017). The extent to which TEs contribute to cell type-specific programs of gene expression is unknown but could be large, given how prevalent TEs are in mammalian genomes. Taken together, preliminary findings suggest that epigenetic regulation of TEs, both suppression and activation, is important for normal development (Sundaram et al. 2014; Sundaram and Wang 2018; Pehrsson et al. 2019). It is necessary to explore this hypothesis more deeply to understand the sequence features that distinguish between transcriptionally active (i.e., can influence gene expression) and inactive TEs and to understand the evolutionary events that lead to TE sequences gaining regulatory functions.

Repetitive elements in genomes are mostly derived from TEs, and they are classified based on their sequence characteristics (Lerat 2010; Bourque et al. 2018; Miao et al. 2020). Annotation of repetitive elements has always been one of the first steps when any genome is sequenced. Efforts represented by Repbase (Bao et al. 2015) and RepeatMasker (Smit et al. 2015) have annotated repeat sequences by grouping them hierarchically into classes, families, and subfamilies. Repeat sequences are categorized into families according to the reconstructed replication history and into subfamilies based on finer features aimed at capturing the evolutionary history within families (Hubley et al. 2016; Carey et al. 2021). In the era of high-throughput sequencing-based functional genomics analysis, however, repetitive elements have presented a challenge. Since repeat elements, as the name suggests, are highly repetitive in the genome, most alignment tools discard reads that cannot be designated to a single location in a given genome assembly by default. Popular alignment tools such as Bowtie 2 (Langmead and Salzberg 2012) report the possible best match among multiple alignments based on the MAPQ (Mapping Quality) score, which is a measure of the confidence in the alignment of a read to a reference genome; STAR (Dobin et al. 2013) disregard reads mapped to multiple locations unless users select specific options to keep that kind of reads; BWA (Li and Durbin 2009) by default keeps those reads but assigns the mapping quality to 0. Specialized tools have been developed to help improve mapping short reads to repeats (Li et al. 2008; Alser et al. 2021; Burkes-Patton et al. 2023). Nonetheless, standards are still lacking, and few platforms exist for users to visualize and interact with genomics data in the context of different types of repeats.

Public consortiums such as ENCODE (The ENCODE Project Consortium 2012) and Roadmap Epigenomics (Roadmap Epigenomics Consortium et al. 2015) have generated a large amount of genomics data using short-read sequencing technology, but efficient ways for exploring such data in the context of repeat elements have been much less developed. Indeed, TEs are routinely deprioritized in large genomic studies, including ENCODE and Roadmap Epigenomics projects. To address this gap, we have extended the

widely used WashU Epigenome Browser to develop the WashU Repeat Browser as an integrative platform for TE genomics data analysis and visualization. We implemented an algorithm called iteres (Xie et al. 2013; Sundaram et al. 2014) to analyze TE profiles using data from public projects, it also enabled users to process their data in BAM or FASTQ formats, including ChIP-seq data such as transcription factor (TF) binding, histone modification, and cap analysis gene expression (CAGE-seq) data, by rescuing non-uniquely mapped reads. An average of 15% more reads can be rescued using this algorithm for constructing epigenomic or expression profiles on TEs (Supplemental Figure S1). The front-end of the WashU Repeat Browser provides intuitive data visualization of TE profiles. It integrates and displays epigenomic data on repetitive elements and provides an easy and intuitive way to explore biological insights about repeats. WashU Repeat Browser also allows investigators to perform their analysis with the provided data processing pipeline and compare their data with publicly available data using the data uploading function. Documentation and tutorials on this Browser can be accessed at https://rb-doc.readthedocs.io/.

Results

Data processing pipeline

It has been a challenge to analyze sequencing reads from TEs, which motivated the development of various methodologies (Rebollo et al. 2012; Notwell et al. 2015; Chuong et al. 2016). A typical approach in profiling the repeat elements in genomic assays, including ChIP-seq, CAGE-seq, ATAC-seq, and DNase-seq, maps chromatin features aggregated by raw reads to the linear genome and focuses on interpreting the epigenomic data along the linear genome. In doing so, reads that cannot be uniquely mapped to specific TE copies, i.e., multi-mapped reads, are typically discarded, resulting in loss of information. A main limitation of this type of analysis strategy is that valuable information on the phylogenetic relatedness among the TE copies is ignored. Even though a substantial amount of reads cannot be mapped to specific TE copies, they can often be confidentially assigned to specific TE subfamilies, thus contributing to the study of the epigenomic patterns of repeat subfamilies. We adopted the strategy to perform the analysis at both the individual copy level and at the repeat subfamily level (Figure 1A). Raw genomic sequencing data in FASTQ format were first mapped to the genome assembly while retaining multi-mapped reads. We then implemented a tool called iteres to access the alignment statistics of repeats in each dataset (Figure 1B).

Iteres rescues the reads mapped to multiple genomic locations by assessing the mapping qualities of all mapped locations and assigning reads to

appropriate subfamilies (Figure 1C). Reads were discarded only when they could not be assigned unambiguously to a subfamily. Mapping locations from individual genomic loci were converted to locations in the consensus sequence of each subfamily based on the RepeatMasker annotation provided by the UCSC Genome Browser (Karolchik et al. 2009). Details of the conversion of genomic coordinates to TE consensus coordinates are provided in Supplemental Method S1. We then calculated the enrichment score based on each repeat subfamily rather than any specific loci. To visualize these results, we set the TE subfamily's consensus sequence as the x-axis coordinate (like a sequence assembly) and aggregated the read information across the consensus to visualize on the y-axis. The cumulative sum of these reads is essentially a distribution of reads across each position within the consensus sequence. Figure 1 illustrates the data analysis and different visualization components, and Figure 1D describes the visualization pipeline for the processed data.

Data visualization components

We developed several novel ways of displaying data on TE: 1) Heatmap view of enrichment scores of TE subfamilies; users can click on a specific cell in the heatmap and jump to the consensus panel of the TE subfamily; 2) Consensus view that displays data anchored on a TE consensus sequence; signals from both unique reads and multi-reads are displayed; 3) Genome view which is a bird’s eye view of enrichment score of individual TE copies across the genome; a filter is provided to set the threshold of enrichment scores to adjust the visual effect. Examples of each visualization component are displayed in case studies as in Figures 2 and 3.

In Figures 2A and 3A, the Heatmap view provides users with an intuitive and interactive way to visualize the data. The Heatmap view functions as an “aerial view” presenting a heatmap-based representation of multiple datasets across multiple repeat subfamilies. The y-axis of the heatmap represents multiple datasets, and the x-axis represents multiple TE subfamilies. We draw the heatmap by the scores that represent the enrichment levels of different datasets relative to the repeat subfamilies. This display enables the visual profiling of multiple datasets, grouped by the associated metadata, and facilitates comparisons across selected TE subfamilies. By utilizing Clustergrammer (Fernandez et al. 2017), the Heatmap view offers interactive features and various sorting functionalities. We also implemented a dynamic hierarchical clustering mechanism to group datasets and subfamilies with similar scores based on cosine similarity using the similarity measurement from the Python library SciPy (Virtanen et al. 2020).

In Figure 2B, the Consensus view displays multiple tracks that represent the enrichment distribution of genomics data and signals across the consensus

sequence of any TE subfamily. A ruler track is on the top, which denotes the nucleotide coordinates of the TE subfamily consensus sequence. The track below the ruler is the genome coverage track, which represents the count of individual TE copies across the genome over each position of the consensus sequence. Below the genome coverage track are the data tracks that display the cumulative counts of reads from the specific dataset overlaying each position along the consensus sequence. Additional tracks from other datasets can be added so that multiple datasets can be compared on specific TE subfamilies. The Consensus view is equipped with interactive features such as zoom in, zoom out, and dragging to facilitate interactive visualization.

Figures 2C and 3B show the Genome view which serves as an interface for a comprehensive overview of the signal of all genomic TE copies of the same TE subfamily using uniquely mapped reads throughout the entire genome in a bird's eye view fashion. We labeled all copies from the same TE subfamily on the chromosome diagrams from the chosen assembly and calculated a score for each copy. The scores are usually a Log Odds Ratio (LOR) of observed signal (read count) on the TE copy over the background or a control dataset. Users can filter TE copies based on the LOR values. Users can also load the selected TE copies into the WashU Epigenome Browser (Li et al. 2019; Li et al. 2022) with the region set view function, where users can choose to view the individual TE copies of interest and display their genomic locations side by side, optionally with customizable upstream and downstream flanking regions, and additional epigenomic data tracks.

We have processed over 7,900 datasets (Supplemental Table S1), including data from the ENCODE, Roadmap Epigenomics, and FANTOM5 projects, and stored them in the cloud server with public access. We also provide a function to visualize the enrichment distribution among all the existing TFs and TEs to help users explore which TFs are most enriched in which TE subfamily. The processed data by individual users can be visualized with the Repeat Browser using the data uploading function. As mentioned above, we classify the TEs into the class-family-subfamily hierarchy. Based on that, we drew a sunburst chart to represent the TE hierarchy through a series of concentric rings, where each ring corresponds to a level in the TE hierarchy. Users can click the different parts on the rings to select the TE subfamilies, TE families, and TE classes. An example of the function to visualize the enrichment distribution and an example of the sunburst chart illustrating TE selection are provided in the Supplemental Tutorial.

Case studies

To demonstrate the utilities of the WashU Repeat Browser, such as verifying predictions or forming new hypotheses, we used ChIP-seq data of transcription factor STAT1 from the human HeLa-S3 cell line (Chuong et al.

2016) and CAGE-seq data of cell lines under DNMTi and HDACi treatment (Brocks et al. 2017) as example case studies.

STAT1 *binds to MER41B elements*

Previous research (Lowe et al. 2007; Wang et al. 2007; Rebollo et al. 2012; Schmidt et al. 2012; Jacques et al. 2013; Sundaram et al. 2014; Notwell et al. 2015) implicates that TEs have the potential to serve as lineage-specific cis-elements with the ability to reconfigure regulatory networks. However, the specific physiological consequences of this process are largely unexplored. One of the current analyses explores how TEs influence the proinflammatory cytokine interferon-γ (IFNG) gene regulatory networks.

Previous research explored the regulatory relationship between the signal transducer and activator of transcription (STAT) and interferon-γ (Schroder et al. 2004; Jorgovanovic et al. 2020; Penrose et al. 2020), and the major STAT protein activated by IFN-γ is STAT1. Several ERV1/MER41 TE subfamilies, especially MER41B, have been previously associated with placental and innate immunity functions, in particular within the IFN-STAT1 pathway (Schmid and Bucher 2010). To explore the interactions between transcription factor (TF) STAT1 and TEs, we used the Repeat Browser to visualize the related ChIP-seq data.

We used the ChIP-seq data (Chuong et al. 2016) from the ENCODE dataset, focusing on three human cell lines treated with IFNG: K562 myeloid-derived cells, HeLa epithelial-derived cells, and primary CD14+ macrophages (Gerstein et al. 2012; Qiao et al. 2013). We selected 27 representative TE subfamilies enriched in the binding sites of IFNG-stimulated cells to visualize the TF-TE relationships in the Repeat Browser. We observed an enrichment of STAT1 ChIP-seq signal over *MER41B* in HeLa cells stimulated with IFNG (Schmid and Bucher 2010), which was displayed as a heatmap using the Heatmap view function (Figure 2A). We observed the peak occurrence over the consensus sequence of the *MER41B* subfamily (Figure 2B), validating the relationship between *MER41B* elements and STAT1 as described by Chuong et al (Schmid and Bucher 2010; Chuong et al. 2016). A genome-wide view of MER41B signals and a sample view of HeLa-S3 on Chr7 are also provided (Figure 2C and 2D). We also utilized the Region Set View feature in the WashU Epigenome Browser to showcase several peaks across multiple regions of interest (Figures 2E and 2F). These visualizations indicate that many *MER41B* elements are likely bound directly by STAT1 in response to IFNG treatment, possibly due to STAT1 binding motifs embedded within their LTR sequences.

Epigenetic treatment induces expression of the LTR12 families

CAGE-seq (Cap Analysis of Gene Expression sequencing) is a high-throughput sequencing-based technique that quantitatively analyzes the 5' end of mRNA molecules, which addresses technical limitations encountered in conventional RNA-seq methodologies, including inadequate coverage of transcript 5′ ends, which is the transcription start site (TSS), and challenges in discerning multiple isoforms and splice variants, frequently overlapping with reference transcripts. Treatments of DNA methyltransferase inhibitors (DNMTi) and histone deacetylase inhibitors (HDACi) result in treatment-induced non-annotated TSSs (TINATs) (Brocks et al. 2017; Goyal et al. 2023; Jang et al. 2024). For example, DAPK1 (death-associated protein kinase 1) is a protein kinase that plays a crucial role in various cellular processes, and epigenetic treatments impact the 5′ region of DAPK1, such that HDACi can activate cryptic TSSs and create alternative promoters (Daskalakis et al. 2018). CAGE-seq data can be used to investigate TINATs on a global scale that is not dependent on reference gene annotation.

Previous studies suggest that more than 80% of TINATs overlapped with TEs (Brocks et al. 2017) and, in particular, the LTR class exhibited a higher rate of association with TINATs than would be expected by chance. Designated as LTR12 (Brind'Amour et al. 2018; Iouranova et al. 2022; Tam and Leung 2023) by Repbase annotation (Bao et al. 2015), these LTRs, notably LTR12C, exhibit a preferential localization around gene promoter regions, demonstrating a strong enrichment of TINATs (Brocks et al. 2017). In one study, LTR12C exhibited the highest enrichment value and was responsible for 50% of all detected TINATs. Given the inherited difficulties in analyzing data derived from TEs, we sought to recapitulate the results from Brocks et al. (Brocks et al. 2017) by using the Repeat Browser.

We imported the CAGE-seq data from Brock et al.'s work (Brocks et al. 2017) into our repeat browser. Data labeled with dimethyl sulfoxide (DMSO) treatment were controls and were compared with data from treatment with DNMTi (DAC), HDACi (SB939), or both DAC and SB939 (DAC+SB). In Figure 3A, focusing on the cells in the LTR12C column, the DAC+SB and DAC treatments showed a strong enrichment level compared to the DMSO treatment. Figure 3B presented the LTR12C genome copies as colored bars along the chromosomes with colors corresponding to the enrichment level. Next, clicking on one chromosome leads to a detailed plot chart. The blue dots represented data for DAC+SB, and the orange dots represented data for DAC across the chromosome. Clicking on a dot will take the users to the WashU Epigenome Browser for a detailed view (Figure 3C).

We selected 5 highly enriched copies of LTR12C and visualized it on the WashU Epigenome Browser with the Region Set View function. We found peaks of H3K4me3 and H3K9ac in the active LTR12C copies (Figure 3D),

which demonstrated epigenetic drugs induced active histone states around genome copies of LTR12C. One of them, highlighted (Figure 3E), corresponded to the dots with three different color tooltips in Figure 3C.

Investigating TF-TE relationships using ENCODE ChIP-seq data

The WashU Repeat Browser offers several tools to facilitate the analysis of genomic data derived from TEs. One key application is the exploration of the TF-TE relationships. Extensive genomics studies (Jordan et al. 2003; Wang et al. 2007; Feschotte and Gilbert 2012; Pehrsson et al. 2019; Du et al. 2024) have substantiated the presence of functional binding sites for specific TFs within specific TEs, underscoring their role in the establishment and rewiring of gene regulatory networks. Given the insights of these studies, we sought to provide a tool to conduct such analyses in a more streamlined fashion.

We selected 45 human ENCODE datasets involving four TFs (EP300, CTCF, SMC3, RAD21) to investigate their binding enrichment over TE subfamilies from all LTR, DNA, SINE, and LINE classes. We generated a heatmap to visualize the enrichment scores for each pair of TF and TE subfamilies using the Repeat Browser, and we displayed the top 100 TE subfamilies in Figure 4A. To validate the relationships between TFs and TEs, we referred to Sundaram et. al.'s comprehensive analysis of the TF-TE relationships (Sundaram et al. 2014), which identified pairwise relationships between specific TE subfamilies and specific TFs. We explored the relationship between TE enrichment level and TE-derived binding peaks (Figure 4B). We chose four representative TFs (CTCF, RAD21, SMC3, and EP300) and compared the TE data identified in panel A of the first figure by Sundaram et al (Sundaram et al. 2014). For each TF, we initially sorted our TE data by enrichment level and established five distinct cutoffs. Consequently, our TE data were divided into five groups, each comprising only the data with enrichment levels exceeding the respective cutoffs. We then checked if our TE data belonged to the identified TF-TE relationships and calculated the precision for each of the five groups. For each group, the precision is the fraction of the number of our TE data belonging to the identified TF-TE relationships among the overall number of our TE data. Our analysis revealed that groups with higher cutoffs exhibited higher precision, indicating that highly enriched TEs are more likely to have TE-derived binding peaks mediated by TFs, particularly CTCF. The TE data from four TFs all present a varying degree of positive correlation ($R > 0$) with Sundaram et al.'s work. The distribution of the enrichment scores for the top 100 TEs in the heatmap is plotted in Figure 4C, with detailed information provided in Supplemental Table S2. The combination of TE subfamilies in Figure 1A of Sundaram et al. is displayed in a pie chart (Figure 4D).

From the heatmap, it is evident that LTR subfamilies significantly outnumbered *LINE*, *SINE*, and DNA elements in these relationships. Recognized for mediating long-range interactions in the genome (Merkenschlager and Odom 2013), CTCF and the cohesion complex (RAD21 and SMC3) were noted for their significant involvement with LTR subfamilies, which contributed many regulatory elements. The CTCF, SMC3, and RAD21 were identified to play more important roles in 3D chromatin structure than other TFs (including EP300), which may account for the much lower correlation of EP300 when analyzed alongside these three TFs (Zhang et al. 2018). The combination of these two observations substantiated the idea that CTCF, RAD21, and SMC3 might have taken advantage of the LTR elements in evolving their target network.

Discussion

We developed the WashU Repeat Browser (https://repeatbrowser.org/) which was designed for researchers to study TE-derived cell type-specific regulatory elements and investigate the contributions of TEs to gene regulation in any cell type or tissue. Using this resource, rich functional information about TEs can be explored, allowing investigators to annotate the chromatin states of TEs and perform their analysis, and compare their data to publicly available data, which expands the access and availability of public genomic data for uncovering the role of TEs in biology and human health.

Additionally, users can explore the epigenomic profiles of TEs across its consensus sequence, rather than solely on the linear reference genome. WashU Repeat Browser offers access to over 7,900 data sourced from various data portals, alongside more than 1,300 distinct TE subfamilies in both human and mouse genomes. We also provide a ranking function to sort the most enriched TFs on TEs or TEs on TFs based on pre-computed results. Multiple visualization components are provided: the Heatmap View provides a graphical representation of the enrichment levels of selected TEs and TFs, equipped with interactive features such as clustering, sorting, and screenshots; the Consensus View presents multiple tracks illustrating the distribution of genomics data enrichment across selected TE consensus sequences. Additionally, the Genome View displays enrichment levels across linear genome copies, allowing users to select multiple data for comparative analysis. Utilizing the filter function, users can identify genome loci where the genome copies exhibit the highest enrichment levels. Integration with the WashU Epigenome Browser facilitates detailed exploration of specific genomic regions, enhancing the depth of analysis and interpretation.

UCSC Repeat Browser and WashU Repeat Browser

The WashU Repeat Browser is conceptually tightly related to the UCSC Repeat Browser (Fernandes et al. 2020) and provides more features, functions, and data. The UCSC Repeat Browser lifts the sequence from a repeat region annotated by RepeatMasker from the human genome assembly to the Dfam (Hubley et al. 2016) consensus of that repeat family and uses the consensus as a scaffold to display data. Thus, the UCSC Repeat Browser provides a similar function to the Consensus View of the WashU Repeat Browser and can be used to investigate one TE subfamily at a time. At this point, the UCSC Repeat Browser does not support non-human genomes and contains only limited pre-processed datasets, and users can also visualize their own data following its tutorial. A comparison of features for the two Repeat Browsers is listed in Supplemental Table S4.

WashU Repeat Browser provides more visualization options including Heatmap, Consensus View, and Genome View, and is linked with the WashU Epigenome Browser (Li et al. 2022) to visualize profiles of each TE copy in the genome. Multiple TE subfamilies and datasets can be loaded at the same time via Heatmap View to explore the enrichment landscape across different datasets. Over 7,900 pre-processed datasets from public consortiums including ENCODE, Roadmap, and FANTOM5 (Kawaji et al. 2017) provide a rich resource for the exploration and comparison of these data. WashU Repeat Browser supports both human and mouse assemblies. In addition, the WashU Repeat Browser provides a mechanism to allow users to use their annotation and custom consensus sequences, making it more flexible to study TE genomics.

Comparison with other TE analysis tools

Many tools designed for processing high-throughput sequencing reads derived from TEs have been developed over the past decade, each with its focus and features. Supplemental Table S5 lists several related tools, including their features. In this study, we focused on developing visualization technologies, and we chose to use iteres for data processing. It is possible to use other tools and format their output data for visualization on the WashU Repeat Browser.

Future directions for this work include the incorporation of additional data types and enhancements to the user interface (UI) design of the WashU Repeat Browser. As RNA-seq data continues to gain prominence in research, one key objective is to develop and integrate data processing pipelines for RNA-seq data into our system. Moreover, we plan to integrate the track design features of the WashU Epigenome Browser into the consensus view to enhance its functionality. Additionally, we will explore improved methods for visualizing the relationship between TE consensus sequences and linear references to achieve more effective data representation.

Methods

Data Processing

We download the raw reads of epigenomics data for human and mouse, including ChIP-seq, DNase-seq, ATAC-seq, and CAGE-seq data, from ENCODE and Roadmap through the ENCODE data portal (The ENCODE Project Consortium 2012). All reads, except for CAGE-seq reads, are mapped to the human (hg38 and hg19) and mouse (mm10) genomes using the BWA aligner (version 0.7.17) with the default setting for uniform processing. Mapping to both hg38 and hg19 allows users to access and utilize both assemblies, providing flexibility to view data in the context of either hg38 or hg19 as needed. The CAGE-seq reads are processed using the STAR aligner (version 2.5.3a) (Dobin et al. 2013). The alignment result files generated by aligner tools are then processed with iteres (Xie et al. 2013; Sundaram et al. 2014) to calculate Reads Per Kilobase per Million mapped reads (RPKM) scores for repeats at each subfamily, family, and class level. Enrichment is then calculated using the RPKM score of the subfamily against the corresponding input/control samples if there was input, like ChIP-seq data, or the average of RPKM of all subfamilies if there was no input, like DNase-seq and ATAC-seq data. In the analysis of CAGE-seq data, mapping locations of TSS are used rather than the entire reads. During the procedure, users have the flexibility to specify the length of the TSS flanking region during data processing. We then compute the RPKM score for the subfamily against the average RPKM of all subfamilies at the defined TSS. The RPKM scores of subfamilies are used for further analysis and visualization.

To ensure efficient storage and accessibility, the processed data are stored in a cloud server at WashU in the Zarr format. Zarr is a format for storing and organizing large, multidimensional arrays of numerical data. Each Zarr file comprises statistical information on each repeat subfamily, including the distribution of reads across the consensus sequences of each subfamily, as well as the genomic location of each segment within the subfamily. Additionally, the Zarr file incorporates metadata such as sample and assay information.

Enrichment Calculation

The computation of enrichment of any TE subfamily in a particular assay varies. The enrichment score is quantified through the RPKM and logarithmic RPKM, as outlined below:

- RPKM:

$$\text{RPKM} = \frac{\text{Number of reads that map to a TE consensus} * 10^9}{\text{Length of the TE consensus sequence}}$$

Note that the quantification of reads is based on the count of mapped reads (unique or all mapped), while the measurement of the TE consensus sequence length is expressed in terms of base pairs.

- Log odds ratio calculation of ChIP-seq data:

$$LOR_{i,j} = \log_2\left(\frac{RPKM_{i,j} \text{ of data signal}}{RPKM_{i,j} \text{ of data control}}\right)$$

- where $LOR_{i,j}$ of ChIP-seq data stands for the ratio between the RPKM values of signal and control through the experiment dataset i and TE subfamily j. $RPKM_{i,j}$ is the RKPM of the reads from ChIP-seq dataset i that mapped to TE subfamily j.
- Log odds ratio calculation of DNase-seq and ATAC-seq:

$$LOR_{i,j} = \log_2\left(\frac{RPKM_{i,j}}{\text{The average } RPKM_i \text{ of all TE subfamilies}}\right)$$

where $LOR_{i,j}$ of DNase-seq and ATAC-seq stands for log-ratio between the $RPKM_{i,j}$ and the average RKPM of reads from dataset i that are mapped to all TE subfamilies. $RPKM_{i,j}$ is the RKPM of the reads from dataset i that mapped to TE subfamily j.

- RPKM and Log odds ratio calculation of CAGE-seq:

$$RPKM_{CAGE} = \frac{\text{Number of reads in TSS that map to a TE consensus} * 10^9}{\text{Length of the TE consensus sequence}}$$

$$LOR_{i,j} = \log_2\left(\frac{RPKM_{CAGE\,i,j}}{\text{The average } RPKM_{CAGE_i} \text{ of all TE subfamilies}}\right)$$

The computational procedure for CAGE-seq is similar to that of DNase-seq and ATAC-seq. However, a distinct preprocessing step is employed wherein the input reads are filtered to retain only in the TSS. Also, the length of TSS can be specified during data processing.

Web interface design

The WashU Repeat Browser uses Python for the backend engine, which supports running fast clustering and sorting algorithms for the Heatmap view. We use HTML5 and JavaScript for frontend web design, incorporating

libraries such as JQuery (https://jquery.com/), Plotly (https://plotly.com/javascript/), and D3.js (https://d3js.org/). The frontend code is based on Svelte (https://svelte.dev), which is a JavaScript framework used for building user interfaces. Graphical rendering is accomplished using HTML5 Canvas and inline SVG (Scalable Vector Graphics). Additionally, visualization on the WashU Repeat Browser can be exported to a JSON session file. This file contains all the information necessary for reproducing the data visualization. Users can easily save and restore their visualization results using this function. The documentation of how to use the web interface can be found in the supplemental file (Supplemental Tutorial) and the documentation site with a tutorial video.

Data access

The data processed by our pipeline is freely available and stored in a Washington University in St. Louis hosted Simple Storage Service (S3) bucket. The accessible URLs for these data are provided in Supplemental Table S1, which can be accessed via direct web access or command line tools.

Software availability

The code for the Repeat Browser website can be found on GitHub (https://github.com/twlab/Repeat-Browser) and the Supplemental Code S1 file. The code for the data processing pipeline can be found on GitHub (https://github.com/twlab/Repeat-Browser_data_processing) and the Supplemental Code S2 file. The tutorial for the Repeat Browser is located at https://rb-doc.readthedocs.io/.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank lab members from Wang lab for testing and feedback on the Browser. This work was partially supported by National Institutes of Health grant numbers R01HG007175, U01CA200060, U01HG009391, U41HG010972, U24HG012070, UM1HG011585, and UM1MH130994.

Author Contributions : All authors helped test the WashU Repeat Browser. J.S., S.C., and D.L. developed the data processing pipelines. J.S. and S.C. analyzed the data. J.S., D.P., and W.Z. coded the web-based visualization platform. X.Z. and A.D. conducted the case studies and related analysis. D.L. and T.W. supervised and coordinated the study. J.S., D.L., and T.W. wrote the manuscript.

Figures

Figure 1. The overall design of the WashU Repeat Browser. (A) Genomics data from public consortiums are downloaded and processed through our in-house pipeline and stored in the cloud server for remote access. In this pipeline, our algorithm iteres uses two steps to process the data (details in panels B and C). (B) In step 1, we treat reads from CAGE-seq differently from the reads from other assays, we extract the transcription start site (TSS) locations from CAGE-seq data and mapping locations of reads from other assays for the next step. (C) In step 2, based on locations from the last step, there are three scenarios when assigning reads to TEs: (1) Uniquely mapped reads are directly allocated to the mapped TE subfamilies. (2) Multi-mapped reads, exhibiting different mapping qualities across distinct TE subfamilies, are assigned to the TE subfamily with better mapping quality. (3) Multi-mapped reads that cannot be assigned to a subfamily with a higher mapping quality are excluded. Reads will be ultimately anchored to the consensus sequences of the corresponding repeat subfamily. (D) Repeat Browser web interface provides two panels for choosing the repeat subfamilies and datasets. After selecting the repeat subfamilies and datasets, statistics of selected datasets over chosen repeat subfamilies can be visualized via the Heatmap View, Consensus View, and Genome View. Profiles of individual repeat copies in the genome can be visualized in the linked WashU Epigenome Browser platform.

Figure 2. Overview of analysis of *MER41B* in HeLa-S3 cell line bound by STAT1. (A) The highlighted cell of Heatmap View indicates strong enrichment of STAT1 ChIP-seq reads in the TE subfamily *MER41B* in HeLa-S3 cells. The heatmap can be updated by sorting the metadata. (B) A Consensus View on *MER41B*, read density plots of two STAT1 ChIP-seq replicates are displayed along with two control tracks on the *MER41B* consensus sequence. Signals from both unique reads and multi-reads are displayed with colors brown and blue, respectively. The first track contains a ruler bar and the color-coded nucleotide sequence of the *MER41B* consensus sequence. The genome coverage track represents the count of individual *MER41B* copies across the whole genome over each position of the consensus sequence. The four tracks below are two signal tracks along with two control tracks. (C) Genome View of *MER41B*. It visually represents the distribution and enrichment RPKM (Reads Per Kilobase Million) of the *MER41B* repeat across different human chromosomes where copies with different values are displayed in different colors, with high enrichment values shown in red. (D) A lollipop plot displays the enrichment scores of TE copies across different genome locations; each dot is a genome copy on the selected chromosome, the dot size depends on its RPKM. The x-axis represents the chromosome loci, and the y-axis represents the enrichment level of the copy. Below is an ideogram, with the red bar indicating the genomic region of the plot. Clicking the dots will navigate the user to the WashU Epigenome Browser for a more detailed view. (E) To utilize the WashU Epigenome Browser, Region Set View in the WashU Epigenome Browser displays multiple regions of interest (in this case, copies of a TE subfamily) side-by-side. We selected 11 highly enriched copies of *MER41B* and visualized them on the WashU Epigenome Browser with the Region Set View function. 11 *MER41B* copies (green) and the 5kb flanking regions up (blue) and downstream (yellow) of each copy are displayed. Tracks included are H3K27ac,

H3K4me1, two replicates of STAT1 ChIP-seq, input reads, and RefSeq gene annotation. We observed the co-enrichment of H3K4me1 and H3K27ac over these *MER41B* elements using the HeLa-S3 cell line data. (F) One of the columns (indicated in the red dotted box in Figure 4E) was explored with the WashU Epigenome Browser to see the detailed distribution on that *MER41B* copy. In this zoom-in view, we can observe the peaks in the two STAT1 tracks corresponding to the loci of the *MER41B* genome copy.

Figure 3. Visualization of LTR12C using CAGE-seq data following treatment of epigenetic drugs. (A) The highlighted cell in Heatmap View indicates strong enrichment of CAGE-seq TSS in the TE subfamily LTR12C in cells treated with DAC+SB. (B) Genome View of LTR12C copies in cells treated with DAC+SB. (C) The single-end (SE) data of LTR12C copies in cells treated with DAC+SB, DMSO, and DAC were used to calculate individual copy enrichment scores, which are displayed in the lollipop plots. (D) Region Set View in the WashU Epigenome Browser displays multiple regions of interest (in this case, copies of the LTR12 subfamily) side-by-side. Five LTR12C copies (dark green) and the 5kb flanking regions upstream (blue) and downstream (yellow) of each copy are displayed. The nine tracks included are three CAGE-seq data and six ChIP-seq data of H3K9ac and H3K4me3 that follow the treatment of DMSO, DAC, and DAC+SB individually. Additionally, the RefSeq gene annotation (MANE Selection v1.0) is provided at the top. (E) Zoomed-in Genome Browser view of one example LTR12C element highlighted in (D) with a red dashed box.

Figure 4. Comprehensive visualization of associations between four TFs (CTCF, SMC3, RAD21, EP300) and TEs. (A) A heatmap depicting the RPKM ratios of signal over control data of four TFs from ENCODE. Each column represents a set of RPKM ratios within a TE subfamily. The ratios were computed for the data of CTCF, SMC3, RAD21, and EP300 in TEs from families such as *ERV1*, *ERVL*, *ERVK*, *L1*, *L2*, *hAT-Tip100*, *TcMAR-Tigger*, *hAT-Charlie*, and *Alu*. We identified 27 TE subfamilies which were most statistically significantly associated with the 4 TFs across 31 datasets. The comprehensive heatmap and detailed table encompassing all the data from the 4 TFs and all TEs were presented in Supplemental Figure S2 and Supplemental Table S3. (B) Precision metrics for utilizing ratio values to predict the correlation between data enrichment levels and the presence of binding peaks. The x-axis represents the cut-off RPKM ratio of the input data. The y-axis refers to the precision of datasets containing binding peaks. (C) Violin graph illustrating the distribution of enrichment scores in the heatmaps. (D) Pie chart presenting the TE subfamilies featured in the heatmap. The inner circle represents the TE class, while the outer ring represents the TE family.

Reference

Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, Yang HT, Xue V, Knyazev S, Singer BD et al. 2021. Technology dictates algorithms: recent developments in read alignment. *Genome Biology* **22**.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**: 11.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS et al. 2018. Ten things you should know about transposable elements. *Genome Biology* **19**.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research* **18**: 1752-1762.

Brind'Amour J, Kobayashi H, Richard Albert J, Shirane K, Sakashita A, Kamio A, Bogutz A, Koike T, Karimi MM, Lefebvre L et al. 2018. LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. *Nature Communications* **9**.

Brocks D, Schmidt CR, Daskalakis M, Jang HS, Shah NM, Li D, Li J, Zhang B, Hou Y, Laudato S. 2017. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nature genetics* **49**: 1052-1060.

Burkes-Patton S, Cooper EA, Schlueter J. 2023. RepBox: a toolbox for the identification of repetitive elements. *BMC Bioinformatics* **24**.

Carey KM, Patterson G, Wheeler TJ. 2021. Transposable element subfamily annotation has a reproducibility problem. *Mobile DNA* **12**.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**: 1083-1087.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**: 71-86.

Chuong EB, Rumi MK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature genetics* **45**: 325-329.

Daskalakis M, Brocks D, Sheng Y-H, Islam MS, Ressnerova A, Assenov Y, Milde T, Oehme I, Witt O, Goyal A. 2018. Reactivation of endogenous retroviral elements via treatment with DNMT-and HDAC-inhibitors. *Cell Cycle* **17**: 811-822.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dominguez AA, Lim WA, Qi LS. 2016. Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nature reviews Molecular cell biology* **17**: 5-15.

Du AY, Chobirko JD, Zhuo X, Feschotte C, Wang T. 2024. Regulatory transposable elements in the encyclopedia of DNA elements. *Nature Communications* **15**.

Fernandes JD, Zamudio-Hurtado A, Clawson H, Kent WJ, Haussler D, Salama SR, Haeussler M. 2020. The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mobile DNA* **11**: 1-12.

Fernandez NF, Gundersen GW, Rahman A, Grimes ML, Rikova K, Hornbeck P, Ma'ayan A. 2017. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data* **4**: 1-12.

Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics* **13**: 283-296.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**: 331-368.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**: 91-100.

Goyal A, Bauer J, Hey J, Papageorgiou DN, Stepanova E, Daskalakis M, Scheid J, Dubbelaar M, Klimovich B, Schwarz D et al. 2023. DNMT and HDAC

inhibition induces immunogenic neoantigens from human endogenous retroviral element-derived transcripts. *Nature Communications* **14**.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**: D81-D89.

Iouranova A, Grun D, Rossy T, Duc J, Coudray A, Imbeault M, De Tribolet-Hardy J, Turelli P, Persat A, Trono D. 2022. KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. *Mobile DNA* **13**.

Jacques P-É, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS genetics* **9**: e1003504.

Jang HJ, Shah NM, Maeng JH, Liang Y, Basri NL, Ge J, Qu X, Mahlokozera T, Tzeng S-C, Williams RB et al. 2024. Epigenetic therapy potentiates transposable element transcription to create tumor-enriched antigens in glioblastoma cells. *Nature Genetics* doi:10.1038/s41588-024-01880-x.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *TRENDS in Genetics* **19**: 68-72.

Jorgovanovic D, Song M, Wang L, Zhang Y. 2020. Roles of IFN-γ in tumor progression and regression: a review. *Biomarker Research* **8**.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics* **9**: 411-412.

Karolchik D, Hinrichs AS, Kent WJ. 2009. The UCSC Genome Browser. *Current Protocols in Bioinformatics* **28**.

Kawaji H, Kasukawa T, Forrest A, Carninci P, Hayashizaki Y. 2017. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Scientific Data* **4**.

Kazazian Jr HH. 2004. Mobile elements: drivers of genome evolution. *science* **303**: 1626-1632.

Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**: 1-24.

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics* **42**: 631-634.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357-359.

Lerat E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**: 520-533.

Li D, Hsu S, Purushotham D, Sears RL, Wang T. 2019. WashU epigenome browser update 2019. *Nucleic acids research* **47**: W158-W165.

Li D, Purushotham D, Harrison JK, Hsu S, Zhuo X, Fan C, Liu S, Xu V, Chen S, Xu J. 2022. WashU epigenome browser update 2022. *Nucleic acids research* **50**: W774-W781.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**: 1754-1760.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**: 1851-1858.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the national academy of sciences* **104**: 8005-8010.

Merkenschlager M, Odom DT. 2013. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**: 1285-1297.

Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. 2020. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biology* **21**.

Notwell JH, Chung T, Heavner W, Bejerano G. 2015. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nature communications* **6**: 6644.

Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. 2019. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nature Communications* **10**.

Penrose HM, Katsurada A, Miyata K, Urushihara M, Satou R. 2020. STAT1 regulates interferon-γ-induced angiotensinogen and MCP-1 expression in a bidirectional manner in primary cultured mesangial cells. *Journal of the Renin-Angiotensin-Aldosterone System* **21**: 147032032094652.

Qiao Y, Giannopoulou EG, Chan CH, Park S-h, Gong S, Chen J, Hu X, Elemento O, Ivashkiv LB. 2013. Synergistic activation of inflammatory cytokine genes by interferon-γ-induced chromatin remodeling and toll-like receptor signaling. *Immunity* **39**: 454-469.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual review of genetics* **46**: 21-42.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.

Schmid CD, Bucher P. 2010. MER41 Repeat Sequences Contain Inducible STAT1 Binding Sites. *PLoS ONE* **5**: e11425.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335-348.

Schroder K, Hertzog PJ, Ravasi T, Hume DA. 2004. Interferon-γ: an overview of signals, mechanisms and functions. *Journal of Leucocyte Biology* **75**: 163-189.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome research* **24**: 1963-1976.

Sundaram V, Wang T. 2018. Transposable Element Mediated Innovation in Gene Regulatory Landscapes of Cells: Re-Visiting the “Gene-Battery” Model. *BioEssays* **40**: 1700155.

Tam PLF, Leung D. 2023. The Molecular Impacts of Retrotransposons in Development and Diseases. *International Journal of Molecular Sciences* **24**: 16418.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Trizzino M, Kapusta A, Brown CD. 2018. Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC genomics* **19**: 1-12.

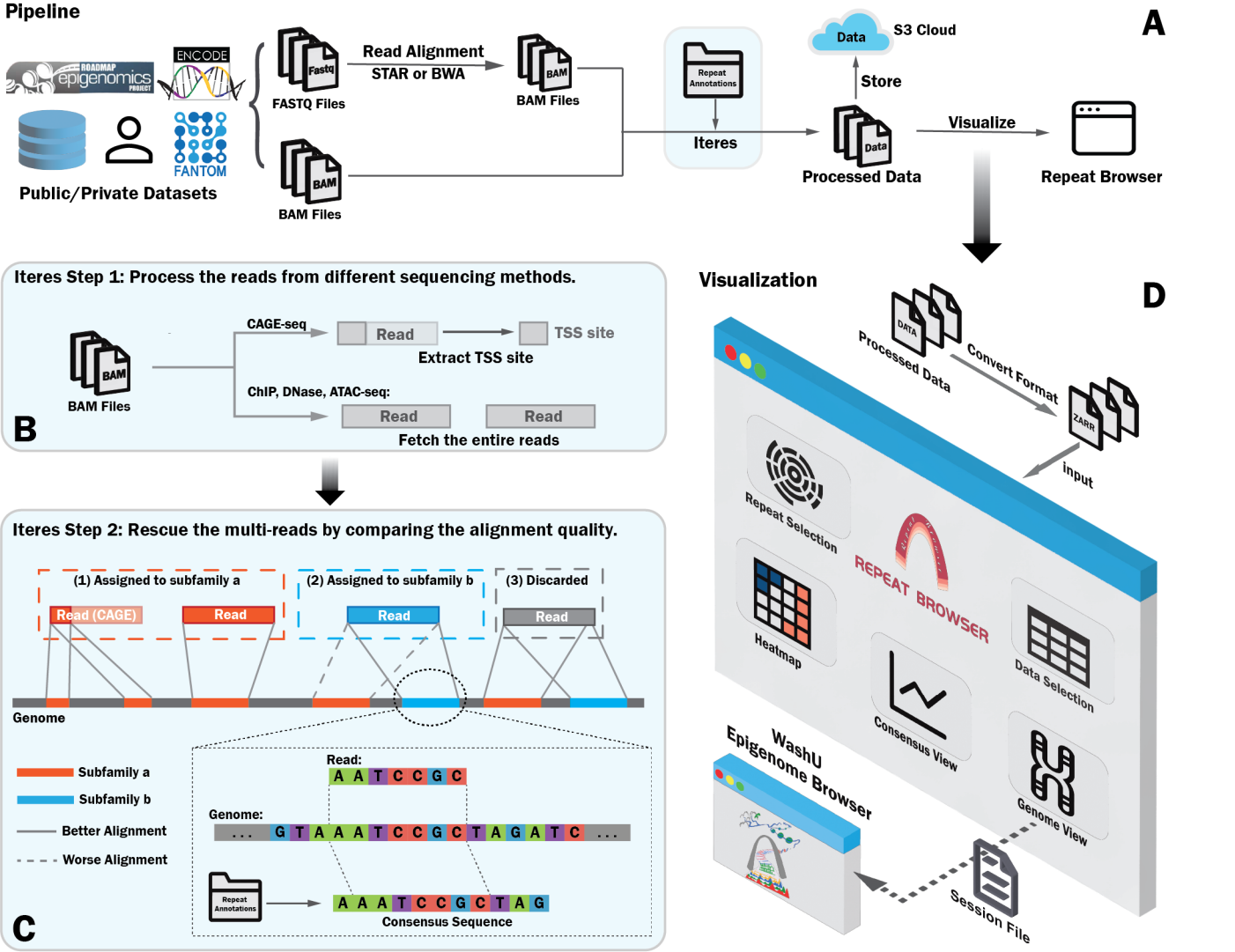
Virtanen P Gommers R Oliphant TE Haberland M Reddy T Cournapeau D Burovski E Peterson P Weckesser W Bright J et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**: 261-272.

Wang J, Vicente-García C, Seruggia D, Moltó E, Fernandez-Miñán A, Neto A, Lee E, Gómez-Skarmeta JL, Montoliu L, Lunyak VV. 2015. MIR retrotransposon sequences provide insulators to the human genome. *Proceedings of the National Academy of Sciences* **112**: E4428-E4437.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences* **104**: 18613-18618.

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature genetics* **45**: 836-841.

Zhang L, Xue G, Liu J, Li Q, Wang Y. 2018. Revealing transcription factor and histone modification co-localization and dynamics across cell lines by integrating ChIP-seq and RNA-seq data. *BMC Genomics* **19**.


Pipeline
A
ROADMAP epigenomics PROJECT
ENCODE
FANTOM
Public/Private Datasets
Fastq
FASTQ Files
Read Alignment
STAR or BWA
BAM
BAM Files
BAM
BAM Files
Repeat Annotations
Iteres
Data
S3 Cloud
Store
Data
Processed Data
Visualize
Repeat Browser
Iteres Step 1: Process the reads from different sequencing methods.
BAM
BAM Files
CAGE-seq
Read
TSS site
Extract TSS site
ChIP, DNase, ATAC-seq:
Read
Read
Fetch the entire reads
B
Iteres Step 2: Rescue the multi-reads by comparing the alignment quality.
(1) Assigned to subfamily a
(2) Assigned to subfamily b
(3) Discarded
Read (CAGE)
Read
Read
Read
Genome
Subfamily a
Subfamily b
Better Alignment
Worse Alignment
Read:
A A T C C G C
Genome:
. . . G T A A A T C C G C T A G A T C . . .
Repeat Annotations
A A A T C C G C T A G
Consensus Sequence
C
Visualization
D
DATA
Processed Data
Convert Format
ZARR
input
Repeat Selection
REPEAT BROWSER
Heatmap
Data Selection
Consensus View
WashU Epigenome Browser
Genome View
Session File