PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: O'Neill K, Brocks D, Hammell MG. 2019 Mobile genomics: tools and techniques for tackling transposons. *Phil. Trans. R. Soc. B* **375**: 20190345.
http://dx.doi.org/10.1098/rstb.2019.0345

Accepted: 20 November 2019

One contribution of 15 to a discussion meeting issue 'Crossroads between transposons and gene regulation'.

Subject Areas:
genomics, computational biology, bioinformatics

Keywords:
transposable elements, computational genomics, retrotransposons, single-cell analysis

Author for correspondence:
Molly Gale Hammell
e-mail: mhammell@cshl.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4796280.

THE ROYAL SOCIETY
PUBLISHING

Mobile genomics: tools and techniques for tackling transposons

Kathryn O'Neill[1], David Brocks[2] and Molly Gale Hammell[1]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2]Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot, Israel

MGH, 0000-0003-0405-8392

Next-generation sequencing approaches have fundamentally changed the types of questions that can be asked about gene function and regulation. With the goal of approaching truly genome-wide quantifications of all the interaction partners and downstream effects of particular genes, these quantitative assays have allowed for an unprecedented level of detail in exploring biological interactions. However, many challenges remain in our ability to accurately describe and quantify the interactions that take place in those hard to reach and extremely repetitive regions of our genome comprised mostly of transposable elements (TEs). Tools dedicated to TE-derived sequences have lagged behind, making the inclusion of these sequences in genome-wide analyses difficult. Recent improvements, both computational and experimental, allow for the better inclusion of TE sequences in genomic assays and a renewed appreciation for the importance of TE biology. This review will discuss the recent improvements that have been made in the computational analysis of TE-derived sequences as well as the areas where such analysis still proves difficult.

This article is part of a discussion meeting issue 'Crossroads between transposons and gene regulation'.

1. Introduction

While several types of genomic repeated sequences exist, the largest fraction of the human genome, approximately half, is comprised of transposable elements (TEs) [1], though some groups estimate much larger TE fractions [2]. These TEs, often called transposons or jumping genes, are DNA sequences that have, or once had, the ability to mobilize within the genome, either directly or through an RNA intermediate. TEs are present, to varying degrees, in the genomes of all known types of organisms, both prokaryotic and eukaryotic, with some species showing more genomic transposons than host sequences [3]. Several excellent reviews have discussed the many and varied types of TEs [4–6]. Briefly, TEs come in two major types. Class I TEs, also called retrotransposons, first transcribe an RNA copy that is then reverse transcribed to cDNA before inserting elsewhere in the genome. Class II TEs, also called DNA transposons, directly excise themselves from one location before reinsertion. In the human genome, the vast majority of TEs are of Class I, retrotransposon type. Nearly, all human TEs have lost the ability to fully mobilize [7–9], with the human-specific LINE-1 element (L1HS) being the only fully autonomous TE with the ability to generate new transposition events to date. However, most TEs have retained some level of functionality, including the ability to direct their own transcription. Thus, transcriptome-wide sequencing assays, like RNA-seq, frequently include transposon-derived transcripts among the set of expressed sequences. Moreover, some transposon transcripts have been co-opted to play a role in host function, particularly during early development, such that some expressed transposon transcripts have been shown to be necessary for proper cell differentiation and maintenance of identity [10–14]. In addition to their roles in

© 2019 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

general cellular function, several types of transposons have become intricately entangled within gene regulatory networks [15], contributing both to *cis*-regulatory sequences [16–18] as well as general chromatin environments [19–21]. For this reason, it is paramount, we consider the contribution of repetitive elements as we unravel the genomic and epigenomic landscapes that control gene expression.

Properly accounting for repetitive regions in most genomics analysis settings requires special considerations for the challenges presented by the number of nearly identical transposon sequences dispersed throughout our genomes. Thus, reads derived from these regions are frequently discarded in most sequencing data analysis protocols owing to the difficulty in properly assigning TE-derived reads to the correct locus of origin. Few packages explicitly support inclusion of repeats and some intentionally discard reads from these regions, as discussed in a recent review [22]. Of the packages designed to address TEs, many tools focus on the detection of novel TE insertions or TE-associated genomic rearrangements. Few tools are developed specifically to address regulatory and transcriptional activity of TEs in common assays, such as RNA-seq, chromatin immunoprecipitation sequencing (ChIP-seq), cross-linking immunoprecipitation sequencing (CLIP-seq) and small RNA-seq (sRNA-seq). In this review, we seek to provide an overview of the packages that explicitly support the inclusion of TE sequences in differential expression and binding analyses, and the strides which have been made to improve our ability to resolve ambiguously mapped reads in genomics analysis.

2. Annotation and de novo detection

A well-assembled and annotated genome is the foundation for effective analysis, as all subsequent analyses discussed below require a reference genome as well as a map of gene and TE positions. While many genomes have near-complete assemblies, and extensive annotation, the quality of both tends to drop over repeat-rich regions for the same reasons discussed above: ambiguity in placing near-identical sequence reads from highly similar copies of related transposons. This ambiguity leads to non-contiguous and erroneous chromosomal assembly, which will feed forward into any genomics analyses using these assemblies [23]. Genome assembly has benefitted immensely from long-read sequencing technologies, particularly in the context of highly repetitive centromeric regions and in nested repeating elements [24,25]. While these long-read technologies are improving the reference genomes used to map new datasets, one caveat is that transposons are often polymorphic within populations, such that each new sample sequenced is expected to have many non-reference transposon-associated insertions, deletions and other structural variants that may be rare or private [26,27].

Once a high-quality assembly is constructed, the process of annotation may begin. Many curated annotation databases have been developed for identifying repeat elements. For an in-depth review of annotation practices and existing repositories, please refer to the review by Goerner-Potvin *et al.* [22]. Here, the distinction between TE-, genome- and polymorphism-focused annotation repositories is emphasized in addition to a list of software for de novo insertion detection. The most widely used database of TE consensus sequences is RepBase [28], which provides the sequences with which genome-specific annotation files are constructed. These annotation files are available through the University of California Santa Cruz Genome Browser (UCSC) and RepeatMasker [29]. While new RepBase consensus sequences require a subscription, several open databases for repeat annotation are available in addition to UCSC including: RepetDB [30], ERVdb [31], Dfam [32], TREP [33], SPTEdb [34], ConTEdb [35] and mips-REdat [36]. The ideal database for analysis will vary depending on the model organism and TEs of interest, as some databases are species and TE type-specific.

3. Mapping

After the construction of a well-annotated reference genome, one is faced with the task of mapping experimental data to the appropriate reference. Even with a perfectly annotated and constructed genome, ambiguously mapped sequencing reads still present a challenging problem. One of the first approaches to address this problem, designed for RNA-seq analysis, was to probabilistically assign multi-mapped reads to regions that also show a higher density of uniquely mapped reads, i.e. reads with a single best genomic alignment under the mapping software's heuristics [37]. However, this was a highly gene-centric model that was primarily focused on host gene expression, and was not explicitly intended for estimating expression from TE loci. Moreover, this approach is biased towards regions that have some uniquely mappable content. Unfortunately, the most recently integrated TE insertions are also the least likely to be uniquely mappable, and are thus the most likely to be lost or underestimated by these methods. To highlight this, figure 1 displays the estimated mappability of several different types of TEs in the human genome, with a specific emphasis on younger types of TEs shown to be active in the human genome [38]. Mappability in this plot was defined as the inverse of the number of times a simulated 76 bp paired-end read mapped to the genome, allowing three mismatches. Mappability was scored per nucleotide with the score assigned to the first nucleotide of the read. This track was procured from an in-depth analysis performed by Sexton and Han which considers the many parameters that contribute to the mappability of a particular sequence, including the mapping software chosen and the length of the sequenced read [39]. These analyses still return to the same basic theme displayed in figure 1: mappability rates vary for different types of transposons, and the most recently inserted transposons are the most likely to be discarded by standard analyses that rely on uniquely mapped reads. In other words, the transposons that present the most problems in genomics analyses are precisely those that are more likely to be functional in terms of: carrying fully functional promoters, encoding for functional proteins, and, rarely, mobilizing within the genome. In addition, many older elements with degraded versions of these components have been recycled to play roles in *cis*-regulatory architecture [40].

Most genome alignment software is aware of the difficulties posed by ambiguously mapped reads, and thus provide extensive parameter sets designed to allow the user to choose the number of alignments considered for each sequenced read. This includes standard genome mapping software applicable to genome resequencing studies as well as ChIP-seq-based studies of protein-DNA binding, such as BWA

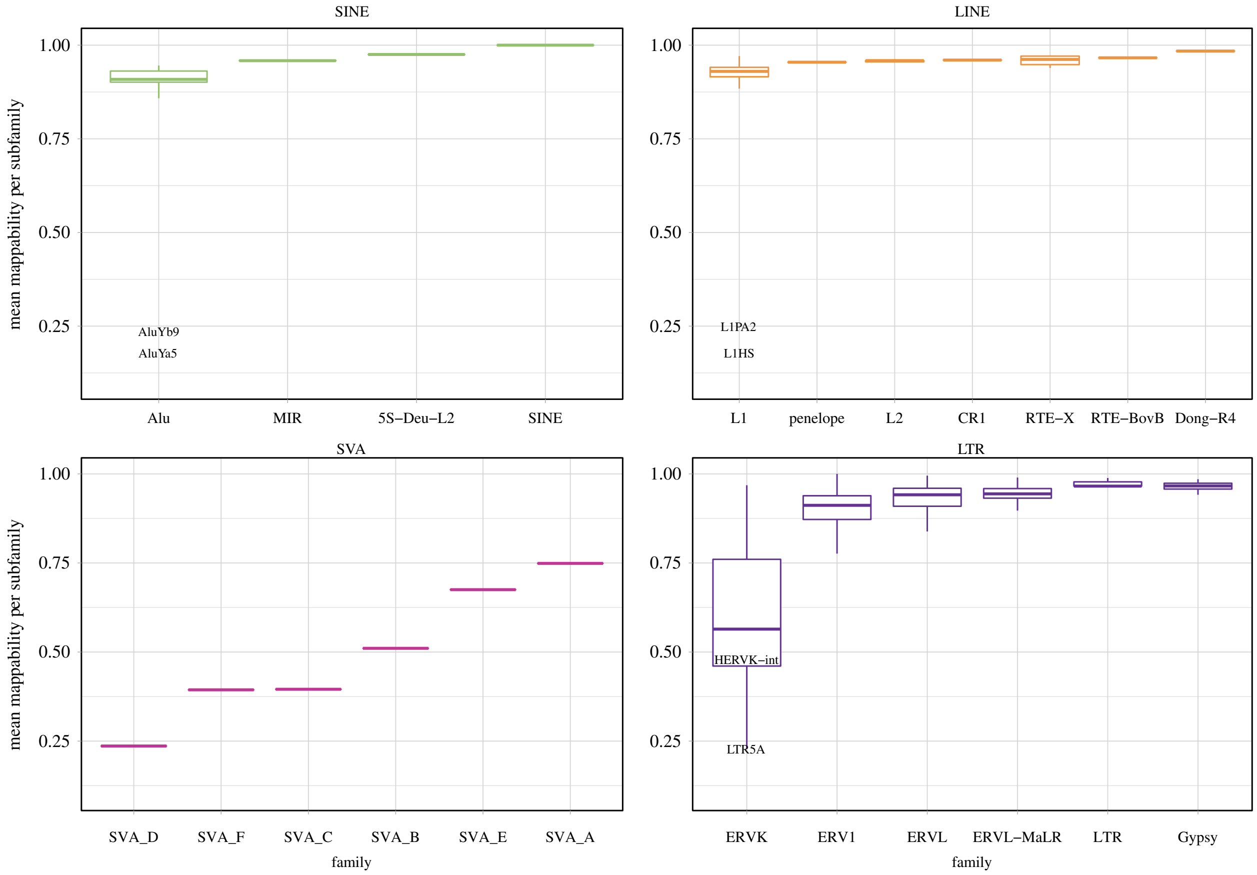

Figure 1. Estimated mean mappability for different types of TEs in the human genome. Mappability tracks from the analysis by Sexton and Han for hg38 were used to construct mean mappability estimates (average probability that a pair of 76 bp reads would map uniquely to a genomic instance of that TE). These were then aggregated by subfamily (L1HS is a human-specific subfamily of the LINE class). Some TEs have accumulated enough mutations across each locus that nearly all copies are uniquely mappable. Very recently inserted, and/or still active TEs, show the lowest mappability rates with many copies still very close to the consensus sequence (e.g. Alu and SVA types). By contrast, many older SINE and LINE TEs have high mappability rates and can easily be assessed using only uniquely aligning reads with standard analysis procedures. Mappability was calculated by counting number of times a 76 bp paired end read (242-mer with an internal gap of 100 nt) would map within the genome at a particular nucleotide where that nucleotide was the beginning of a 242-mer.

[41], Bowtie [42] and NovoAlign (http://novocraft.com/). For RNA-seq aligners, there are two approaches, those that align to reference transcriptomes and those that align to genomes. Transcriptome methods like kallisto [43] and Salmon [44] perform pseudoalignments with transcript derived k-mers and can attempt to build the reference transcriptome from the RNA-seq data itself. Salmon can be specified to report unmapped reads, kallisto does not include this option. While pseudoalignment is very fast, computationally less intensive, and helpful in organisms without a reference genome, it can be complicated in the context of repetitive elements, where all of the caveats that make genome assembly difficult (discussed above) would also apply to de novo transcriptome assembly. With regard to genome-based RNA-seq aligners, there are a number of packages available including: STAR [45], HISAT2 [46], GSNAP [47], NovoAlign, RUM [48], Minimap2 [49] and others [50]. In the context of sRNA-seq data, short-read genome-based aligners (BWA [41], Bowtie [42] and SCRAM [51]) that do not consider splice junctions tend to work as well or better than RNA-seq tailored algorithms, with SCRAM being specifically designed for small RNA analysis pipelines. Another approach to improve mappability would be to incorporate long-read sequencing methods, as longer reads contain more information and can serve as a way to reduce ambiguity in the context of RNA-seq. Many of the previous aligners like STAR, HISAT2 and GSNAP have been applied to long-read sequencing data after error correction [52] and have been shown to work well. In addition, algorithms like BLASR [53], GraphMap [54], rHAT [55], LAMSA [56], Kart [57], NGLMR [58] and lordFAST [59] have been developed specifically to address the increased length and error rates associated with long-read technologies.

Some tools designed to improve mapping rates for repetitive regions work after an initial analysis with one of the tools listed above. These standalone tools can use alignment files as input and then attempt to statistically redistribute the ambiguous reads based on distributions of neighbouring alignments. One such algorithm is MMR [60] which iteratively redistributes ambiguously mapped reads across their respective loci to maximize smoothness of multi-mapped read distribution in the context of unique reads, or reduce the variance in coverage. Another is a Gibbs sampling method [61] which uses stochastic redistribution of multi-mapped reads, normalized to the background distribution, in order to iteratively search for the most likely locus of origin. This type of iterative statistical technique for optimal assignment of reads to the correct loci has been picked up and elaborated on by several different groups, and represents a theme throughout the review. While it does not employ the statistical redistribution of reads, CoCo [62] is a package which corrects and salvages multimapped reads by taking into consideration nested genomic architecture, a common feature associated with TEs.

4. Analysis

The next step in a general next-generation sequencing (NGS) sequencing analysis pipeline is to annotate and quantify those reads which mapped to the genome. The mapping profiles will vary widely based on molecular context of the sequencing library. Each type of NGS data comes with its own challenges in the context of highly repetitive elements. The remaining sections will go through analysis strategies for each of the most common NGS data types in detail. The tools in these sections are listed for reference in figure 2, where they are grouped by the experimental assays used to generate the data. Electronic supplementary material, table S1 gives references and links to the software for all tools described.

5. RNA-seq

RNA-seq for expression analysis is one of the most well-studied areas in genomics, and this is also reflected in the diversity of tools available for analysis of transcripts from repetitive regions. RNA-seq data derived from short-read sequencing platforms is comprised of small fragments, derived from short single- or paired-end reads tiled across the region of a transcript of origin. Of the tools which have been developed to facilitate transcriptional analysis of repetitive elements, here we will focus on those which take into consideration ambiguously mapped reads. How to address ambiguously mapped reads is an old problem in genome science particularly when using older sequencing technologies from which reads were much shorter (approx. 36 nt) than what we currently consider a short read (approx. 150 nt). These early RNA-seq packages were largely gene-centric, as investigation of repetitive elements with these earlier technologies was (and remains) a challenge. However, the basic principles for probabilistic redistribution of ambiguously mapped reads emerged at this time. The first strategies employed a single-step multimapped read redistribution based on the number of uniquely mapped reads at each locus. [37] This was followed quickly by an expectation-maximization (EM) algorithm to iteratively estimate the most likely expression levels of gene transcripts based on relative counts of unique and multimapped reads [63]. In addition to probabilistic redistribution of reads, packages like Cufflinks [64] and HTSeq [65] have multimapper modes where ambiguously mapped reads are weighted by the relative number of genomic alignments (as $1/n$, where n is the number of potential alignments in the genome). The package Scavenger [66] considers multimapped reads and uses an intermediate consensus assignment with remapping to rescue unmapped reads. Differences in strategies used to address multimapped reads and their associated limitations are outlined in detail by Treangen & Salzberg [23].

As interest broadened to begin investigating transposon expression through RNA-seq explicitly, several packages were developed to handle transposons separately from the rest of the transcriptome. Among the first TE-centric packages was RepEnrich [67] which functions by creating repetitive element pseudochromosomes, which are a series of contigs that represent all of the genomic instances of each transposon subfamily annotated in RepeatMasker, concatenated onto a single region. These subfamily pseudochromosomes were then used to identify reads that mapped only to one subfamily of transposons, such as the human-specific LINE element L1Hs, even if the exact generating locus was still ambiguous. This was able to separate the level of uncertainty to finer detail, such that reads could be described as: unique in the genome, unique to a particular subfamily or ambiguously mapping to multiple types of transposons. Similar to RepEnrich, TETools [68] is another transcript quantification method which uses a detailed annotation file or 'rosette' to facilitate quantification from TE-derived reads, and which again aggregates reads at the subfamily level. TeXP [69] is a package which focuses on LINE-1 elements specifically and models spurious genome transcription to more accurately quantify LINE-1 expression. TEtranscripts [70] was the first TE-centric algorithm to implement statistical read redistribution to handle multimapped reads. TEtranscripts uses an expectation maximization algorithm to find the most likely distribution of ambiguously mapped TE-derived RNA-seq reads, and also includes expression estimates for both host genes and TEs in the output. After TEtranscripts, other packages have been developed to expand the methods used for statistical read redistribution including MMR [60] and SalmonTE, [71] with SalmonTE being unique in its use of a pseudoalignment strategy from the authors of the original Salmon [44] package in order to bypass the mapping step typically used in RNA-seq analysis. Yanagi [72] expands on this pseudoalignment strategy by mapping to a segmented version of the transcriptome to reduce ambiguity of mapping.

In the packages described above, quantification was performed at the subfamily level, as determining the specific expressed genomic loci within a subfamily is quite difficult for TEs that are close to the consensus sequence. However, several newer packages have been released to address the need for locus-specific quantification of TE-derived transcripts. TE-centric packages include SINEsFIND [73], and ERVmap [74] which are specialized for their respective TE family of interest. Two pipelines used genome-guided de novo transcriptome assembly with Trinity [75] to quantify TE expression at a locus-specific level: TEcandidates [76] and a pipeline described by Guffanti *et al.* [77] More recently, SQuIRE [78] (software for quantifying interspersed repeat expansion), and Telescope [79] adapted the EM-based read redistribution strategies described above to infer originating loci of ambiguously mapped reads, using uniquely mapped reads surrounding the locus to guide the EM read redistribution.

One of the motivating reasons to study TEs is for their influence over regulatory networks in our genome. To address this specifically, a final type of RNA-seq analysis package has been released at the interface of gene-centric and TE-centric models. LIONS [80] is a novel package which detects novel fusion events that connect TE promoter sequences to downstream coding gene sequences. These chimeric TE/gene transcripts represent one of the many ways that TE promoter elements might affect regulation of adjacent genes.

6. Small RNA-seq

Cells regulate TE expression using multiple strategies. The most potent silencers of TEs in germline cells are small RNAs (sRNAs) of the PIWI-interacting RNA (piRNA) class [81]. In somatic tissues, two additional classes of small

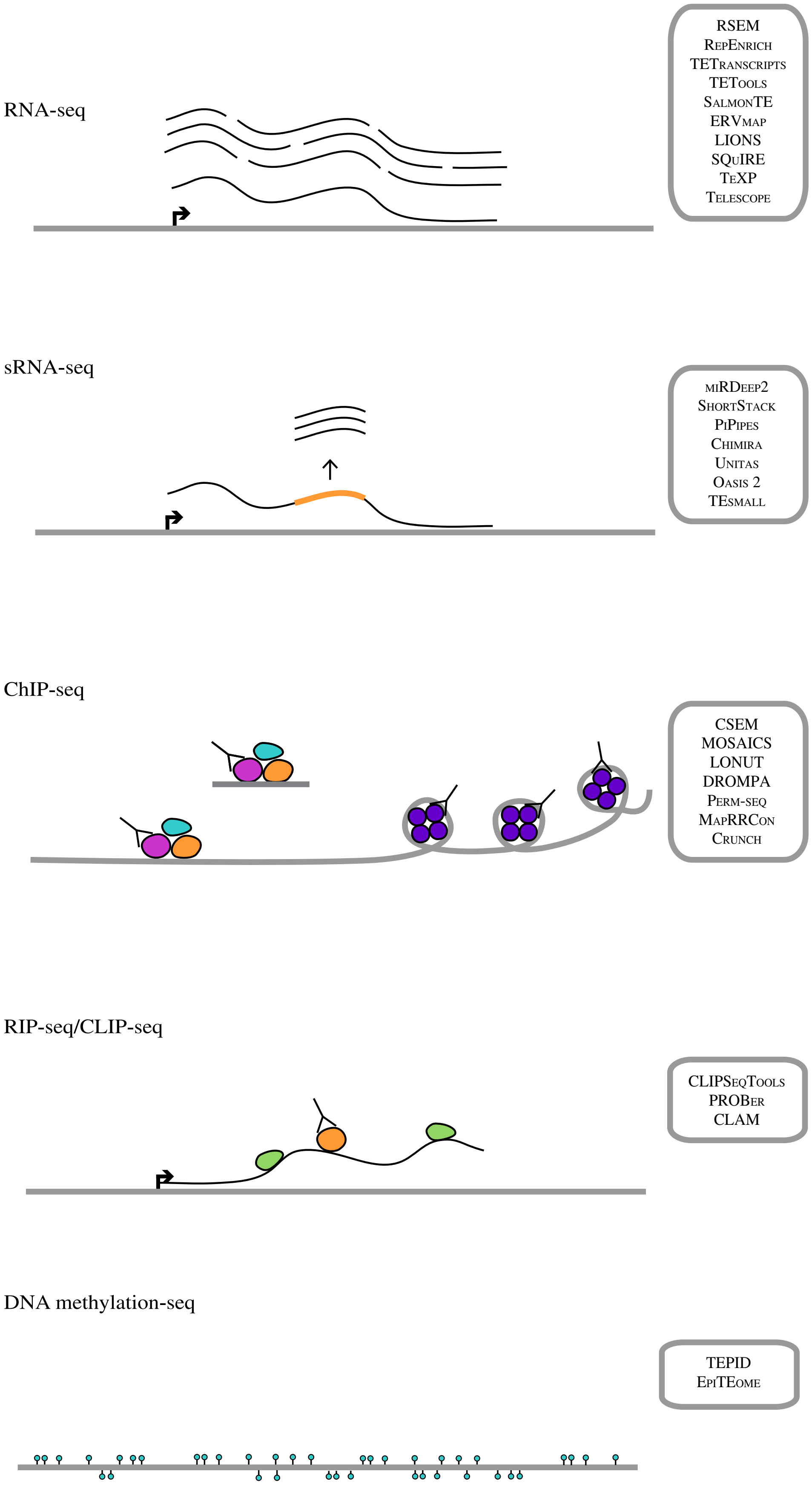

Figure 2. Published tools available for including repetitive regions in several common genomics analysis protocols. These have been divided into those that are geared towards RNA expression analysis (RNA-seq), small RNA expression analysis (sRNA-seq), genome and chromatin binding factors (ChIP-seq), RNA-binding factors (RIP/CLIP-seq) and DNA methylation analysis (DNA methylation-seq) A table describing these tools (electronic supplementary material, table S1) also provides links and references for the software and associated publications.

RNAs contribute to TE silencing: short interfering RNAs (siRNAs) derived from expressed transposon transcripts [81] and the more recently described 3′ tRNA derived fragments (3′ tRFs) [82]. Therefore, it is integral to the study of transposon biology to consider sRNAs and accurately quantify their production. To this end, several packages have been released to investigate sRNA species, which prove particularly challenging when derived from repetitive loci in the genome as they are short in length, typically between 18–36 nucleotides. Packages like MiRDeep2 [83], ShortStack [84], PiPipes [85], Chimira [86], sRNAtoolbox [87], Oasis 2 [88] and Manatee [89] have been developed to detect specific types of sRNA loci in the genome and quantify their differential expression. While microRNAs (miRNAs) are not known to play a large role in transposon regulation, a large fraction of miRNAs and other known TE regulatory sRNAs are present in multiple copies in the genome, making TE-focused strategies for multimapped read resolution useful, even for non-TE-derived sRNAs. Statistical techniques, including machine learning, have already been extensively employed in the arena of piRNA prediction, a critical step for the ultimate quantification of piRNA reads accumulation in packages like piRNAPredictor [90], Piano [91] and a k-mer-based method described by Zhang *et al.* [92] ShortStack after publication was updated to include Butter [93] which now performs statistical redistribution of multimapped reads.

These methods described above have largely considered sRNA classes separately, however, several packages including Unitas [94] and TEsmall [95] have strived to consider sRNA classes comprehensively to facilitate proper normalization of heterogeneous sRNA libraries, and to facilitate differential expression analysis across classes while taking into consideration ambiguously mapped reads.

While several iterative statistical methods have been employed in the study of sRNAs for annotation and target prediction [96], there is still much room for improvement in the handling of ambiguously mapped reads for small RNA expression analysis. Many of these issues have been nicely reviewed by Bousios *et al.* [97] particularly in the context of plants whose genomes are highly enriched in TEs and where sRNAs form a large component of the TE silencing machinery. Briefly, the chief challenge for applying probabilistic read redistribution algorithms for sRNA loci is that many types of sRNAs accumulate as very short transcripts cut from larger precursors. Often the precursors are rapidly processed and/or would not be caught by sRNA library preparation protocols. For miRNAs, for example, typically only the guide and passenger strands are detected in sRNA-seq libraries, leaving only two short approximately 22 nucleotide sRNAs and few surrounding reads from the precursor transcript to help guide decisions about the true originating locus. Thus, some loci may be more amenable to statistical inference algorithms, while others need additional assays in order to determine the precise source of sRNA biogenesis.

7. Immunoprecipitation-sequencing (ChIP, CLIP and RIP)

In this section, we have grouped together multiple disparate genomics data types that all involve immunoprecipitation-based steps in order to find protein binding sites in nucleic acids. These data can be derived from chromatin-bound factors (ChIP-seq) or RNA-binding proteins (CLIP-seq/RIP-seq), but are grouped here as IP-seq because of the similar challenges these data types present for computational analysis pipelines. Typically, the published pipelines for IP-seq data analysis begin by discarding multimapped reads in order to achieve higher specificity and resolution for the protein binding sites. This can be troublesome when studying proteins which bind to regions rich in repetitive elements. For example, H3K9me3 histone markers are known to be enriched in constitutive heterochromatin [98], a region of the genome highly enriched in repeat elements. Therefore, when calling H3K9me3 peaks using only uniquely mapped reads, the actual enrichment above background levels may be significantly higher than what is reported, skewing the estimates of background levels and discarding many truly bound regions. While this is a known issue for heterochromatin binding proteins, recent surveys of DNA- and RNA-factors have shown that transposon-derived regulatory elements form a significant fraction of both transcription factor binding sites [18,99] as well as RNA-binding protein recognition elements [100,101].

For ChIP-seq-based datasets, it is important to acknowledge the differences and difficulties associated with attempting to detect binding elements for chromatin binding factors and marked histones that typically bind broadly over large areas (broad peaks) when compared with transcription factors, which typically display sharp, narrow peaks. H3K9me3 typically shows a broad peak profile, as these histone marks are found on nucleosomes spread across wide stretches of chromatin. This distribution warrants a different detection strategy than that used for a typical transcription factor, such as MYC, which might occupy narrow binding regions, on the order of approximately 50–150 nucleotides in a typical assay. This is particularly relevant when these different peaks occur in repetitive genomic regions. The larger the bound region, the more likely it is that some of that genomic sequence will be uniquely mappable, which can guide the inference about read accumulation in adjacent sequences.

To address multimapped reads specifically, packages like the peak caller CSEM [102] have used expectation maximization to redistribute ambiguously mapped ChIP-seq reads based on the distribution of surrounding uniquely mapped reads. Owing to the reliance on uniquely mappable reads, these methods function best on broader peaks because they query a larger region, which may be more likely to contain uniquely mappable content. LONUT [103] calls a set of unique peaks and a set of non-unique peaks, then aggregates both call sets together to remove any redundancy. MOSAiCS [104], while not specifically developed to handle repetitive regions, recommends using the CSEM algorithm as a pre-processing step in order to include multimapped reads. DROMPA [105] and Crunch [106] take into account multimapped reads using a simple $1/n$ fractional distribution strategy. Crunch subsequently places a large emphasis on motif prediction and annotation. The analysis pipeline MapRRcon [107] uses unique and multimapped reads, but resolves the issue of multimapped read ambiguity by calling peaks on the consensus sequence of transposon subfamilies.

There is still significant room for progress in the arena of ChIP-seq analysis in repetitive regions. It is still difficult to call narrow peaks in repetitive regions, owing to the lack of sufficient reads surrounding the locus of interest to guide the inference algorithms. Perm-seq [108] addresses this issue by using the orthogonal dataset of DNAase hypersensitivity profiling for better resolution in repetitive regions of the genome. As sufficient reference datasets become available in multiple cell types and conditions, this may make this strategy feasible as a general method. By contrast, while broad peak callers tend to include more information within the locus of interest to help guide inference across repetitive regions, the data from these methods tend to have a lower signal-to-noise ratio, such that improvement of broad peak callers generally is still an active area of computational development.

The problems described above in the context of ChIP-seq analysis are compounded in the context of CLIP- and RIP-seq datasets, where one must also normalize for differences in the expression level of the bound transcript substrates. If the bound transcripts contain repetitive regions, or are entirely composed of repetitive elements, one must first find a way to accurately distribute ambiguous reads among the input transcriptome dataset before calling enriched binding sites in particular transcripts. CLIPPER [109] was one of the first CLIP-seq pipelines, but was restricted to uniquely mapped reads only. CLIPSEQTOOLS [110] is a CLIP-analysis pipeline which randomly assigns ambiguously mapped reads to one of their candidate mapping loci. CLAM [111] uses expectation maximization algorithms, as described above, to redistribute ambiguously mapped reads between expressed transcripts, but the algorithm works only on the alignment file and does not include information about enriched peaks in its statistical weights. PROBER [112] has been developed as a general-purpose algorithm for detecting sites of RNA binding or modification (termed 'toeprint' profiling) and includes an algorithm for handling multimapped reads using a Gibbs sampler approach to iteratively infer a single 'best' alignment for each read. While PROBER does include steps to handle multimapped reads, it was not developed specifically for TEs, and thus has not been tested on highly repetitive regions, such as TEs that are very close to the consensus.

8. DNA methylation-sequencing

We have detailed several methods to asses differential expression, and protein binding in the context of repetitive elements. However, a critical component to the understanding of transposon biology is the analysis of DNA methylation as it is the main mechanism by which transposons are transcriptionally silenced long term [113]. To assess DNA methylation, particularly the 5-methylcytosine (5-mC) modification, several techniques have been developed and compared [114]. In brief, the most common method to assess DNA methylation is bisulfite sequencing: whole-genome DNA sequencing following bisulfite conversion of all non-methylated cytosine residues to uracil. Bisulfite sequencing-based methods can be non-directional [115], or directional [116] allowing one to reduce the ambiguity of strand of origin. One of the first analysis pipelines developed for high-throughput bisulfite sequencing was in *Arabidopsis* [116] and analysis was performed in conjunction with sRNA-seq datasets. In this pipeline, ambiguously mapped reads were discarded by mapping to a repeat masked version of the genome, a technique once commonly used in animal systems to reduce mapping ambiguity in the context of bisulfite induced C > T conversions [117]. Bisulfite sequencing analysis differs significantly from other analysis pipelines in that often two reference genomes are used, one which contains converted cytosines in addition to the original reference genome. In this context, what are considered ambiguous reads are those reads which map to both the converted and unconverted reference genomes. This compounds the difficulty of assigning multimapped reads, such that many published bisulfite sequencing software packages choose not to include multimapped reads to avoid this confounded ambiguity (electronic supplementary material, table S1). The most commonly used pipelines for bisulfite sequencing reads including BSMAP [118], BISMARK [119], MOABS [120] and BS-SEEKER3 [121], none of which include probabilistic handling of multimapper reads. For a more comprehensive list of non-TE-specific methylation pipelines, please see the review by Adusumalli *et al*. [122] and the supplemental material of a recently published pipeline, BICYCLE [123]. Here, confounding between ambiguity in bisulfite conversion rates, non-reference polymorphisms and read non-uniqueness can complicate the statistical tests used to determine if a site in the genome is differentially methylated. Thus, this represents an area of computational genomics that could benefit greatly from further development.

Because DNA methylation is a critical mechanism by which transposons are silenced, several groups have used new methods to improve methylation analysis for TEs. TEPID [124] and EPITEOME [125] were designed to improve analysis of TE methylation levels by including the analysis of split reads that cross junctions between TEs and uniquely mappable genome regions. An approach employed to asses the low mappability of young TEs, like L1-Ta, in the human genome was repurposed to align bisulfite reads to a consensus sequence as described in Shukla *et al*. [126]. One interesting method to improve methylation analysis is to first rigorously determine the average bisulfite conversion rates genome-wide, then use this as a parameter to tease apart mapping ambiguities from differences in conversion rates, as done by Noshay *et al*. [127]. Despite these improvements, DNA methylation analysis is still a difficult bioinformatic challenge that would benefit from further study.

9. Single-cell RNA-sequencing

All of the software described above has been geared towards genomics datasets generated from bulk tissue samples. However, bulk profiling of heterogeneous cell populations only provides averages that obscure underlying variability of TE expression across cell types, as illustrated in figure 3. This problem is further amplified when aggregating transcriptional signal across numerous loci within high copy-number TE families. It remains largely unknown how TE de-repression varies between individual cells, what factors drive such differences, and how this variability might affect cellular phenotypes. Single-cell RNA-sequencing (scRNA-seq) promises to answer some of those questions and has already redefined our knowledge about tissue composition and gene regulatory networks [128]. While its broad application has so far been largely restricted to the study of gene activity patterns, a few pioneering studies have used first-generation protocols to identify TE expression dynamics across single pre-implantation embryonic cells [129,130]. Those early efforts were largely limited by small cell numbers, high sequencing burden per cell, and lack of molecular barcode counts to estimate true transcriptional output, thus preventing broad-scale adaptation. Since then, the increasing demand in single-cell transcriptome data has

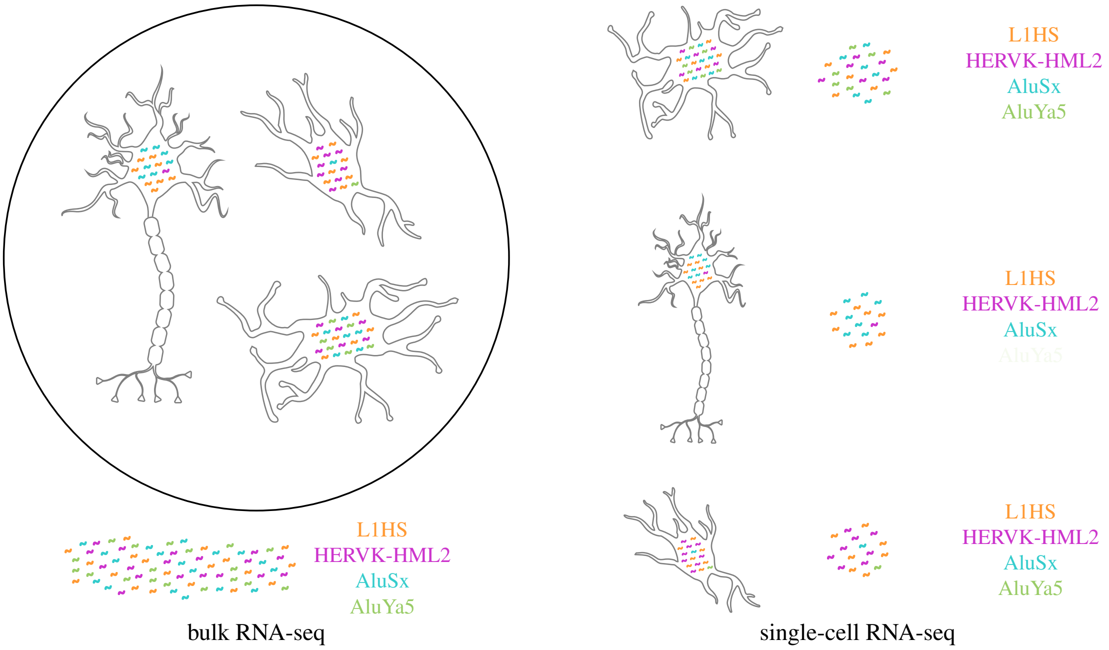

Figure 3. Comparison of bulk RNA-seq versus single-cell RNA-seq. Heterogeneity in expression profiles across cell types is masked by bulk sequencing methods. Transposable element (TE) expression may vary across cell types, between cells of the same type, and within the same cells across time. Single-cell methods are necessary to reveal this heterogeneity, but software for single-cell data analysis is not currently optimized for handling TEs.

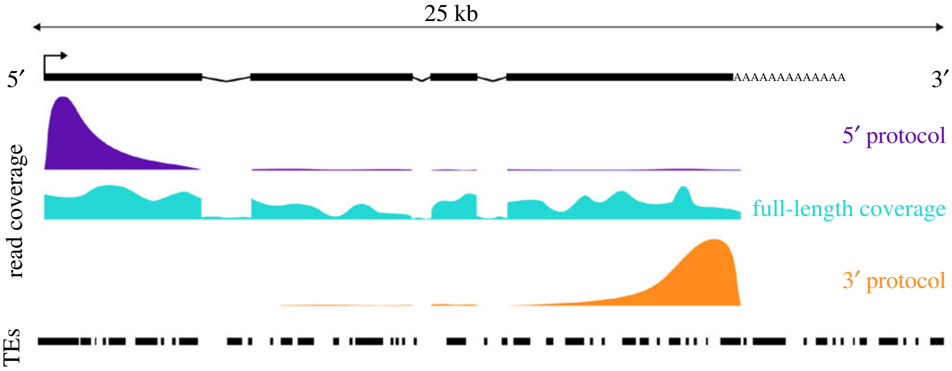

Figure 4. Impact of different RNA-seq library strategies on read coverage along a TE-derived transcript with exon/intron structure (top). TE intervals are shown at the bottom. Briefly, scRNA-seq protocols that offer full-length transcript coverage provide the best means to identify full length transcribed TEs in a locus-specific manner, but this method suffers from noise owing to intronic TEs in host genes that might be mistaken for expressed TE transcripts as well as the inability to barcode individual mRNA molecules. 5′- and 3′-based protocols allow for barcodes that enable mRNA molecule counting, with 5′ protocols also offering the ability to detect TE transcripts originating from proper TE promoters.

seen an unprecedented expansion of available scRNA-seq protocols with considerably improved throughput, robustness, and error-rates [131]. One such publication was by Guo *et al.* [132], where the number of cells were scaled up allowing for investigation into TE dynamics in spermatogenesis.

Despite such experimental advancements, inherent design principles of scRNA-seq protocols that cooperate with the well-known challenges of TE transcriptome analysis have so far prevented their common application for the study of TE expression at single-cell resolution (figure 4). For example, many popular methods quantify RNA molecules at the 3′ end of polyadenylated mRNAs [133–137] and therefore depend on accurate reference models to bridge the gap between polyadenylation sites and the corresponding transcript isoform and/or promoter. This is problematic for TE-derived transcripts, which are generally poorly annotated in many species. While protocols with full-length transcript coverage might alleviate some of those problems, the naive assignment of reads to the nearest TE interval can still lead to erroneous assignment, misattribution of intronic reads from unprocessed pre-mRNAs, and hence misinterpretation of TE de-repression. Full-length protocols additionally suffer from higher sequencing burden, often lack of unique molecular identifiers to account for PCR duplicates, and potentially higher background TE read coverage owing to intronic signal originating from pre-mRNAs [138,139].

A potential solution to minimize misattribution problems are 5′ end based scRNA-seq protocols that incorporate a template switch oligo (TSO) towards the start of transcription initiation [140,141]. Although incomplete processing and premature TSO incorporation during library preparation might vary between transcripts and cells, such protocols have already been successfully used to map alternative transcription start sites between individual cells [142]. Importantly, a recent study also demonstrated its utility to quantify unexpected variability in TE promoter activity between thousands of single cancer cells following epigenetic therapy [143]. However, the problem of premature TSO incorporation, combined with the pervasive nature of TEs, and technical noise inherent to all current scRNA-seq protocols requires dedicated strategies to mitigate the danger of spurious estimates of TE cell-to-cell variation. To the best of our knowledge, no peer-reviewed computational pipeline currently combines such features with the reliable quantification of TEs at single-cell resolution, but unpublished efforts already aim to facilitate TE single-cell analysis for a wide array of available scRNA-seq protocols (https://tanaylab.github.io/Repsc/). With the continuous methodological advancements and the increasing interest in TE biology, we anticipate a rapid progress towards the routine quantification of TEs in individual cells that will be accompanied by the discovery of unprecedented heterogeneity in TE transcription patterns.

10. Conclusion

(a) What is now doable?

The last years have seen a general improvement in sequencing read length, making it possible to study the majority of TEs in a genome-wide fashion. For particularly young and less diverged families, we have discussed at length the strides made in genome biology to address the difficulties of treating ambiguously mapped sequencing fragments for differential expression and binding analyses. In the context of highly repetitive regions of the genome, these difficulties are compounded, particularly for the most active TEs, which remain close to their consensus sequence and thus are the most difficult to map. The greatest progress has been made with RNA-seq data analysis, as we have progressed from using simple fractional assignments of multimapped reads within genes to approaching true locus-specific resolution in the most repetitive regions of the genome—such as the L1HS subfamily, active Alu families and composite SVA elements. Progress has been made in the realm of sRNA analysis as these improved algorithms for RNA-seq analysis have now been incorporated into sRNA-seq data analysis pipelines. In immunoprecipitation based assays, for ChIP- and CLIP-seq datasets, efforts have been made to use probabilistic read redistribution for peaks within repetitive regions, but challenges remain.

(b) What is still hard?

sRNA-seq data contains a large proportion of multimapped reads, and while significant effort has been put forth to leverage advanced iterative statistical methods for novel sRNA discovery and target prediction, these methods have not been as widely applied to sRNA-seq transcript quantification. This may be attributed to the tight distribution of sRNA reads across their mapping loci, making it difficult to garner locus-specific information from adjacent reads. Moreover, these much shorter reads (18–30 nt) are intrinsically less unique in the genome than longer sequences.

In ChIP-seq data, the expected profile of read distributions can vary widely from the typically tall, narrow peaks associated with most transcription factor profiles or RNA-binding proteins to the broader, shorter and noisier peaks associated with some marked histones, such as H3K9me3. Algorithms have been developed to address both types of ChIP-seq profiles. Yet the lines between these categories can be blurred, and there is a large trade-off between the window size in peak calling and the ability to use uniquely mapped reads to probabilistically reassign all other reads to a particular locus. One area of active research for broader regions would be to incorporate multimapped reads into segmentation models which allow for the detection of changes in peak landscape, as opposed to simply calling the absence or presence of individual peaks.

scRNA-seq represents one of the newest genomic assays to be used for TE expression profiling, and as such, remains an area of greatest need for improvements in software packages specifically designed to handle the complexities inherent in TE genomics. Efforts are already underway, but as yet no published software packages for scRNA-seq are available. That said, many standard scRNA-seq packages could be adapted for this use, as in the example protocol described above. However, as discussed in detail, differences in the experimental protocols used to generate scRNA-seq libraries will have a large impact upon the interpretability of the data, and this is particularly problematic for TE expression analysis.

Two types of analysis which largely do not include multimapped reads are assays for transposase accessible chromatin using sequencing (ATAC-seq) [144] and Hi-C [145], an extension of chromosome conformation capture (3C). The read distributions for ATAC-seq data greatly resemble those of ChIP-seq and this analysis encounters similar computational difficulties when studying repetitive regions of the genome. Fortunately, as this analysis is similar to ChIP-seq there has already been significant effort which could be incorporated into ATAC-seq analysis. Adapting Hi-C pipelines to take into account multimapped reads is still a difficult task as this type of analysis already requires the resolution of chimeric reads representing genomic proximity. mHiC [146] has been developed to address this issue, but the relative sensitivity to highly repetitive transposon regions is unclear. Significant work has been done using these methods to address the role of transposons in genome architecture and the transition from the embryonic cell state to early embryonic-like cells [99,147,148]. These analyses can only improve as better methods for handling repetitive reads are included.

(c) What new technology needs to be developed?

Long-read sequencing technologies promise to solve many issues inherent in the assays described above. Once the issues with throughput and error rates can be solved, long-read sequencing would enable the isolation of entire transcripts and, if correctly barcoded, would also allow for accurately calibrated expression estimates. These technologies could also be combined with antibody-based pulldowns and endonuclease-based footprinting assays, to accurately call *cis*-regulatory regions derived from TEs. Finally, long-read genome resequencing assays that sequence through highly repetitive genome regions may allow for better genomic annotations that will benefit all of the applications described above. To this end, not only must new experimental protocols be developed which emphasize longer reads but new computational pipelines must also be developed to ensure that these long read analysis pipelines properly handle and account for the complications inherent in addressing TE genomics.

Data accessibility. This article has no additional data.

Authors' contributions. All authors contributed to the writing and editing of this manuscript.

Competing interests. The authors have no competing interests to declare.

Funding. M.G.H. receives support from the Rita Allen Foundation and the Chan-Zuckerberg Initiative. K.O. acknowledges funding from the National Science Foundation Graduate Research Fellowship and NIH training grant no. (2T32GM065094-16). D.B. is supported by an EMBO long-term fellowship grant no. (ALTF 900-2017).

Acknowledgements. We acknowledge Ying Jin for her helpful conversations about TE genomics. We apologize for any software we failed to cite in this review pertinent to the study of repetitive elements in genome biology. Schematic images were adapted with permission from Servier Medical Art (https://smart.servier.com).

References

1. Pace II JK, Feschotte C. 2007 The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.* **17**, 422–432. (doi:10.1101/gr.5826307)
2. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011 Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384. (doi:10.1371/journal.pgen.1002384)
3. Schnable PS *et al.* 2009 The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115. (doi:10.1126/science.1178534)
4. Feschotte C, Pritham EJ. 2007 DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331–368. (doi:10.1146/annurev.genet.40.110405.090448)
5. Slotkin RK, Martienssen R. 2007 Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285. (doi:10.1038/nrg2072)
6. Levin HL, Moran JV. 2011 Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627. (doi:10.1038/nrg3030)
7. Boissinot S, Chevret P, Furano AV. 2000 L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17**, 915–928. (doi:10.1093/oxfordjournals.molbev.a026372)
8. Sheen F, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD. 2000 Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496–1508. (doi:10.1101/gr.149400)
9. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH. 2003 Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA* **100**, 5280–5285. (doi:10.1073/pnas.0831042100)
10. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014 The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21**, 423–425. (doi:10.1038/nsmb.2799)
11. Grow EJ *et al.* 2015 Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225. (doi:10.1038/nature14308)
12. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. 2017 LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet.* **49**, 1502–1510. (doi:10.1038/ng.3945)
13. Rowe HM *et al.* 2013 TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res.* **23**, 452–461. (doi:10.1101/gr.147678.112)
14. Percharde M *et al.* 2018 A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell* **174**, 391–405. (doi:10.1016/j.cell.2018.05.043)
15. Chuong EB, Elde NC, Feschotte C. 2017 Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86. (doi:10.1038/nrg.2016.139)
16. Imbeault M, Helleboid P-Y, Trono D. 2017 KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554. (doi:10.1038/nature21683)
17. Chuong EB, Elde NC, Feschotte C. 2016 Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1088. (doi:10.1126/science.aad5497)
18. 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 Res.* **24**, 1–15. (doi:10.1101/gr.168872.113)
19. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019 Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell* **24**, 724–735.e5. (doi:10.1016/j.stem.2019.03.012)
20. Venuto D, Bourque G. 2018 Identifying co-opted transposable elements using comparative epigenomics. *Dev. Growth Differ.* **60**, 53–62. (doi:10.1111/dgd.12423)
21. Raviram R, Rocha PP, Luo VM, Swanzey E, Miraldi ER, Chuong EB, Feschotte C, Bonneau R, Skok JA. 2018 Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol.* **19**, 1–19. (doi:10.1186/s13059-018-1598-7)
22. Goerner-Potvin P, Bourque G. 2018 Computational tools to unmask transposable elements. *Nat. Rev. Genet.* **19**, 688–704. (doi:10.1038/s41576-018-0050-x)
23. Treangen TJ, Salzberg SL. 2012 Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46. (doi:10.1038/nrg3117)
24. Jain M *et al.* 2018 Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323. (doi:10.1038/nbt.4109)
25. Gordon D *et al.* 2016 Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344. (doi:10.1126/science.aae0344)
26. Wong TN *et al.* 2018 Cellular stressors contribute to the expansion of hematopoetic clones of varying leukemic potential. *Nat. Commun.* **9**, 1–10. (doi:10.1038/s41467-017-02088-w)
27. Chaisson MJP *et al.* 2019 Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16. (doi:10.1038/s41467-018-07882-8)
28. Bao W, Kojima KK, Kohany O. 2015 Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 4–9. (doi:10.1186/s13100-015-0035-7)
29. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002 The human genome browser at UCSC. *Genome Res.* **12**, 996–1006. (doi:10.1101/gr.229102)
30. Amselem J *et al.* 2019 RepetDB: a unified resource for transposable element references. *Mob. DNA* **10**, 4–11. (doi:10.1186/s13100-019-0150-y)
31. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, Stoye J, Tristem M, Johnson WE. 2018 Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* **15**, 1–11. (doi:10.1186/s12977-018-0442-1)
32. 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 Res.* **44**, 81–89. (doi:10.1093/nar/gkv1272)
33. Wicker T, Matthews DE, Keller B. 2002 TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.* **7**, 561–562. (doi:10.1016/S1360-1385(02)02372-5)
34. Yi F, Jia Z, Xiao Y, Ma W, Wang J. 2018 SPTEdb: a database for transposable elements in salicaceous plants. *Database* **2018**, 1–8. (doi:10.1093/database/bay024)
35. Yi F, Ling J, Xiao Y, Zhang H, Ouyang F, Wang J. 2018 ConTEdb: a comprehensive database of transposable elements in conifers. *Database* **2018**, 1–7. (doi:10.1093/database/bay131)
36. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M. 2013 MIPS PlantsDB: a database framework for

comparative plant genome research. *Nucleic Acids Res.* **41**, 1144–1151. (doi:10.1093/nar/gks1153)

37. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008 Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628. (doi:10.1038/nmeth.1226)
38. Mills RE, Bennett EA, Iskow RC, Devine SE. 2007 Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191. (doi:10.1016/j.tig.2007.02.006)
39. Sexton CE, Han MV. 2019 Paired-end mappability of transposable elements in the human genome. *Mob. DNA* **10**, 1–11. (doi:10.1186/s13100-019-0172-5)
40. Feschotte C. 2008 Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405. (doi:10.1038/nrg2337)
41. Li H, Durbin R. 2010 Fast and accurate long-read alignment with Burrows–wheeler transform. *Bioinformatics* **26**, 589–595. (doi:10.1093/bioinformatics/btp698)
42. Langmead B. 2010 Aligning short sequencing reads with Bowtie. *Curr. Protocol Bioinf.* **32**, 1–24. (doi:10.1002/0471250953.bi1107s32)
43. Bray NL, Pimentel H, Melsted P, Pachter L. 2016 Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527. (doi:10.1038/nbt.3519)
44. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017 Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419. (doi:10.1038/nmeth.4197)
45. 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. (doi:10.1093/bioinformatics/bts635)
46. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915. (doi:10.1038/s41587-019-0201-4)
47. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016 GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. In *Stat. Genomics methods protoc* (eds E Mathé, S Davis), pp. 283–334. New York, NY: Springer.
48. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. 2011 Comparative analysis of RNA-seq alignment algorithms and the RNA-seq unified mapper (RUM). *Bioinformatics* **27**, 2518–2528. (doi:10.1093/bioinformatics/btr427)
49. Li H. 2018 Sequence analysis Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. (doi:10.1093/bioinformatics/bty191)
50. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, Fitzgerald GA, Grant GR. 2017 Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* **14**, 135–139. (doi:10.1038/nmeth.4106)
51. Fletcher SJ, Boden M, Mitter N, Carroll BJ. 2018 SCRAM: a pipeline for fast index-free small RNA read alignment and visualization. *Bioinformatics* **34**, 2670–2672. (doi:10.1093/bioinformatics/bty161)
52. Krizanovic K, Echchiki A, Roux J, Sikic M. 2018 Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**, 748–754. (doi:10.1093/bioinformatics/btx668)
53. Chaisson MJ, Tesler G. 2012 Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinf.* **13**, 238. (doi:10.1186/1471-2105-13-238)
54. Sovic I, Sikic M, Wilm A, Fenlon SN, Chen S, Nagarajan N. 2016 Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* **7**, 11307. (doi:10.1038/ncomms11307)
55. Liu B, Guan D, Teng M, Wang Y. 2015 rHAT: fast alignment of noisy long reads with regional hashing. *Bioinformatics* **32**, 1625–1631. (doi:10.1093/bioinformatics/btv662)
56. Liu B, Gao Y, Wang Y. 2017 LAMSA: fast split read alignment with long approximate matches. *Bioinformatics* **33**, 192–201. (doi:10.1093/bioinformatics/btw594)
57. Lin H, Hsu W. 2017 Kart: a divide-and-conquer algorithm for NGS read alignment. *Bioinformatics* **33**, 2281–2287. (doi:10.1093/bioinformatics/btx189)
58. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, Schatz MC. 2018 Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468. (doi:10.1038/s41592-018-0001-7)
59. Haghshenas E, Sahinalp SC, Hach F. 2019 lordFAST: sensitive and fast alignment search tool for long noisy read sequencing data. *Bioinformatics* **35**, 20–27. (doi:10.1093/bioinformatics/bty544)
60. Kahles A, Behr J, Rätsch G. 2016 MMR: a tool for read multi-mapper resolution. *Bioinformatics* **32**, 770–772. (doi:10.1093/bioinformatics/btv624)
61. Wang J, Huda A, Lunyak VV, Jordan IK. 2010 A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics* **26**, 2501–2508. (doi:10.1093/bioinformatics/btq460)
62. Deschamps-Francoeur G, Boivin V, Sherif AE, Scott MS. 2019 CoCo: RNA-seq read assignment correction for nested genes and multimapped reads. *Bioinformatics* **35**, 5039–5047. (doi:10.1093/bioinformatics/btz433)
63. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010 RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500. (doi:10.1093/bioinformatics/btp692)
64. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010 Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 516–520. (doi:10.1038/nbt.1621)
65. Anders S, Pyl PT, Huber W. 2015 HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169. (doi:10.1093/bioinformatics/btu638)
66. Yang A, Tang JYS, Troup M, Ho JWK. 2019 Scavenger: a pipeline for recovery of unaligned reads utilising similarity with aligned reads. *F1000 Res* **8**, 1587. 1–20. (doi:10.12688/f1000research.19426.1)
67. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. 2014 Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, 1–17. (doi:10.1186/1471-2164-15-583)
68. Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. 2017 TETOOLS facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* **45**, 1–12. (doi:10.1093/nar/gkw1046)
69. Navarro FC, Hoops J, Bellfy L, Cerveira E, Zhu Q, Zhang C, Lee C, Gerstein MB. 2019 TeXP: deconvolving the effects of pervasive and autonomous transcription of transposable elements. *PLoS Comput. Biol.* **15**, 1–19. (doi:10.1371/journal.pcbi.1007293)
70. Jin Y, Tam OH, Paniagua E, Hammell M. 2015 TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599. (doi:10.1093/bioinformatics/btv422)
71. Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. 2018 An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. *Biocomputing* **2018**, 168–179. (doi:10.1142/9789813235533_0016)
72. Gunady MK, Mount SM, Bravo HC. 2018 Fast and interpretable alternative splicing and differential gene-level expression analysis using transcriptome segmentation with Yanagi. *bioRxiv* 1–23. (doi:10.1101/364281)
73. Carnevali D, Conti A, Pellegrini M, Dieci G. 2017 Whole-genome expression analysis of mammalian-wide interspersed repeat elements in human cell lines. *DNA Res.* **24**, 59–69. (doi:10.1093/dnares/dsw048)
74. Tokuyama M, Kong Y, Song E, Jayewickreme T, Kang I, Iwasaki A. 2018 ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc. Natl Acad. Sci. USA* **115**, 12 565–12 572. (doi:10.1073/pnas.1814589115)
75. Haas BJ *et al.* 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512. (doi:10.1038/nprot.2013.084)
76. Valdebenito-Maturana B, Riadi G. 2018 TEcandidates: prediction of genomic origin of expressed transposable elements using RNA-seq data. *Bioinformatics* **34**, 3915–3916. (doi:10.1093/bioinformatics/bty423)
77. Guffanti G, Bartlett A, Klengel T, Klengel C, Hunter R, Glinsky G, Arkhipova I. 2018 Novel bioinformatics approach identifies transcriptional profiles of lineage-specific transposable elements at distinct loci in the human dorsolateral prefrontal cortex.

Mol. Biol. Evol. **35**, 2435–2453. (doi:10.1093/molbev/msy143)

78. Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. 2019 SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res.* **47**, 1–16. (doi:10.1093/nar/gky1229)
79. Bendall ML *et al.* 2019 Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput. Biol.* **15**, e1006453. 1–25. (doi:10.1371/journal.pcbi.1006453)
80. Babaian A, Thompson IR, Lever J, Gagnier L, Karimi MM, Mager DL. 2019 LIONS: analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinformatics* **35**, 3839–3841. (doi:10.1093/bioinformatics/btz130)
81. Malone CD, Hannon GJ. 2009 Small RNAs as guardians of the genome. *Cell* **136**, 656–668. (doi:10.1016/j.cell.2009.01.045)
82. Schorn AJ, Gutbrod MJ, LeBlanc C, Martienssen R. 2017 LTR-retrotransposon control by tRNA-derived small RNAs. *Cell* **170**, 61–71. (doi:10.1016/j.cell.2017.06.013)
83. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012 miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52. (doi:10.1093/nar/gkr688)
84. Axtell MJ. 2013 ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19**, 740–751. (doi:10.1261/rna.035279.112)
85. Han BW, Wang W, Zamore PD, Weng Z. 2015 piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics* **31**, 593–595. (doi:10.1093/bioinformatics/btu647)
86. Vitsios DM, Enright AJ. 2015 Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics* **31**, 3365–3367. (doi:10.1093/bioinformatics/btv380)
87. Rueda A, Barturen G, Lebrón R, Gómez-Martín C, Alganza Á, Oliver JL, Hackenberg M. 2015 sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.* **43**, W467–W473. (doi:10.1093/nar/gkv555)
88. Rahman RU, Gautam A, Bethune J, Sattar A, Fiosins M, Magruder DS, Capece V, Shomroni O, Bonn S. 2018 Oasis 2: improved online analysis of small RNA-seq data. *BMC Bioinf.* **19**, 1–10. (doi:10.1186/s12859-018-2047-z)
89. Handzlik JE, Tastsoglou S, Vlachos IS, Hatzigeorgiou AG. 2020 Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data. *Sci. Rep.* **10**, 705. (doi:10.1038/s41598-020-57495-9)
90. Li D, Luo L, Zhang W, Liu F, Luo F. 2016 A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinf.* **17**, 1–11. (doi:10.1186/s12859-015-0844-1)
91. Wang K, Liang C, Liu J, Xiao H, Huang S, Xu J, Li F. 2014 Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinf.* **15**, 1–8. (doi:10.1186/1471-2105-15-S12-S1)
92. Zhang Y, Wang X, Kang L. 2011 A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics* **27**, 771–776. (doi:10.1093/bioinformatics/btr016)
93. Axtell MJ. 2014 Butter: High-precision genomic alignment of small RNA-seq data. *bioRxiv* 1–16. (doi:10.1101/007427)
94. Gebert D, Hewel C, Rosenkranz. D 2017 unitas: the universal tool for annotation of small RNAs. *BMC Genomics* **18**, 1–14. (doi:10.1186/s12864-017-4031-9)
95. O'Neill K, Liao WW, Patel A, Hammell M. 2018 TEsmall identifies small RNAs associated with targeted inhibitor resistance in melanoma. *Front. Genet.* **9**, 461. (doi:10.3389/fgene.2018.00461)
96. Hadi LHA, Lin QXXL, Minh TT, Loh M, Ng HK, Salim A, Soong R, Benoukraf T. 2018 miREM: an expectation-maximization approach for prioritizing miRNAs associated with gene-set. *BMC Bioinf.* **19**, 1–8. (doi:10.1186/s12859-017-2006-0)
97. Bousios A, Gaut BS, Darzentas N. 2017 Considerations and complications of mapping small RNA high-throughput data to transposable elements. *Mob. DNA* **8**, 1–13. (doi:10.1186/s13100-017-0086-z)
98. Zhang T, Cooper S, Brockdorff N. 2015 The interplay of histone modifications – writers that read. *EMBO Rep.* **16**, 1467–1481. (doi:10.15252/embr.201540945)
99. Cao Y *et al.* 2019 Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res.* **29**, 40–52. (doi:10.1101/gr.241141.118)
100. Attig J *et al.* 2018 Heteromeric RNP assembly at lines controls lineage-specific RNA processing. *Cell* **174**, 1067–1081. (doi:10.1016/j.cell.2018.07.001)
101. Kelley DR, Hendrickson DG, Tenen D, Rinn JL. 2014 Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol.* **15**, 1–16. (doi:10.1186/s13059-014-0537-5)
102. Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keleş S. 2011 Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-seq data. *PLoS Comput. Biol.* **7**, e1002111. (doi:10.1371/journal.pcbi.1002111)
103. Wang R *et al.* 2013 Locating non-unique matched tags (LONUT) to improve the detection of the enriched regions for ChIP-seq Data. *PLoS ONE* **8**, 1–10. (doi:10.1371/annotation/5fa9cfb4-9964-4586-845d-d8205f318d68)
104. Sun G, Chung D, Liang K, Keles S. 2013 Statistical Analysis of ChIP-seq data with MOSAiCS. In *Deep Seq. Data anal.* (ed N Shomron), pp. 193–212. New York, NY: Springer Science+Business Media.
105. Nakato R, Itoh T, Shirahige K. 2013 DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. *Genes Cells* **18**, 589–601. (doi:10.1111/gtc.12058)
106. Berger S, Pachkov M, Arnold P, Omidi S, Kelley N, Salatino S, Van Nimwegen E. 2019 Crunch: integrated processing and modeling of ChIP-seq data in terms of regulatory motifs. *Genome Res.* **29**, 1164–1177. (doi:10.1101/gr.239319.118)
107. Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, Mita P, Fenyö D, Boeke JD. 2018 Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc. Natl Acad. Sci. USA* **115**, E5526–E5535. (doi:10.1073/pnas.1722565115)
108. Zeng X, Li B, Welch R, Rojo C, Zheng Y, Dewey CN, Keleş S. 2015 Perm-seq: mapping protein-DNA interactions in segmental duplication and highly repetitive regions of genomes with prior-enhanced read mapping. *PLoS Comput. Biol.* **11**, 1–23. (doi:10.1371/journal.pcbi.1004491)
109. Lovci MT *et al.* 2013 Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442. (doi:10.1038/nsmb.2699)
110. Maragkakis M, Alexiou P, Nakaya T, Mourelatos Z. 2016 CLIPSeqTools: a novel bioinformatics CLIP-seq analysis suite. *RNA* **22**, 1–9. (doi:10.1261/rna.052167.115)
111. Zhang Z, Xing Y. 2017 CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res.* **45**, 9260–9271. (doi:10.1093/nar/gkx646)
112. Li B, Tambe A, Aviran S, Pachter L. 2017 PROBer provides a general toolkit for analyzing sequencing-based toeprinting assays. *Cell Syst.* **4**, 568–574. (doi:10.1016/j.cels.2017.04.007)
113. Deniz Ö, Frost JM, Branco MR. 2019 Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431. (doi:10.1038/s41576-019-0106-6)
114. Harris RA *et al.* 2010 Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097. (doi:10.1038/nbt.1682)
115. Cokus SJ *et al.* 2008 Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219. (doi:10.1038/nature06745)
116. Lister R, Malley RCO, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008 Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536. (doi:10.1016/j.cell.2008.03.029)
117. Lister R *et al.* 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322. (doi:10.1038/nature08514)
118. Xi Y, Li W. 2009 BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinf.* **10**, 1–9.
119. Krueger F, Andrews SR. 2011 Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572. (doi:10.1093/bioinformatics/btr167)
120. Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. 2014 MOABS: model based

analysis of bisulfite sequencing data. *Genome Biol.* **15**, 1–12. (doi:10.1186/gb-2014-15-1-r1)

121. Huang KYY, Huang Y-J, Chen P-Y. 2018 BS-Seeker3: ultrafast pipeline for bisulfite sequencing. *BMC Bioinf.* **19**, 2–5. (doi:10.1186/s12859-017-2004-2)

122. Adusumalli S, Feroz Mohd Omar M, Soong R, Benoukraf T. 2014 Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief. Bioinform.* **16**, 369–379. (doi:10.1093/bib/bbu016)

123. Graña O, López-Fernández H, Fdez-Riverola F, Pisano DG, Glez-Peña D. 2018 Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics* **34**, 1414–1415. (doi:10.1093/bioinformatics/btx778)

124. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. 2016 Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* **5**, 1–27. (doi:10.7554/eLife.20777)

125. Daron J, Slotkin RK. 2017 EpiTEome: simultaneous detection of transposable element insertion sites and their DNA methylation levels. *Genome Biol.* **18**, 1–10. (doi:10.1186/s13059-017-1232-0)

126. Shukla R *et al.* 2013 Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111. (doi:10.1016/j.cell.2013.02.032)

127. Noshay JM *et al.* 2019 Monitoring the interplay between transposable element families and DNA methylation in maize. *PLoS Genet.* 15, e1008291. 1–25. (doi:10.1371/journal.pgen.1008291)

128. Tanay A, Regev A. 2017 Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338. (doi:10.1038/nature21350)

129. Göke J, Lu X, Chan Y, Ng H, Ly L-H, Sachs F, Szczerbinska I. 2015 Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135–141. (doi:10.1016/j.stem.2015.01.005)

130. Boroviak T *et al.* 2018 Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development* **145**, 1–18. (doi:10.1242/dev.167833)

131. Ziegenhain C *et al.* 2017 Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643. (doi:10.1016/j.molcel.2017.01.023)

132. Guo J *et al.* 2018 The adult human testis transcriptional cell atlas. *Cell Res.* **28**, 1141. (doi:10.1038/s41422-018-0099-2)

133. Jaitin DA *et al.* 2014 Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779. (doi:10.1126/science.1247651)

134. Macosko EZ *et al.* 2015 Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214. (doi:10.1016/j.cell.2015.05.002)

135. Hashimshony T *et al.* 2016 CEL-Seq2: sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* **17**, 1–7. (doi:10.1186/s13059-016-0938-8)

136. Nakamura T *et al.* 2015 SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Res.* **43**, 1–17. (doi:10.1093/nar/gkv134)

137. Cao J *et al.* 2017 Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667. (doi:10.1126/science.aam8940)

138. Manno GL *et al.* 2018 RNA velocity of single cells. *Nature* **560**, 494–498. (doi:10.1038/s41586-018-0414-6)

139. Ding J *et al.* 2019 Systematic comparative analysis of single cell RNA-sequencing methods. *bioRxiv*. (doi:10.1101/632216)

140. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J, Lönnerberg P, Linnarsson S. 2012 Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828. (doi:10.1038/nprot.2012.022)

141. Cole C, Byrne A, Beaudin AE, Forsberg EC, Vollmers C. 2018 Tn5Prime, a Tn5 based 5′ capture method for single cell RNA-seq. *Nucleic Acids Res.* **46**, 1–12. (doi:10.1093/nar/gky182)

142. Karlsson K, Lönnerberg P, Linnarsson S. 2017 Alternative TSSs are co-regulated in single cells in the mouse brain. *Mol. Syst. Biol.* **13**, 1–10. (doi:10.15252/msb.20167374)

143. Brocks D, Chomsky E, Mukamel Z, Lifshitz A, Tanay A. 2018 Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression. *bioRxiv*. (doi:10.1101/462853)

144. Buenrostro J, Wu B, Chang H, Greenleaf W. 2015 ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 1–10. (doi:10.1002/0471142727.mb2129s109)

145. Lieberman-Aiden E *et al.* 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–294. (doi:10.1126/science.1181369)

146. Zheng Y, Ay F, Keles S. 2019 Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. *Elife* **8**, e38070. (doi:10.7554/elife.38070)

147. Kruse K, Díaz N, Enriquez-Gasca R, Gaume X, Torres-Padilla M-E, Vaquerizas JM. 2019 Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *bioRxiv* 1–28. (doi:10.1101/523712)

148. Rodriguez-Terrones D *et al.* 2018 A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119. (doi:10.1038/s41588-017-0016-5)