COMPUTATIONAL TOOLS

Computational tools to unmask transposable elements

Patricia Goerner-Potvin[1] *and Guillaume Bourque*[1,2,3]*

Abstract | A substantial proportion of the genome of many species is derived from transposable elements (TEs). Moreover, through various self-copying mechanisms, TEs continue to proliferate in the genomes of most species. TEs have contributed numerous regulatory, transcript and protein innovations and have also been linked to disease. However, notwithstanding their demonstrated impact, many genomic studies still exclude them because their repetitive nature results in various analytical complexities. Fortunately, a growing array of methods and software tools are being developed to cater for them. This Review presents a summary of computational resources for TEs and highlights some of the challenges and remaining gaps to perform comprehensive genomic analyses that do not simply 'mask' repeats.

TE annotation
Assembled genomes are annotated to indicate which sequences are derived from transposable elements (TEs). The annotation reveals which families of TEs are present as well as the percentage of TE-derived sequences in a genome.

Repression mechanisms
Active transposable elements contain promoters that can initiate transcription. They are 'silenced' through various repression mechanisms to prevent transcription and further mobilization.

[1]*Department of Human Genetics, McGill University, Montréal, Canada*

[2]*Canadian Centre for Computational Genomics, Montréal, Canada*

[3]*McGill University and Génome Québec Innovation Centre, Montréal, Canada*

**e-mail: guil.bourque@mcgill.ca*

https://doi.org/10.1038/s41576-018-0050-x

Discovered through pioneering work in maize by Barbara McClintock in 1948 (REFS[1,2]), transposable elements (TEs), also known as mobile elements, have since been detected in all plants and animals[3], along with various prokaryotic species[4]. TEs are generally categorized into two classes, on the basis of the intermediate substrate propagating insertions (RNA or DNA), and further split into families and subfamilies on the basis of various structural features[5] (BOX 1). TEs have been responsible for major genomic expansions, and TE-derived sequences constitute a large portion of most eukaryotic genomes[3], including approximately half of the human genome[6] and up to 95% of some plant genomes[7]. Furthermore, as TE annotation tools improve and new TE families and instances are discovered, some elements challenge existing nomenclature[8–10] and the proportion of genomes they occupy[11] (BOX 1).

Most TEs have accumulated mutations and truncation events, rendering them transposition incompetent. In addition, different species have evolved various repression mechanisms, including TE promoter methylation[12,13], to prevent further transposition events[14,15]. However, in most genomes, a number of families remain active and can generate new insertions, termed polymorphic insertions. For instance, the class I long interspersed nuclear element 1 (LINE-1; also known as L1) family is still active in humans and is also responsible for the mobilization events of a number of other families (class I Alu elements and SVA elements). It has been estimated that new germline insertions are as frequent as 1 in every 95 births for L1 (REF.[16]) and one in 21 births for Alu[17]. These LINE-mediated insertions can in turn disrupt genes at their site of integration, and already 124 such insertions have been associated with human diseases[18]. Similarly to the L1, Alu and SVA families in humans, many TE families remain active in other species, such as long terminal repeat (LTR) families in the genomes of most plants, flies and mice and DNA transposons in plants and non-mammalian animal species[19] (BOX 1). TE insertions can also occur in somatic cells and have been observed in plants, *Caenorhabditis elegans* and in mammalian genomes, leading to genomic mosaicism among cells of an individual[19–21]. In humans, somatic insertions of L1, Alu and SVA elements have been observed in neuronal and cancer cells[22–27].

Beyond their impact through transposition events, some TEs, or parts of their derived sequences, have been domesticated to perform various crucial cellular functions[28,29]. For instance, the regulatory elements they contain can be co-opted for necessary cis-regulation in host regulatory pathways[30]. In particular, in mammalian genomes, a large fraction of regulatory binding sites[31] and over half of the regions of open chromatin in the human genome[32] have been provided by TE-derived sequences. It should be noted, however, that only a small fraction of these are likely to have a functional impact on the host[33]. In the rest of the text, full-length TEs and TE-derived sequences will be collectively referred to as TEs. TEs can also create alternative transcripts[34] and have trans-regulation activities[35], such as regulation of transcript nuclear accumulation by Alu-enriched sequences[36]. Alternatively, the cis and trans activity of TEs can be deleterious, as seen for example with endogenous retrovirus (ERV)-driven oncogene expression in cancer[37]. Overall, multiple key human cellular processes have been affected by TEs, including pluripotency[38], placenta formation[39], X chromosome inactivation[40], the immune system[41] and cancer[42].

Box 1 | **Classification models and abundance of TEs in different species**

The main classes of transposable elements (TEs) date back from the initial efforts to classify mobile elements by Finnegan in 1989 (REF. [5]). In this classification proposition, TEs were separated into two main categories according to the RNA or DNA nature of the mobilization intermediate of the element. Additionally, autonomous TEs encode the necessary mobilization proteins, whereas non-autonomous TEs do not but rely instead on the transposition machinery of other elements. Class I elements, also referred to as RNA transposons, are retrotransposons that use RNA as an intermediate for their copy-and-paste mechanism. Class II elements are DNA transposons mobilizing in a cut-and-paste fashion. Class I elements are further divided into elements flanked with long terminal repeats (LTR TEs) and those without (non-LTR TEs). LTR TEs are eukaryotic elements that are present from *Saccharomyces cerevisiae*[145] to all plants and animals. They remain active in plants and account for up to 95% of some plant genomes[7]. LTR elements remain active in mice and fly genomes[19]; however, in humans, these elements are no longer active and occupy about 10% of the genome[6]. Class I non-LTR elements, which include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), are present in high copy numbers in the genomes of all amniotes with a few exceptions[146] and account for about 34% of the human genome[6]. Class II elements are far less populous, and DNA transposons are estimated to occupy only 3% of the human genome.

As more and more species were sequenced, species-specific TEs were discovered that did not always fit in this binary classification. Examples of these recently discovered elements in eukaryotes are Helitrons, Cryptons, Mavericks and DIRS elements. For instance, DIRwS1 elements were classified as LTRs, but as more instances of DIRS1 were discovered, their mechanism of propagation came to be understood as different from that of other LTRs as they use an internal recombinase similar to that of the bacteriophage-λ, suggesting that they incorporate new insertions through recombination[9]. Furthermore, bacterial DNA transposons using a DNA transposon circle as a mobilization intermediate are effectively propagating via a copy-and-paste mechanism without an RNA intermediate[10]. These new mechanisms challenge the original two-class system. In 2007, a unified classification system was proposed[8]. In this proposed model, class I and class II remain, as distinguished by the presence or absence of RNA intermediates, respectively. However, the class II transposons now have two subclasses that reflect the mobilization mechanism. The first subclass is for elements that move in the pre-established cut-and-paste mechanism, whereas the second subclass contains the Helitrons and Maverick elements that cut only a single strand of the donor TE. Further down the classification hierarchy, orders are introduced. In class I elements, for example, LTRs, DIRS1s, LINEs, Penelopes-like elements (PLEs) and SINEs all represent individual orders. Although comprehensive and logical, this new classification system has yet to become mainstream. Finally, other classification systems completely forgo the RNA–DNA binary class system[147].

Notwithstanding these important functions in normal and disease processes, TEs are often ignored or 'masked' in genomic studies because their repetitive nature makes them challenging to analyse, especially using short-read sequencing technologies. For instance, as reads from TEs are often ambiguously mapped, discarding multi-mapped reads could exclude them from downstream analyses. Additionally, genome assemblies from short reads often struggle to correctly place TEs, resulting in incomplete downstream annotation. Nonetheless, we are at an exciting time of technological advancement regarding TE detection and analysis. Indeed, reduced costs have led to an explosion of large-scale sequencing projects, in which TE polymorphisms and their impact can be studied. Additionally, novel long-read sequencing technologies are also emerging, which are reducing the complexity of TE detection and genome assembly. Capitalizing on these advances, numerous methods and software tools are being developed to facilitate the inclusion of TEs in genomic studies.

In this Review, we offer a comprehensive guide to bioinformatics tools that have been developed to detect and analyse TEs (FIG. 1). First, we introduce the various classification tools and databases available. Next, we focus on tools for the annotation of TEs in genomic sequences, which rely on both de novo and targeted approaches. After that, we describe the main strategies employed for polymorphic TE insertion detection and explore some key examples. We also present emerging methods being developed to characterize and predict the functional impact of TEs. Finally, we present a number of standard tools that can be customized to take into account TEs while performing genomic analyses and summarize some of the current challenges and remaining gaps when analysing TEs.

TE classification and repositories

TEs are grouped into two main classes, which are further divided into families and subfamilies (BOX 1). Information on TEs is catalogued into three types of repositories: TE-centric, genome-centric and polymorphism-centric. TE-centric repositories collect information about the consensus sequence associated with each TE family, genome-centric repositories catalogue all individual TE instances within a reference genome, and polymorphism-centric repositories contain insertions in individuals diverging from the annotated reference genome of that species.

TE-centric repositories. Focusing on the TEs themselves, these databases contain a consensus sequence for each family and subfamily. They are used for classification purposes, annotation of TEs in genomes, and by various other bioinformatics tools that require TE reference sequences. RepBase Update is the most popular consensus repository of mobile elements in eukaryotic genomes and aims to contain either a consensus sequence or a representative instance for each TE family[43] (TABLE 1). RepBase Update classifies TEs into three groups: DNA transposons, LTR retrotransposons and non-LTR retrotransposons. Dfam is a more recent eukaryotic TE-centric database, in which TE families are more formally defined and are gathered as multiple sequence alignments through hidden Markov models (HMMs)[44]. Dfam also facilitates the annotation of individual TE instances that are related to known TE families but that have accumulated mutations and become distant from the consensus sequence (see below). Both RepBase Update and Dfam have been used alongside RepeatMasker (see Related links) — a tool that identifies repetitive sequences by conducting genome-wide searches for sequences homologous to those present in the databases — for annotating the human genome and most other eukaryotic genomes[45].

Genome-centric repositories. Genome-centric databases catalogue individual TE instances annotated in reference genomes. They display TE diversity within genomes and within TE families as they contain all genetic variations and truncations introduced over time and through individual mobilization events. These TE catalogues also allow for more accurate TE queries by providing exact TE sequences instead of a merged consensus. Dfam is the only TE repository for individual

Polymorphic insertions
Individual transposable element instances that have not been fixed in a species genome and are present in some re-sequenced genomes but absent from others, such as the reference genome. Polymorphic insertions can be either germline or somatic.

Long interspersed nuclear element 1
(LINE-1; also known as L1). Autonomous class I transposons that encode reverse transcriptase, endonuclease and RNA-binding proteins that effectively mobilize RNA sequences and create novel insertions.

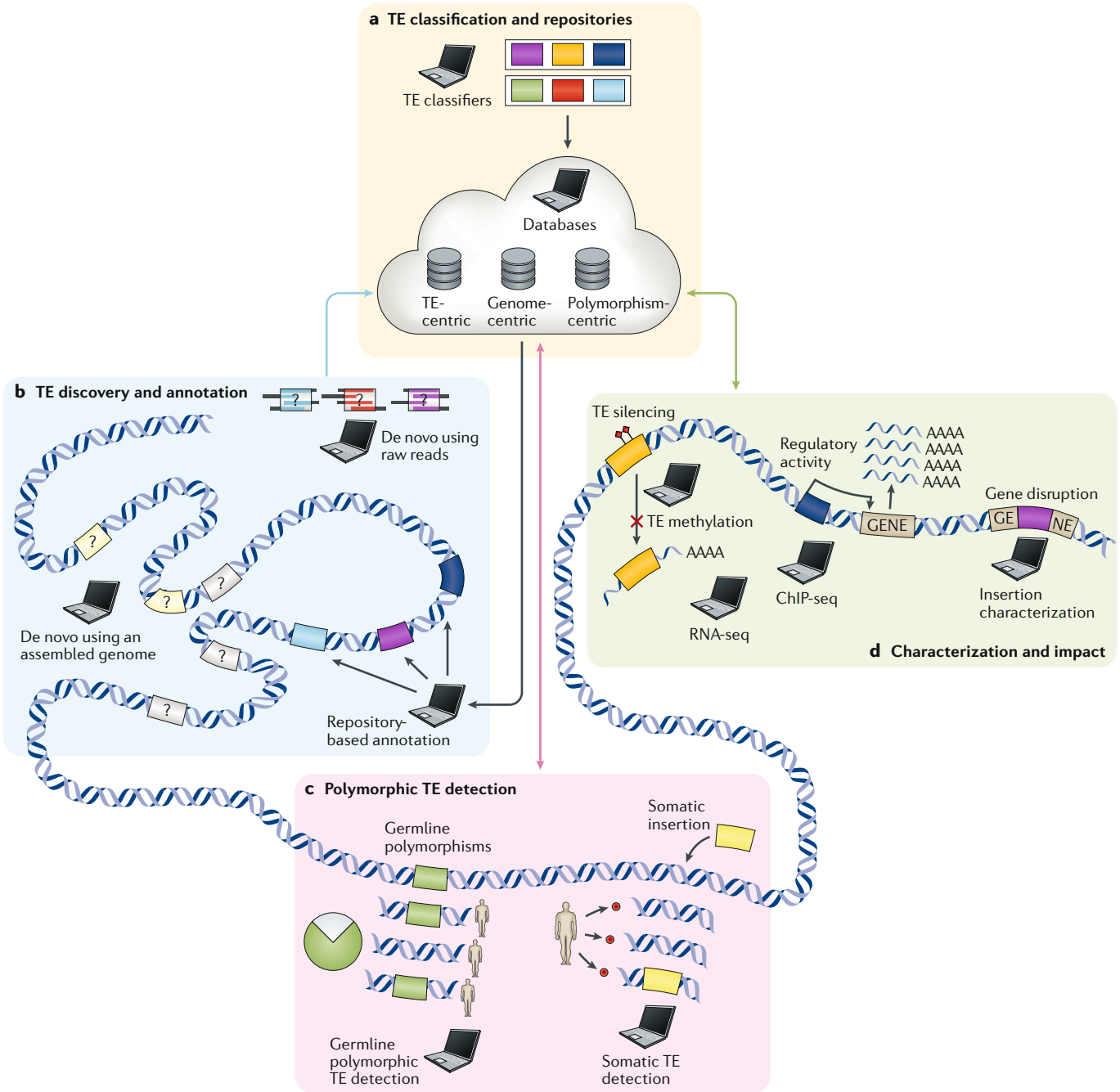

Fig. 1 | **Computational tools to analyse TEs.** This Review covers four broad categories of transposable element (TE)-specific computational tools. **a** | TE phylogeny can be established using TE classification software, and representative or consensus sequences of these TE families are stored in repositories. These databases can be either TE-centric, genome-centric or polymorphism-centric. **b** | TE discovery and annotation tools label TEs in assembled genomes either through sequence homology from TEs found in databases or de novo methods. Alternatively, TEs can be annotated de novo from raw reads. **c** | Polymorphic TE detection tools identify novel insertions that are absent from the reference genome; these can be germline insertions in different individuals of a population (green insertion, with population frequency represented in a pie chart) or somatic insertions within different cells of an individual (yellow insertion). **d** | TE characterization tools either provide information on novel TE insertions or assess the functional impact of genomic TEs through RNA sequencing (RNA-seq), chromatin immunoprecipitation followed by sequencing (ChIP-seq), methylation or DNA sequencing analysis.

TE instances annotated in mammalian genomes; however, these sequences and genomic locations are also available through the RepeatMasker tracks in genome browsers. By contrast, plants are the hosts of highly populous taxon-specific TEs and, in this context, a number of repositories have been developed specifically for them (TABLE 1). For instance, TREP[46] is a dual repository for plant and fungal TEs. TREP contains both a consensus

Table 1 | **TE repositories**

Repository	Host	TE class	Consensus	Entries		Characteristics
				Families	Insertions	
TE-centric						
RepBase[43]	Eukaryotes	All	√	>44,000	–	Main source for TE consensus sequence
Dfam[45]	Eukaryotes	All	HMM	4,150	>9 million	Family profiles to include distant TEs
GyDB[148]	Eukaryotes	LTRs	√ and HMM	314	413	Phylogenetic evolution study of viruses and TEs
SINEBase[149]	Eukaryotes	SINEs	√	217	–	SINEBank collection of consensus + search tools
Genome-centric						
RepeatMasker	Eukaryotes	All	–	1,348 in Hs	>5 million in Hs	Species annotations + TE protein database
TREP[46]	Plants and fungi	All	√	323	4,714	Used for TE classification by some annotation tools
RiTE[48]	Rice	All	√	124	>170,000	Specific to rice genomes
P-MITE[47]	Plants	MITEs	√	3,527	>2 million	One of the few resources for MITE elements
MASiVEdb[49]	Plants	Sireviruses	√	–	17,903	One of the few resources for Sireviruses
MnTEdb[150]	Mulberry	All	√	1,062	5,925	Database and tools for TEs in mulberry genomes
FishTEDB[151]	Fish	All	√	–	30,344	Consensus sequences of TE families found in fish
BmTEdb[152]	Silkworm	All	√	1,308	15,432	TE database and tools for TE families of silkworm
DPTEdb[153]	Dioecious plants	All	–	–	31,340	Sequences of TEs from eight plant species + browser
Transposon Registry[154]	Bacteria and archaea	Autonomous elements	√	–	–	Tn numbers to new TEs and encoded resistance
gEVE[155]	Mammals	Endogenous viral elements	HMM	–	774,172	Genome annotations + ORF and proteins
TranspoGene[50]	Eukaryotes	All	√	–	>1 million TEs in 14,783 human genes	Compiles TE insertions located in gene regions
HERVd[156]	Human (Hs)	HERVs	√	150	725,763	HERV repeats and families in human
HESAS[51]	Human (Hs)	HERVs	–	352	26,981 genes	HERV insertions in genes and expression data
LINE FUSION GENES[34]	Human (Hs)	LINEs	–	205	1,329 genes	LINEs in human genes and functional impact
RCPedia[112]	Human (Hs)	Retrocopies	–	–	–	TE-derived pseudogenes and expression data
Polymorphic-centric						
dbRip[54]	Human	Class I	–	36	3,605	Previous polymorphic database (not maintained)
euL1db[55]	Human	L1Hs	–	1	142,495	L1Hs-specific main repository for human
dbVar[157]	Human and mouse	All	–	–	517,506	Structural variants + 1000 Genomes TEs
TIDAL-Fly[158]	Fly	All	–	135	>300 genomes	TE landscape of fly genomes from TIDAL
PGSB PlantDB[52]	Plants	All (mostly LTRs)	√	97	23,420	Group of various plant TE databases now merged
BrassicaTED[53]	*Brassica* species	mTEs	–	41	–	Small TEs: SINEs, MITEs and TRIMs

HERV, human endogenous retrovirus; HMM, hidden Markov model; Hs, *Homo sapiens*; L1Hs, human-specific LINE-1 element; LINE, long interspersed nuclear element; LTR, long terminal repeat; MITE, miniature inverted repeat TE; mTEs, miniature TEs; ORF, open reading frame; SINE, short interspersed nuclear element; TE, transposable element; TIDAL, Transposon Insertion and Depletion Analyser; Tn, transposon; TRIM, terminal repeat retrotransposons in miniature.

repository (nrTREP), with at most two sequences per subgroup, and a complete repository of individual TE insertions (total_TREP). In addition, this repository also provides a database of hypothetical proteins derived from the insertion sequences (PTREP). P-MITE[47] is a plant-specific set of databases, with MITErepdb used for consensus sequences and MITEdb used for annotated miniature inverted repeat TE (MITE) instances in 41 plant genomes[47]. RiTE[48] is specialized for rice and related genomes, hosting full TE sequences obtained from genome sequencing, rebuilt consensus sequences and individual reference insertions[48]. Similarly, MASiVEdb[49] is a plant-specific database that contains annotated insertions of Sireviruses along with characterization information such as age of insertions[49]. Finally, a few additional databases catalogue TEs that are inserted in coding regions of genomes — these include TranspoGene[50], HESAS[51] and LINE FUSION GENES[34].

Polymorphism-centric repositories. Germline and somatic polymorphic insertions detected in individuals but absent from the reference genome are reported in these dedicated databases. As more individuals are sequenced and new insertions are discovered, the population frequency of these insertions can be determined, which helps to assess the validity of novel observations. In addition, these databases offer a larger pool of individual TE instances to explore TE diversity. Polymorphic repositories can also help associate TEs with different phenotypes, and some of these databases also report putative functional impacts of insertions. The databases are all host-specific and are often created in the context of large resequencing projects that aim to profile the diversity of a given species. For instance, PGSB PlantDB is one such repository that is composed of the Repeat Element Database (PGSB-Redat) and Catalogue (PGSB-Recat)[52] (TABLE 1). PGSB PlantDB combines TREP with the previous database TIGR repeats (now defunct), tandem repeats database PlantSat and GenBank, along with polymorphic LTR insertions. Similarly, BrassicaTED hosts miniature TEs (mTEs) including short interspersed nuclear elements (SINEs), MITEs and terminal repeat retrotransposons in miniature (TRIM) reference and polymorphic insertions in *Brassica* species[53]. Genes located at TE insertion sites are also included in the database.

Two major polymorphism-centric databases exist for the human genome. The dbRip repository includes polymorphic instances of any active retrotransposons organized by genomic loci[54]. The dbRip repository currently contains 3,605 elements, including 800 L1 insertions, but has not been updated since 2012. In 2014, the euL1db was developed specifically for cataloguing L1 non-reference insertions in human genomes and currently contains 142,495 entries[55]. euL1db hosts individual germline and somatic polymorphic L1 insertions reported per sample from healthy and cancer-affected donors but also merges them across samples in events called meta-retrotransposon insertion polymorphisms (MRIPs). Finally, whereas human-specific L1 (L1Hs) insertions from the 1000 Genomes Project are present in euL1db, polymorphic Alu, L1 and SVA were deposited in the National Center for Biotechnology Information (NCBI) structural variant database dbVar[56].

Challenges and remaining gaps. Two aspects of TE databases remain suboptimal. First, species-specific repositories are essential to account for the sequence diversity of TEs across organisms, but there is some overlap between databases, and merging repositories with shared host or TE type would be beneficial to avoid the need for multiple queries and increase cohesiveness. Second, there is a demand for an integrated resource for polymorphic TE findings in human genomes. Whereas human polymorphic L1 insertions are readily available through the dedicated euL1db repository, the broad structural variant repository dbVar remains the only option for other TE families. Given, for instance, the high rate of Alu germline insertions[17] and that somatic Alu insertions are found in large numbers alongside L1s in certain cancers[27], a resource that would enable the exploration of all TE polymorphisms, including non-L1 insertions, would be greatly beneficial.

Approaches for TE discovery and annotation

TE discovery and annotation are essential steps that can be performed with or without a genome assembly, and various tools exist to perform such tasks (TABLE 2). Two main strategies rely on assembled genomes: first is repository-based annotation, whereby sequences are queried against known TE consensus sequences or TE motifs; and second is de novo annotation. An alternative method that does not require a genome assembly is de novo annotation using raw reads (FIG. 2). Although repository-based RepeatMasker is currently the gold standard for TE annotation, de novo approaches offer the potential to identify novel TE families. As it is still useful to classify de novo elements, such approaches can be used as a first step ahead of a repository-based search to provide more comprehensive results, a feature that is already included in some de novo annotation tools to annotate their findings.

Repository-based annotation. The idea of repository-based annotation tools is to perform genome-wide searches of either TE consensus sequences or TE motifs associated with the different families of mobile elements. Their performance is linked to the quality and specificity of the databases used. The most widely known TE sequence database query tool is RepeatMasker, which queries against the RepBase Update and Dfam databases and provides an annotated list of interspersed repeats and low-complexity DNA for a wide range of eukaryotic organisms. This tool currently annotates 48% of the human genome as TEs (FIG. 2), and this percentage increases to 53% when coupled with Dfam2.0 (REF.[45]). RepeatMasker also provides tracks that are an integral part of the UCSC and other genome browsers. Tracks can be downloaded to provide the genomic locations and sequences of TE instances, and such data have been crucial for many TE-centric genomic analyses. Also produced by RepeatMasker is a modified version of the target sequence in which all TE regions have been replaced by Ns.

Alu
Primate-specific non-autonomous short interspersed nuclear element retrotransposon. Alus are highly abundant in primate genomes and can mobilize through the long interspersed nuclear element (LINE) retrotransposition machinery.

SVA
Primate-specific non-autonomous retrotransposons composed of fragments of Alus and retroviral long terminal repeat elements. The SVA name comes from the fact that they are derived from short interspersed nuclear elements, variable number tandem repeats (VNTRs) and Alu elements. They mobilize through long interspersed nuclear element (LINE) mobilization proteins.

Germline insertions
Transposable element insertions occurring in the parental germ line or during embryogenesis and shared between all cells of an individual.

Somatic insertions
Transposable element insertions occurring later in life in a specific tissue. These insertions are unique to one or a subset of cells of an individual.

Domesticated
(Also known as co-opted). A transposable element (TE) for which at least part of its sequence has been recruited to perform a specific function for the host, such as providing a TE-encoded protein with physiological functions. The co-opted sequence has been domesticated.

Cis-regulation
A transposable element modulating the expression of nearby genes by having part of its sequence acting as a regulatory element.

Trans-regulation
A transposable element modulating cellular processes distant from its genomic location. Trans-regulation is done via its transcript or encoded protein.

Cryptons
DNA transposons initially identified in fungi that are characterized by the use of tyrosine recombinase instead of transposase for transposition.

Table 2 | **TE discovery and annotation tools**

Tool	Method	TEs detected	TE classification	Genomes
Repository-based — using an assembled genome				
RepeatMasker	TE sequence query	All	RepBase + Dfam databases	Eukaryotes
CLARI-TE[57]	TE sequence query	All + nested TEs	ClariTERep database	Wheat
MASiVE[58]	TE motif query	Sirevirus	In-house annotation	Plants
HelitronScanner[59]	TE motif query	Helitrons	In-house annotation	Plants
LTR Annotator[60]	TE motif query	LTR	TREP + PGSB	Plants
MGEScan[61]	TE motif query	Class I	In-house annotation	Eukaryotes
LTRdigest[63]	TE motif query	LTR	In-house annotation	Eukaryotes
De novo — using an assembled genome				
RECON[65]	Multiple alignment clustering	All	RepeatMasker	Eukaryotes
RepeatScout[66]	Consensus seed clustering	All	RepeatMasker	Eukaryotes
RepeatModeler[67]	Pipeline using RECON and RepeatScout	All	RepeatMasker	Eukaryotes
phRAIDER[68]	PatternHunter spaced seeds	All	–	Eukaryotes
Red[69]	Machine learning	All + simple repeats	–	Bacteria + eukaryotes
P-Clouds[11]	Oligonucleotide clustering	All	In-house annotation	Large eukaryotic genomes
TEdenovo[71]	Multiple alignment clustering	All	PASTEC	Eukaryotes
detectMITE[159]	Complex-number-based numeric calculation	MITEs	CD-HIT clustering	Eukaryotes
De novo — using raw reads				
RepeatExplorer[73]	Quantitative assembly	All	In-house annotation + RepeatMasker	Eukaryotes
dnaPipeTE[74]	Quantitative assembly	All	RepeatMasker	Eukaryotes
Tedna[75]	Assembly	All	–	Eukaryotes
RepARK[76]	Assembly	All	–	Eukaryotes
REPdenovo[77]	Assembly	All (long repeats)	–	Eukaryotes
RepLong[78]	Pairwise alignment	All (long repeats)	RepBase + RepARK library	Eukaryotes

CD-HIT, Cluster Database at High Identity with Tolerance; LTR, long terminal repeat; MITEs, miniature inverted repeat TEs; PASTEC, Pseudo Agent System for Transposable Element Classification; TE, transposable element.

This can be an effective strategy to bypass analytical issues associated with repetitive sequences (see below) but effectively 'masks' repeats from downstream analyses. Other tools also exist for specific genomes or TE families. Such software include CLARI-TE[57], which was recently developed to annotate complex genomes of plants, including annotating frequent nested repeats in the wheat genome.

Additional tools in this category search for known motifs or genomic structures linked to particular TE families. They include MASiVE[58], HelitronScanner[59] and LTR Annotator[60], which are able to detect Sirevirus, Helitrons and LTRs, respectively, in plant genomes (TABLE 2). Similarly, for eukaryotic genomes, MGEScan[61] is a Galaxy-based[62] set of tools to detect both non-LTR and LTR genomic instances on the basis of the protein domain and structure of the elements, and LTRdigest[63] is an annotation tool that classifies LTRs according to internal sequence structure. TE database annotation tools also exist to classify pre-identified TEs. For instance, LTRclassifier[64] was built to address LTR classification complexities in plant genomes and classifies LTRs according to the similarity of HMM motifs, DNA or protein motifs. The strength of LTRclassifier lies in its added features for functionality prediction.

De novo annotation using assembled genomes. In the past, some of the most popular de novo annotation tools for assembled genomes have been RECON[65], RepeatScout[66] and RepeatModeler[67], which use either pairwise similarity or consensus seeds as starting points to cluster repetitive sequences. Built on the same principles as RepeatScout, the recent phRAIDER tool is reportedly ten times faster[68], whereas Red is a machine-learning detection tool that enables the identification of TEs and of simple repeats[69]. These tools have been applied to annotate newly sequenced genomes, as well as to identify missing TEs from RepeatMasker genome annotations. In fact, human genome annotation with these de novo methods often results in an additional 10% or more being assigned to TEs[11,69] (FIG. 2). This probably includes false positives but also novel TE families and TE instances that have diverged from their family consensus beyond recognition. The tool in this category that finds the highest percentage of repetitive sequences in the human genome is P-clouds[11], which is used to scan

Mavericks
Recently identified eukaryotic large DNA transposons (also known as Polintons) encoding up to ten proteins, including some that are similar to virus capsid.

Multi-mapped reads
Sequencing reads that map ambiguously to more than one location on the reference genome. These are common for repetitive regions including transposable elements.

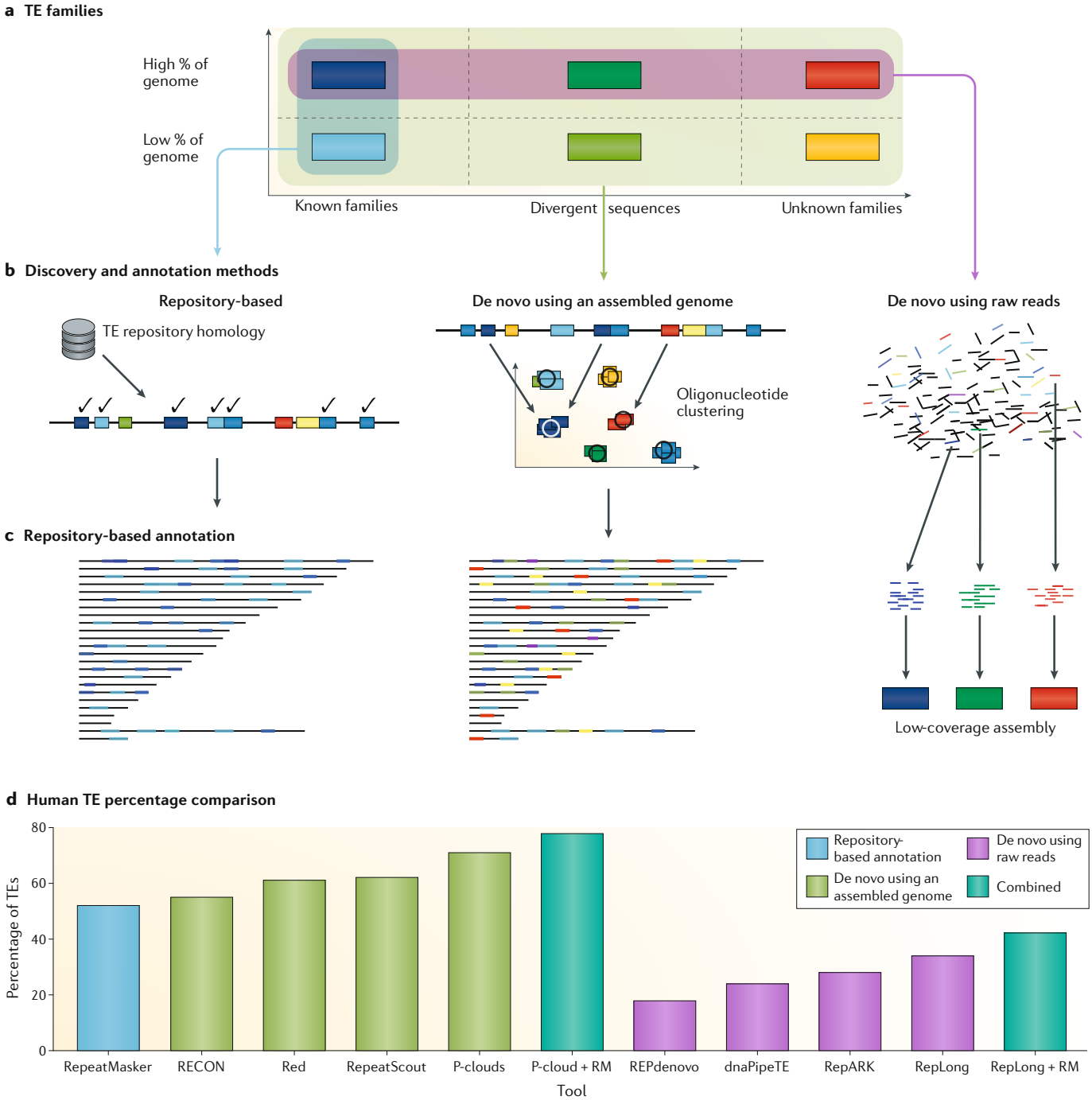

Fig. 2 | **Discovery and annotation of TEs and repeats in genomes. a** | Transposable elements (TEs) are shown according to their sequence matching — known families (a known TE family consensus sequence; left), divergent sequences (TE instances diverging extensively from a known consensus sequence, middle) or unknown families (TE families not present in repositories; right) — and their frequency in the genome (high, top; or low, bottom). **b** | Three main approaches for TE annotation. In repository-based annotation, an assembled genome is compared with TE consensus sequences from repositories; in de novo methods, oligonucleotides from assembled genomes are clustered by sequence similarity (some divergent TE sequences shown in green can cluster with known families); and in de novo methods for raw reads, sequencing reads are directly assembled into TE contigs. **c** | Resulting annotations of each method. Repository-based tools detect sequences that are homologous to consensus sequences of known TE families, excluding instances that became too divergent, and return TE annotation of genomes. De novo methods using assembled genomes detect known or unknown TEs and can also return TE annotation of genomes (although they require TE repositories to classify discovered TEs into families). De novo TE assembly using raw reads detects TEs on the basis of their genomic frequency and returns an assembly of TEs without genome annotation. **d** | Plot of the percentage of the human genome annotated as TEs by various methods. Percentages are either self-reported or generated by tool comparison papers[11,45,69,74,78]. '+RM' indicates a second run was conducted by the tool author to capture additional TEs with RepeatMasker.

Long-read sequencing
Can be achieved by directly sequencing long DNA molecules, such as by using Pacific Biosciences or Oxford Nanopore Technologies platforms. Alternatively, linked-read sequencing of 10X Genomics generates synthetic long reads by barcoding long molecules of DNA and sequencing interspersed short fragments each retaining the originating long molecule barcode, effectively linking these short reads into longer contigs.

Consensus sequence
Nucleotide sequence representing an approximation of the active transposable element (TE) that gave rise to a group of interspersed repeats. They are generated from a multiple alignment of instances from the same TE family that have accumulated mutations over time.

Miniature inverted repeat TE
(MITE). A recently coined name for non-autonomous short terminal inverted repeat DNA transposons.

Short interspersed nuclear elements
(SINEs). Non-autonomous elements for which their propagation is dependent on the retrotransposition machinery of long interspersed nuclear elements (LINEs) in the same genome. They contain an internal RNA polymerase III promoter derived from a small RNA gene, usually a tRNA.

Nested repeats
Transposable elements (TEs) that inserted in or near previous TE insertions. These are very challenging to detect with short reads.

Terminal inverted repeats
(TIRs). Repeated sequences that are present in the terminal regions of various transposable elements (TEs) are specific for particular TE families. These motifs contain transposase and DNA binding sites that are essential for transposition of the TE.

large eukaryotic genomes for the presence of TEs and other repetitive sequences on the basis of a probability approach that groups highly repeated oligonucleotides. Elements are then annotated using an in-house TE database. P-clouds attributes 66–69% of the human genome to mobile elements, an estimate that includes a self-reported 12.6% false discovery rate (FDR). Notably, the FDR of P-clouds has also been reported to potentially be as high as 54.0%[70]. By contrast, TEdenovo of the REPET suite uses a more classical multiple alignment and clustering method to provide comprehensive TE annotation[71]. See TABLE 2 for additional examples.

De novo annotation using raw reads. Tools in this category capitalize on the high abundance of TE sequences compared with other genomic sequences and use low-coverage sequencing data to assemble TEs and other repetitive sequences directly from raw reads. Starting this trend, RepeatExplorer[72] was initially published as a graph-based clustering algorithm to identify TEs from sequencing reads from the 454 pyrosequencing platform, which are then annotated per cluster with RepeatMasker. A later version of RepeatExplorer was released with extended features and deployed on a Galaxy server[62,73]. Following the same principle, dnaPipeTE was developed for next-generation sequencing (NGS) short reads using the Trinity assembler. Similar to RepeatExplorer, dnaPipeTE generates quantitative annotation of TEs but also provides the age of TE instances through read-contig comparison[74]. Other tools of this type include Tedna[75] and RepARK[76]; however, these two tools lack genomic quantification of TEs. Finally, REPdenovo, which also assembles TEs from raw reads, was developed partly to address the assembly of long repeats and identified missing reference TE instances in the human genome[77]. Together, these TE assembly tools represent a great opportunity to annotate TEs in newly sequenced genomes that lack a reliable corresponding reference genome. Although less sensitive than tools based on repository-based annotation or de novo annotation of assembled genomes, these raw-read annotation tools are able to assign 20–30% of the human genome to repeats and include TE instances that are absent from RepBase[78]. Finally, RepLong was recently published as the first long-read de novo TE annotation tool[78]. In a method similar to RepeatExplorer, repeat clusters are built from networks of read-pair comparisons, from which a representative read of each cluster is then used to identify the community networks. Capitalizing on long reads, RepLong is currently the most sensitive TE assembly tool and assigns 35% of the human genome to repeats (FIG. 2).

Challenges and remaining gaps. Determining the fraction of a genome that is repetitive is challenging as no ground truth exists. A previous generation of annotation tools were reviewed in 2010 and it was deemed necessary to run multiple tools in a complementary manner in order to get exhaustive annotations[79]. For example, the proportion of the human genome that is classified as repetitive varies depending on the tools used (FIG. 2). Although de novo tools have the potential to identify new repeat families and report missing TE instances, they are also more prone to false positives[70]. Methods relying on only raw reads return lower repeat content in the human genome[78], but they have the advantage of being able to annotate TEs in species lacking an assembled genome. A recent comparative study of TE annotation tools also highlighted missing TE instances from current genome annotations[80]. All of this points to the need for systematic benchmarking of annotation tools.

Tools to detect TE polymorphisms

Once a reference genome is available and its mobile elements are annotated, TE polymorphisms (that is, insertions and deletions of elements compared with the reference genome) can be detected by sequencing different individuals (FIG. 1). Similar to single-nucleotide polymorphisms (SNPs), such polymorphic insertions can be associated with different phenotypes, and some have already been linked to diseases such as haemophilia[81] and Rett syndrome[82,83]. Numerous software tools have been developed to detect both germline and somatic TE insertions (TABLE 3). These tools utilize various strategies to circumvent the challenges posed by the abundance and homology of TEs, their length and their complex mechanisms of replication.

Short-read approaches for germline TE polymorphism detection. Most TE detection workflows use NGS data as input. NGS reads are typically 100–250 nucleotides in length and can be single or paired-end reads. Finding novel insertions using these short reads is a challenge that is typically addressed by one, or a combination, of three strategies: using split-read (SR) information, relying on discordant read pair (DRP) alignment or identification of TE-specific motifs (FIG. 3). The SR methods detect reads containing breakpoints overlapping insertion sites, where one segment of the read aligns to the reference and the rest contains the beginning or the end of a TE sequence. The DRP strategy looks at paired-end reads and flags instances where the two ends align far from each other, are in opposite orientations or where only one read aligns on the reference genome. DRP and SR analyses can be done as a first detection step or on the preselected TE mapping reads (for example, NGS TE Mapper[84]). SR tools are expected to provide a higher positional accuracy because the junction between the TE and host sequence is directly identified, whereas DRP tools are expected to have a higher sensitivity with more reads available to support the insertion. This expectation has been confirmed by a workflow integrating six previously published detection tools[85]. Tools combining the SR and DRP approaches mostly start with DRP to identify sites and refine junction position with SR, whereas a few detection tools truly combine both approaches (TABLE 3).

A third strategy used by TE polymorphic insertion detection tools is to look for known TE motifs in sequencing reads as opposed to alignment characteristics used by the SR and DRP methods. Different motifs can be associated with TEs, such as LTRs at either end of LTR elements, or terminal inverted repeats (TIRs) present in DIRS and class II elements. Transposition

Table 3 | **Polymorphic TE detection tools**

Tools	Input	TE detected	Method			Characteristics
			DRP	SR	Motifs	
Germline short-read tools						
RetroSeq[160]	Pre-aligned reads	Class 1	√	√	–	Low memory requirement
VariationHunter-CR[161]	–	Non-LTR	√	–	–	Accounts for diploidy
TE-Tracker[162]	–	All	√	–	–	Identifies source elements
Mobster[163]	Pre-aligned reads	All	√	–	–	Requires alignment with MOSAIK
TIF[88]	–	TSD	–	√	√	Identifies TEs by their TSD
splitreader[164]	Pre-aligned reads	All	–	√	√	TE detection in *Arabidopsis*
Alu-detect[165]	Pre-aligned reads	Alu	√	√	√	Specific to Alus
TIGER[166]	–	L1	√	–	√	Identifies L1s and transduction events
HelitronFinder[167]	–	Helitron (HelA)	–	–	√	Specific to Helitrons
ITIS[128]	–	All	√	√	–	Reports zygosity
Short reads — pooled population						
NGS TE Mapper[84]	–	TEs with TSD	–	√	√	Overall target sites
TE-locate[91]	–	All	√	–	–	Needs TE database or annotation
TEMP[92]	–	All	√	√	–	Reports frequency of insertions
T-Lex2[168]	–	All	–	√	√	Pooled or individual samples
Pool-seq[169]	–	All	√	–	–	TE detection in flies
PoPoolationTE2[170]	Pre-aligned on masked genome	All	√	–	–	TEs across population or different tissues
TIDAL[158]	Single reads	All	–	√	–	TE detection in flies
ME-Scan[100]	Capture	AluYb8/9	–	√	–	Capture-based method amplifying Alu junction from pooled samples
Short reads — individuals in population						
Melt[56]	–	Non-LTR	√	√	–	1000 Genomes tools
Tangram[90]	Pre-aligned with MOSAIK	Class 1	√	√	–	Needs specific alignment
pecnv teclust[171]	–	All	√	–	–	TE detection in flies
RelocaTE2[172]	–	All	–	√	√	Detects TSDs
Somatic short-read tools						
Jitterbug[95]	Tumour option	All	√	√	–	Pairwise tumour–normal comparison of insertions
TranspoSeq[94]	Tumour	Non-LTR	√	√	√	
TranspoSeq-Exome[94]	Tumour	Non-LTR	√	√	√	
TraFiC[93]	Tumour	Transduction	–	–	√	
Short-read capture tools						
RC-seq[26]	Capture	Somatic	√	–	–	Tools for specific wet-laboratory experiment data sets
L1-seq[16]	Capture	L1 + somatic	–	–	–	
Tip-seq[97]	Capture	L1 + somatic	√	√	–	
SLAV-seq[99]	Capture	L1 + somatic	√	–	–	
TE-NGDS[101]	Capture	L1Hs, AluYa5/8, and AluYb8/9	–	√	–	
SIMPLE[102]	Capture	Full L1s	–	–	–	
Long-read tools						
LoRTE[103]	PacBio long reads	All	–	–	–	TE detection in flies
SMRT-SV[104]	PacBio long reads	SVs	–	–	–	SVs including TEs
SNIFFLES[105]	PacBio long reads	SVs	–	–	–	SVs including TEs
Other tools						
epiTEome[129]	MethylC-Seq	All + methylation	√	√	√	Assess methylation
TEPID[173]	Populations	All + methylation	√	√	–	Assess methylation
McClintock[85]	–	All	Multiple DRP and SR tools			Pipeline of many TE detection tools

DRP, discordant read pair; L1, long interspersed nuclear element; L1Hs, human-specific L1 element; LTR, long terminal repeat; PacBio, Pacific Biosciences; SR, split read; SVs, structural variants; TE, transposable element; TIDAL, Transposon Insertion and Depletion Analyser; TSD, target site duplication.

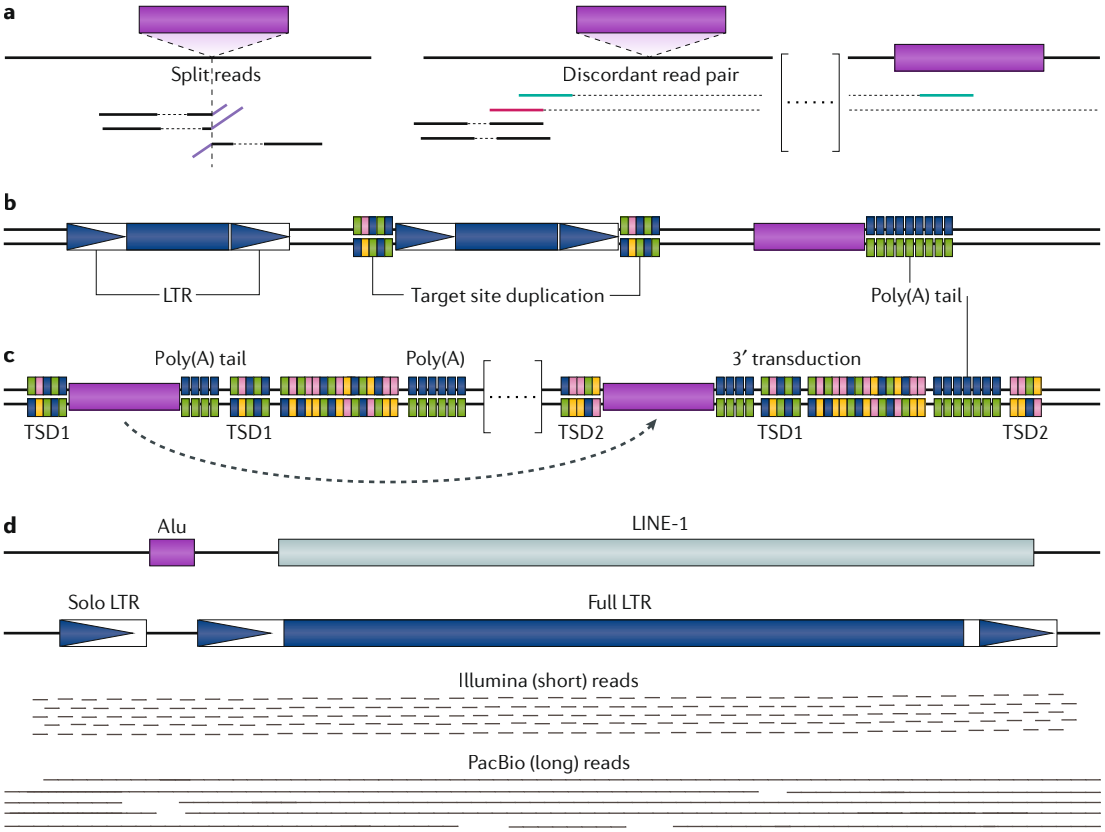

Fig. 3 | **Detection of polymorphic TE insertions.** Short-read and long-read approaches for transposable element (TE) detection. **a** | Short-read evidence of TEs include split reads (SRs) and discordant read pairs (DRPs). SRs partially align to the reference (black) whereas the other portion of the read is soft-clipped (purple) as it maps to a TE that is absent from the reference. DRPs (shown in green and pink) have only one read aligning at the site of insertion. The other aligns on another TE (green) or does not map (absent second pink read). **b** | Long terminal repeats (LTRs), target site duplications (TSDs) and polyadenylation (poly(A)) tails are TE-associated motifs. TSDs are identical sequences flanking a TE. **c** | Example of a new mobilization event. Long interspersed nuclear element-1 (LINE-1)-associated TE transcription ends at poly(A) sites. Transcription can bypass the transcription termination site (and poly(A) tail) and continue on flanking genomic DNA. This combined TE–poly(A)–TSD–host–poly(A) sequence inserts at a new location, effectively copying genomic DNA in a transduction event. This highlights the complexity of TE signature detection. **d** | Short and long sequencing reads scaled down proportionally to human transposons. Accordingly, it is challenging to map short reads to individual TEs and even harder to detect novel insertions. Long reads such as those from Pacific Biosciences (PacBio) can be used to solve this problem as they can span entire TE instances.

mechanisms also differ between TE families, leading to family-specific insertion signatures. Three common insertion signatures are used by detection tools: the presence of target site duplications (TSDs) flanking most TE insertions, long stretches of adenosines known as poly(A) tails in L1-mediated retrotransposition[86], and 3′ transduction[87] in L1 insertions. TSD length varies between TE families. Class I LTRs and class II elements have fixed-length TSDs (for example, four to six nucleotides for LTRs), whereas they are absent in DIRS and of variable lengths in LINE and SINE insertions[8]. For example, the TE detection tools TIF[88] and NGS TE Mapper[84] directly search for reads mapping over TE ends and retain reads containing TSDs. An array of tools is also available for individual families of TEs (TABLE 3). Additionally, insertion signatures can be used to validate insertions discovered through other methods.

Using these various strategies, tools can detect polymorphic germline insertions between one individual and the reference genome. In turn, populations can be analysed either by making calls in each sample independently or as a whole by pooling the samples and making joint calls. For example, the 1000 Genomes Project[89] used Tangram[90] and later MELT[56], both of which offer options for polymorphic insertion detection in each individual of a large population through a combined SR and DRP approach. Population analysis at the individual level is useful to link novel TE insertions with phenotypes, or even just to speed up the analysis of multiple samples. By contrast, tools to look for the presence of TEs in pooled samples, such as TE-Locate[91] (a DRP tool) and TEMP[92] (a combined SR and DRP tool), aim to observe the movement of TEs within genomes by detecting the locations of novel insertion events across populations as a whole. Similarly, NGS TE Mapper[84] does not perform an exhaustive TE detection per individual but rather identifies overall target sites for insertions through a motif and SR approach.

DIRS
Dictyostelium intermediate repeat sequence (DIRS) are classified as a superfamily of long terminal repeat transposons in the RepBase database and as a distinct order and superfamily in the 2007 Wicker unified transposable element classification system.

Target site duplications
(TSDs). Occur at insertion sites of most transposable elements (TEs), where the host genomic sequence is duplicated surrounding the new TE instance. As the two DNA strands are not cleaved at the exact same location, a few bases in between the two cuts will become duplicated during the second strand synthesis closing the insertion site.

Transduction
Host genomic DNA that is transcribed and inserted elsewhere in the genome through transposable element (TE) retrotransposition events. These duplicated sequences can be found with or without adjacent TE sequences as TE reverse transcription is often prematurely stopped.

SMRT
A PCR-free, single-molecule real-time (SMRT) sequencing platform from Pacific Biosciences that produces long reads. Reads are 1–60 kb in length, with a median of 10 kb.

Short-read tools for somatic TE detection. Numerous software tools have been developed to detect somatic TE insertions specifically in tumour samples, such as TraFiC[93], TranspoSeq[94], TranspoSeq-Exome[94] and Jitterbug[95]. These tools need to take into account that the event might be found only in a subset of the cells and, in many cases, they will use a normal sample from the same individual as a control. TraFiC detects L1 insertions and 3′ or solo transduction events through a motif detection approach[94], whereas TranspoSeq and TranspoSeq-Exome combine motif, SR and DRP on cancer whole-genome sequencing (WGS) and whole-exome sequencing data[94]. Finally, Jitterbug uses an SR and DRP approach for detecting somatic L1 insertions in cancer, as well as polymorphic germline insertions in various species[95].

Capture-based short-read methods. The identification of TE polymorphisms can also be facilitated by experimental DNA capture protocols targeting specific TE sequences. Methods including L1-seq[96], RC-seq[26], and TIPseqHunter[97] use such a strategy and are able to reduce sequencing costs and enrich for DNA reads that are the most informative for TE detection. These methods can generate sufficient TE fragments and junction reads to detect insertions that are present in only a small number of cells[98] and have been used to identify somatic TE insertions in cancer (TIPseqHunter[97] and L1-seq[96]) and in neuronal cells (SLAV-seq[99] and RC-seq[26]). Although the specific enrichment protocol differs, most of these tools employ a similar strategy. L1-seq[96], SLAV-seq[99] and TIPseqHunter[97] amplify L1Hs 3′ ends and flanking nucleotides and are specific to this subfamily by relying on a short motif that distinguishes it from all older elements. Similarly, ME-Scan[100] amplifies AluYb8/9 3′ junction sites from pooled human samples, and the novel TE-NGS[101] tool amplifies the 3′ ends of L1Hs and AluYa5/8–AluYb8/9 elements in a single experiment. Alternatively, RC-seq[26] uses a microarray to capture either end of L1, Alu and SVA insertions, avoiding PCR amplification bias, and SIMPLE[102] primes the 5′ end of L1s to extend only full-length L1 to identify mobilization-competent instances. Subsequently, most tools use an SR and/or DRP approach to call TE insertions, whereas SLAV-seq and TIPseqHunter also integrate a machine-learning approach to increase accuracy. Compared with L1-seq, RC-seq identifies a larger number of L1 insertion sites but has fewer sites supported by many reads[98]. Whereas L1-seq, SLAV-seq, TIPseqHunter and TE-NGS only identify the junction sequence at the 3′ insertions, SIMPLE and RC-seq can also detect 5′ insertion junction sequences for full-length or heavily truncated L1s.

Long-read tools. Finally, an alternative to avoid the challenges of using short reads to map TE insertions is to take advantage of emerging long-read sequencing technologies (FIG. 3). Indeed, long reads can span entire TE insertions to facilitate their detection and have the potential to detect even more complex structural variation events such as nested and tandem insertions, duplications and inversions. LoRTE is an example of a polymorphic TE detection tool that relies on the Pacific Biosciences (PacBio) single-molecule real-time (SMRT) long-read sequencing platform[103]. This tool aligns long reads on a TE consensus sequence, searches for the flanking sequences of those reads against the reference genome and compares the internal sequences from flanking pairs in the same orientation and location back to the consensus sequence. Tested on 1,950 pooled sequenced fly genomes, LoRTE detected 7 TE deletions and 14 insertions. Compared with short-read tools, such long-read methods are better adapted to return the full sequence of novel TE insertions, which enables better functional downstream analysis and bioinformatic validation of the insertions. Other software tools that can be used with long reads for TE detection include the structural variant callers SMRT-SV[104] and SNIFFLES[105]. SMRT-SV identifies structural variants using existing local de novo assembly tools to generate a consensus at every variant locus[104], whereas SNIFFLES makes calls through SRs and the identification of regions that are high in mismatches or coverage discrepancies. SMRT-SV was developed and tested on two human haploid genomes, and >89% of the polymorphic TEs that were discovered were new relative to the 1000 Genomes Project SV catalogue[106]. This finding highlights the power of long reads to detect insertions that would have been missed by standard NGS approaches. A human diploid long-read assembly was also published the same year and detected 2,664 TE insertions, of which only 893 were shared with the previous study[107].

Challenges and remaining gaps. Short-read TE detection tools such as MELT currently represent the most convenient methods to detect TE insertions in existing data from WGS experiments. However, even with these various recent tools, TE insertion detection does not yet produce standard findings, with diverging calls among tools[108,109]. Recent benchmarking efforts agree that using multiple tools in a complementary manner is necessary to achieve better predictions[85,110]. SimulaTE is a simulation tool that was recently released to generate various types of TE insertions[111]. SimulaTE creates Illumina or PacBio sequences simulating TE insertions in population cohorts and could prove useful for benchmarking TE detection tools. Short-read capture methods, combining wet-laboratory TE enrichment and custom bioinformatics analysis such as RC-seq and TE-NGS, offer higher sensitivity but require the sequencing of targeted regions, resulting in data sets that do not represent the full spectrum of possible insertions. Finally, long-read TE detection tools have the potential to resolve more insertions and complex rearrangements; however, relevant data sets remain sparse.

TE characterization and impact

TE insertions can have direct consequences for the host genome such as providing novel genes or transcripts[38], modulating gene expression[30], generating genomic instability[99] or actively transposing[19] (FIG. 4). Here, we describe some of the tools that have been developed to look at the potential impact of both fixed and polymorphic TEs.

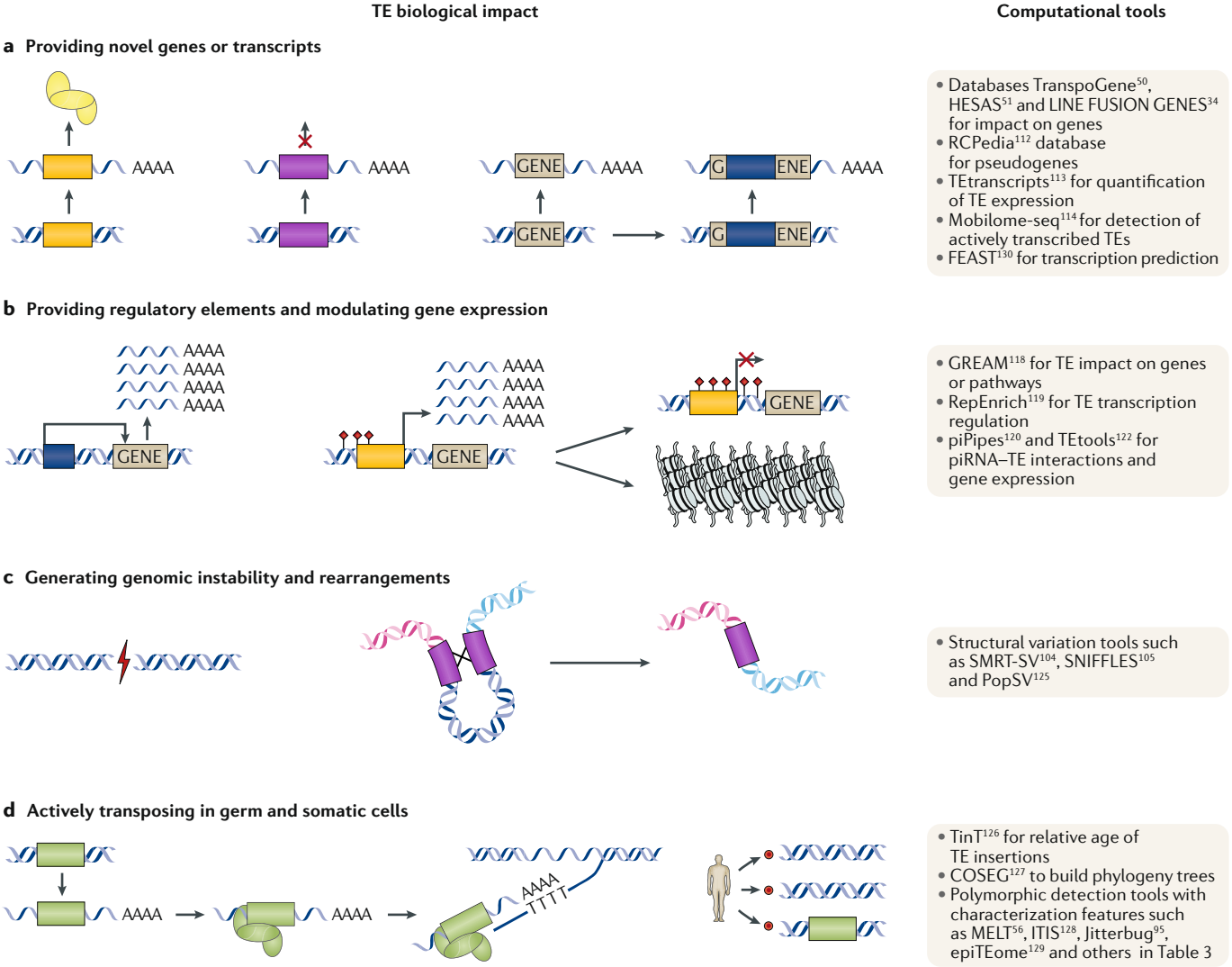

Fig. 4 | **Functional impacts of TEs.** Transposable elements (TEs) and TE-derived sequences impact host genomes through various mechanisms. Computational tools presented on the right can be used to assess the four impact category sections illustrated in parts **a**–**d**. **a** | Examples of TEs providing novel genes and transcripts. From left to right: protein-coding TE RNA, TE-derived non-coding RNA, and gene transcript altered after TE insertion. **b** | TEs provide regulatory elements and modulate gene expression. From left to right: examples of TE promoter cis-regulating gene expression, and methylation silencing of TEs leading to silencing of a nearby gene through spreading of DNA methylation or heterochromatin formation. **c** | Examples of TE-associated genomic instabilities. From left to right: double-stranded breaks and a recombination event leading to a large deletion. **d** | Example of TE mobilization. From left to right: TE mobilization through transcription, translation and reverse transcription, and somatic insertions leading to mosaicism. piRNA, PIWI-interacting RNA.

Predicting TE-derived and TE-disrupted genes and transcripts. TEs annotated within gene regions are reported in numerous dedicated databases (TABLE 1). These databases list affected genes and resulting transcripts to determine the potential impact of TE insertions in host genomes. In particular, TranspoGene is a database for TEs inserted in protein-coding genes of human, mouse, chicken, zebrafish, fruitfly, nematode and sea squirt genomes[50]. TranspoGene also contains information on affected genes and associated diseases. The same group also released microTranspoGene for the collection of TE-derived microRNAs[50]. Looking at the impact on gene transcripts, HESAS catalogues human endogenous retrovirus (HERV) insertions in gene regions and contains expression data[51]. LINE FUSION GENES is another database that records LINE insertions in human genes and annotates them on the basis of their impact as an alternative promoter, alternative polyadenylation signal or exonization[34]. Alternatively, translated TE machinery can also mobilize transcripts of genomic DNA, leading to retrocopies or pseudogenes. Human retrocopies are collected in the RCPedia database along with expression data from various species and tissues[112]. All of these impact-focused TE databases are important resources, but their usefulness depends on the number of findings uploaded to them.

Focusing on TE transcription, TEtranscripts[113] is a differential expression analysis software tool to handle ambiguously mapped RNA, which includes TE RNA reads. In a comparison against various bioinformatic

ChIP-seq
Chromatin immunoprecipitation followed by sequencing (ChIP-seq) consists of the capture and sequencing of DNA that is bound by a protein of interest, such as a transcription factor or modified histone.

RACE-seq
Rapid amplification of cDNA ends (RACE) is a method to amplify complete RNA molecules. RACE-seq involves sequencing the RNA molecules amplified through the RACE protocol. It is often used to detect novel transcripts.

CAGE-seq
Cap analysis of gene expression (CAGE) sequencing is a method to identify transcription start sites through sequencing of 5′ RNA transcripts.

PIWI-interacting RNA
(piRNA). piRNAs are short non-coding RNA molecules that bind to PIWI proteins. They are established as part of transposable element silencing mechanisms in animals.

tools for RNA sequencing (RNA-seq) data analysis, TEtranscripts was shown to be the most accurate for TE expression. Additionally, Mobilome-seq[114] can detect transcribed TEs in plant and potentially other eukaryotic genomes from sequenced extrachromosomal circular DNA (eccDNA) reads mapping to annotated TEs. Although eccDNAs have not been shown to reintegrate into chromosomal DNA, they can arise from extrachromosomal linear forms of active TEs[115].

Predicting the impact of TEs on gene regulation. Strategies used to identify impactful or functional TEs are quite diverse and include measuring overlap with other annotated genomic features, looking for signs of negative or positive selection[116,117], identifying over-represented transcription factor binding sites within their sequences[31,42] or looking at association with chromatin or transcription data sets[38,41]. For instance, the web-based tool GREAM was developed to identify TE instances harbouring a potential effect on a specific pathway or group of genes[118]. This tool outputs short-lists of suspect genomic repeats or TEs that are over-represented or under-represented near genes from a user-specified gene list.

Many TE-derived sequences are still transcribed and can affect the expression of nearby genes[30]. A few tools combine chromatin immunoprecipitation followed by sequencing (ChIP-seq) and RNA-seq data to link expression of TEs and genes with regulatory binding sites. For instance, RepEnrich was developed to study the transcriptional regulation of TEs and to identify active instances[119]. The tool extracts RNA polymerase I (Pol I) and Pol II binding sites specifically in class I TEs from ChIP-seq data and looks for TE enrichment in RNA-seq data[119]. Combining DNA sequencing, ChIP-seq, rapid amplification of cDNA ends sequencing (RACE-seq), cap analysis of gene expression sequencing (CAGE-seq), small RNA-seq and RNA-seq data sets, piPipes[120] was developed to facilitate and standardize combined analysis of TEs and PIWI-interacting RNA (piRNA) gene expression. In metazoans, piRNAs contribute to TE silencing through various mechanisms including direct targeting and degradation of TE transcripts[14] and are inherited through meiosis to help re-form heterochromatin[121]. piPipes includes independent tools to analyse each of these types of data sets, providing information on TE and gene expression, TE insertions, structural variation and piRNA-induced cleavage products, among others, for a comprehensive study of piRNA–TE interactions. Finally, TEtools[122] extends mRNA and small RNA analyses to unassembled reads and unannotated genomes, and it has been used to show linked expression between TEs and piRNA precursor genes.

Detection of structural variation in repetitive regions. The presence of TEs can also lead to genomic instability via, for instance, non-allelic homologous recombination[123]. However, mapping of short reads to TEs is often ambiguous, making structural variant calling in these regions more challenging. As an attempt to include multi-mapped short reads to detect copy number variants (CNVs) in repetitive regions, a study used a Poisson formula to account for all mapping locations of multi-mapped reads and to assign correct mapping[124]. This method was able to recover single CNVs in repetitive regions but was less efficient when multiple CNV events occurred in the same region. Another tool, PopSV, compares coverage at each locus to a set of reference samples and is able to make calls in low-mappability regions[125]. Notably, the authors found an enrichment of polymorphic CNVs in these regions. Alternatively, structural variants associated with TEs can also be detected through dedicated long-read tools such as SMRT-SV[104].

Characterization of novel insertions. Although very few TE instances remain capable of mobilization, numerous families are still active in various genomes[19]. Assessing the relative age of insertions (fixed or polymorphic), TinT[126] analyses TEs inserted within other TEs (nested TEs) and establishes the chronology of insertions. These data can be used to determine whether a family is still active in the host genome. Older and conserved insertions can also point towards host-beneficial functions. TinT uses annotated nested repeats from RepeatMasker as input and establishes the order of insertion events through orientation, position of the start and end of TEs and sequence interruption, among other criteria. Potentially active TEs can also be identified by building phylogenetic trees for subfamilies of TEs identified from de novo annotation tools or polymorphic TE detection tools, through programs such as COSEG from RepeatMasker (original code from REF.[127]), and investigating the youngest members.

Once polymorphic TE insertions are identified, the most important features are the presence of biological insertion signatures, the length of insertions, zygosity and surrounding genomic elements. In the context of a population, the frequency of individual insertions is also a key characteristic. Various insertion detection tools can provide some of this information (TABLE 3), and here we present some of the tools with the most characterization features. ITIS[128] and Jitterbug[95], for example, determine the zygosity of the insertions identified. Other tools go even further to predict the functional impacts of insertions. For example, MELT[56] not only detects de novo TE insertions but also characterizes each call in terms of position, orientation, classification, insertion signatures, length and so forth. The potential impact can also be assessed by looking at adjacent genes and reporting the precise gene features affected. Notably, to assess both the presence and the methylation state of mobile elements, epiTEome[129] is a tool that combines de novo TE detection and DNA methylation analysis using the same MethylC-seq data set. This information can help to determine which insertions have been silenced and which ones still have the ability to mobilize.

Challenges and remaining gaps. Looking at existing TE functional analysis tools, current methods can identify putative active TEs, assess silencing state through methylation or piRNA, link TE and gene expression and identify regulatory binding sites. Nevertheless, many tools focus on TE consensus sequences rather than individual TE instances, which sometimes limits

the interpretability of their results. Available software tools also do not cover the extent of TE impact, especially when looking at trans-regulatory activity of TEs. For instance, new approaches such as the transcription prediction tool FEAST[130] are needed for the annotation and prediction of actively transcribed TE-initiated long non-coding RNAs (lncRNAs). Finally, and most notably, when looking for impact, it is very important to distinguish between TEs that are designated as active on the basis of only a biochemical signature, such as active chromatin marks, versus TEs that have more robust evidence for functionality, as was discussed in publications around the Encyclopedia of DNA Elements (ENCODE) project and a recent review[33,131,132].

General genomic tools adapted for TE analysis

There is growing appreciation of the usefulness of genomic analyses that extend beyond coding regions. This trend represents an opportunity to further our understanding of TEs. We now present a number of general-purpose genomic tools that can be adapted for the analysis of repetitive sequences.

Alignment tools for low-mappability regions. Regions of low mappability include TEs but also other genomic elements such as simple repeats, satellite repeats and segmental duplications. Systematic discarding of multi-mapped reads associated with low-mappability regions results in biased genomic analyses that overlook potential candidate loci. The alignment step is the starting point of most genomic analyses, and inclusion of multi-mapped reads can have a considerable impact on results in terms of both increasing sensitivity and potentially raising the number of false positives. The popular index-based BowTie[133] and BowTie2 (REF. [134]) programs allow multi-mapping by reporting multiple hits per read using a quality-aware backtracking method. Multiple alignments for a given read are scored according to the number of mismatches through greedy selection for minimal quality values that represent sequencing errors. If there are more equivalent matches than specified, BowTie will return a random set of these mapping positions. These aligners are used by many TE tools; however, in a benchmark of repetitive region alignment, the commercial tool NovoAlign (see Related links) showed the most accuracy regardless of read length compared with four other tools, including BowTie[135]. NovoAlign is a hashing-based aligner that assigns mapping quality using posterior alignment probability and defines optimal mapping through dynamic adjustment of gap and mismatch penalties. In cases of ambiguity, similar to BowTie, NovoAlign will return a random subset of equivalent mapping positions.

Tools for using ambiguous reads in RNA-seq and ChIP-seq. The Multiple Mapper Resolution (MMR) tool[136] and the Gibbs sampling strategy of Wang et al.[137] are methods to resolve ambiguously mapped reads and assign them to unique positions, allowing them to be used in standard downstream analysis tools. MMR assigns unique tags to RNA-seq reads on the basis of mapping density of other reads, whereas the Wang et al. strategy uses an iterative approach to assign unique positions to multi-mapped ChIP-seq reads. In some cases, an active TE instance might account for a peak in transcription of a certain TE family, but the reads will get assigned in a normalized way across other TE instances from the same family. Information is returned about an active family overall, as the reads are no longer discarded, but it might not identify the precise instances that are active. Alternatively, numerous ChIP-seq and RNA-seq tools were designed specifically to be able to handle multi-mapped tags directly. For ChIP-seq data, MOSAiCs[138] uses a fraction scoring and weighted alignment to call peaks. By contrast, LONUT[139] first identifies peaks with unique tags and then looks at the distance of multi-mapped reads to these peaks and potential enrichment to decide where reads go, whereas DROMPA[140] divides ambiguous tags equally among locations. For RNA-seq, both Cufflinks[141] and HTSeq-count[142] have multi-mapped read modes. However, they were shown to provide less-accurate TE expression level estimations at the subfamily level (for example, L1Hs) than TEtranscripts[113]. Finally, recent tools such as Kallisto[143] and Salmon[144] completely forgo the use of aligned reads and perform their own pseudo-alignments to provide faster results. These pseudo-alignment steps can also score ambiguous reads, such that these tools will be able to report estimated expression abundance even for low-mappability regions.

Challenges and remaining gaps. Optimal methods to extend analyses into low-mappability regions are not yet established. A promising approach is the use of long-read sequencing technologies that can resolve the complexity of these regions and remove the analytical challenges associated with ambiguous reads. Genomic applications using these technologies have been mainly geared towards closing genome assemblies and detecting structural variation, both of which contribute to better annotation and detection of TEs, but more analytical tools to use them for different applications are also needed. In particular, it would be interesting to generate long-read transcriptome and chromatin data sets to better assess the performance of multi-mapped short-read approaches in repeat-rich regions.

Conclusions

TEs are diverse, abundant and active components of most genomes. Not only have TEs been successful at contributing a substantial proportion of the DNA of many species, they have also become an integral part of many cellular processes for their host. In some cases, especially when TEs become dysregulated, they can also be associated with disease. Nonetheless, genomic studies have often excluded them because their repetitive nature leads to various analytical challenges, especially when relying on short-read technologies. Thankfully, new long-read technologies and wide-ranging computational tools are being developed specifically for TEs. Through these, new TE families and instances are being discovered, and more cellular functions and diseases have now been shown to be associated with TEs. These discoveries solidify the need to systematically include TEs in standard genomic analyses to avoid missing important information.

There is still room for improvement in many areas of TE computation. One trend, which also applies to other types of genomic analyses, is to move towards ensemble approaches to address a given problem, such as TE insertion detection, by combining the strengths of multiple bioinformatics tools that use complementary strategies. These ensemble approaches improve sensitivity and enable a better control of false positives. Systematic TE benchmarking efforts will also be needed to resolve important questions such as in TE detection and annotation. Indeed, only a few de novo TE detection methods currently exist, but it would valuable to have further developments in this area, linked to systematic benchmarks, to improve this process. Additionally, it is important that TE findings are compiled in dedicated online repositories, and although there are many options, there is a lack of unified TE repositories that have a long-term sustainability plan. The need for this is especially acute for TE-associated human polymorphisms and mutations, particularly in the context of the millions of genomes currently being sequenced.

Finally, performing multifaceted and multi-omics analyses will be key to better understand the impact of fixed and polymorphic TEs in our genomes and to distinguish selfish behaviour from co-opted functions.

Published online: 19 September 2018

1. McClintock, B. Mutable loci in maize. *Carnegie Inst. Wash.* **47**, 155–169 (1948).
2. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36**, 344–355 (1950).
3. Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
4. Garrett, R. A., She, Q., Brügger, K., Faguy, D. & Redder, P. in Mobile DNA II (eds. Craig N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 1060–1073 (American Society of Microbiology, Washington, DC, 2002).
5. Finnegan, D. J. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**, 103–107 (1989).
6. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Kronmiller, B. A. & Wise, R. P. TEnest: automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol.* **146**, 45–59 (2008).
8. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
9. Goodwin, T. J. & Poulter, R. T. The DIRS1 group of retrotransposons. *Mol. Biol. Evol.* **18**, 2067–2082 (2001).
10. Duval-Valentin, G., Marty-Cointin, B. & Chandler, M. Requirement of IS911 replication before integration defines a new bacterial transposition pathway. *EMBO J.* **23**, 3897–3906 (2004).
11. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLOS Genet.* **7**, e1002384 (2011).
12. Hata, K. & Sakaki, Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227–234 (1997).
13. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).
14. Malone, C. D. & Hannon, G. J. Small RNAs as guardians of the genome. *Cell* **136**, 656–668 (2009).
15. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627 (2011).
16. Ewing, A. D. & Kazazian, H. H. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* **20**, 1262–1270 (2010).
17. Xing, J. et al. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009).
18. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).
19. Huang, C. R. L., Burns, K. H. & Boeke, J. D. Active transposition in genomes. *Annu. Rev. Genet.* **46**, 651–675 (2012).
20. Emmons, S. W. & Yesner, L. High-frequency excision of transposable element Tc 1 in the nematode Caenorhabditis elegans is limited to somatic cells. *Cell* **36**, 599–605 (1984).
21. Fernandez, L., Torregrosa, L., Segura, V., Bouquet, A. & Martinez-Zapater, J. M. Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. *Plant J.* **61**, 545–557 (2010).
22. Miki, Y. et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* **52**, 643–645 (1992).
23. van den Hurk, J. A. et al. L1 retrotransposition can occur early in human embryonic development. *Hum. Mol. Genet.* **16**, 1587–1592 (2007).
24. Muotri, A. R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
25. Coufal, N. G. et al. L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).
26. Baillie, J. K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).
This study is the first mapping of somatic retrotransposition events in the human brain and is performed with the capture-based polymorphic TE detection tool RC-seq.
27. Goodier, J. L. Retrotransposition in tumors and brains. *Mob. DNA* **5**, 11 (2014).
28. Volff, J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913–922 (2006).
29. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, aac7247 (2016).
30. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
31. Bourque, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
32. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLOS Genet.* **9**, e1003504 (2013).
33. Venuto, D. & Bourque, G. Identifying co-opted transposable elements using comparative epigenomics. *Dev. Growth Differ.* **60**, 53–62 (2018).
34. Kim, D.-S. et al. LINE FUSION GENES: a database of LINE expression in human genes. *BMC Genomics* **7**, 139 (2006).
35. Mariner, P. D. et al. Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* **29**, 499–509 (2008).
36. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).
37. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* **7**, 24 (2016).
38. Lu, X. et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21**, 423–425 (2014).
39. Naville, M. et al. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clin. Microbiol. Infect.* **22**, 312–323 (2016).
40. Lyon, M. F. Do LINEs have a role in X-chromosome inactivation? *J. Biomed. Biotechnol.* **2006**, 59746 (2006).
41. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
42. Wang, T. et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
43. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
This article presents the most comprehensive collection of TE consensus sequences from eukaryotic genomes, used with references 44 and 45 in RepeatMasker genome annotations.
44. Wheeler, T. J. et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2013).
45. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
References 44 and 45 present a eukaryotic TE consensus database with added HMM profiles used to improve genomic annotation of TEs.
46. Wicker, T., Matthews, D. E. & Keller, B. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.* **7**, 561–562 (2002).
47. Chen, J., Hu, Q., Zhang, Y., Lu, C. & Kuang, H. P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.* **42**, D1176–D1181 (2014).
48. Copetti, D. et al. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**, 538 (2015).
49. Bousios, A. et al. MASiVEdb: the sirevirus plant retrotransposon database. *BMC Genomics* **13**, 158 (2012).
50. Levy, A., Sela, N. & Ast, G. TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. *Nucleic Acids Res.* **36**, D47–D52 (2007).
51. Kim, T.-H., Jeon, Y.-J., Kim, W.-Y. & Kim, H.-S. HESAS: HERVs expression and structure analysis system. *Bioinformatics* **21**, 1699–1700 (2005).
52. Spannagl, M. et al. PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **44**, D1141–D1147 (2016).
This article presents a combination of multiple plant databases containing TE consensus sequences, annotated instances and polymorphic insertions.
53. Murukarthick, J. et al. BrassicaTED - a public database for utilization of miniature transposable elements in Brassica species. *BMC Res. Notes* **7**, 379 (2014).
54. Wang, J. et al. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.* **27**, 323–329 (2006).
55. Mir, A. A., Philippe, C. & Cristofari, G. euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res.* **43**, D43–D47 (2015).
The euL1db database contains the most comprehensive collection of polymorphic L1Hs insertions in human genomes.
56. Gardner, E. J. et al. The mobile element locator tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
This paper presents a great example of a polymorphic TE detection tool that also provides characterization of insertions, and it was used for the 1000 Genomes Project.

57. Daron, J. et al. Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.* **15**, 546 (2014).
58. Darzentas, N., Bousios, A., Apostolidou, V. & Tsaftaris, A. S. MASiVE: mapping and analysis of sirevirus elements in plant genome sequences. *Bioinformatics* **26**, 2452–2454 (2010).
59. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111**, 10263–10268 (2014).
60. You, F. M., Cloutier, S., Shan, Y. & Ragupathy, R. LTR annotator: automated identification and annotation of LTR retrotransposons in plant genomes. *IJBBB* **5**, 165–174 (2015).
61. Lee, H. et al. MGEScan: a Galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics* **32**, 2502–2504 (2016).
62. Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
63. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002–7013 (2009).
64. Monat, C., Tando, N., Tranchant-Dubreuil, C. & Sabot, F. LTRclassifier: a website for fast structural LTR retrotransposons classification in plants. *Mob Genet. Elements* **6**, e1241050 (2016).
65. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
66. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21** (Suppl. 1), 351–358 (2005).
67. Smit, A. & Hubley, R. RepeatModeler 1.0.11. *RepeatModeler* http://www.repeatmasker.org/RepeatModeler/ (2018).
68. Schaeffer, C. E., Figueroa, N. D., Liu, X. & Karro, J. E. phRAIDER: pattern-hunter based rapid ab initio detection of elementary repeats. *Bioinformatics* **32**, i209–i215 (2016).
69. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16**, 860 (2015).
70. Caballero, J., Smit, A. F. A., Hood, L. & Glusman, G. Realistic artificial DNA sequences as negative controls for computational genomics. *Nucleic Acids Res.* **42**, e99 (2014).
71. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLOS ONE* **6**, e16526 (2011).
72. Novák, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**, 378 (2010).
This paper presents the first method to discover TEs in unassembled sequencing reads, on which many recent tools are based.
73. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
74. Goubert, C. et al. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome Biol. Evol.* **7**, 1192–1205 (2015).
75. Zytnicki, M., Akhunov, E. & Quesneville, H. Tedna: a transposable element de novo assembler. *Bioinformatics* **30**, 2656–2658 (2014).
76. Koch, P., Platzer, M. & Downie, B. R. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Res.* **42**, e80 (2014).
77. Chu, C., Nielsen, R. & Wu, Y. REPdenovo: inferring de novo repeat motifs from short sequence reads. *PLOS ONE* **11**, e0150719 (2016).
78. Guo, R. et al. RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics* **34**, 1099–1107 (2018).
79. Lerat, E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**, 520–533 (2010).
This detailed review discusses bioinformatics tools for TE annotation and classification.
80. Hoen, D. R. et al. A call for benchmarking transposable element annotation methods. *Mob. DNA* **6**, 13 (2015).
81. Kazazian, H. H. et al. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
82. Yu, F., Zingler, N., Schumann, G. & Strätling, W. H. Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res.* **29**, 4493–4501 (2001).
83. Muotri, A. R. et al. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443–446 (2010).
84. Linheiro, R. S. & Bergman, C. M. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. *PLOS ONE* **7**, e30008 (2012).
This article presents a polymorphic TE detection method for fly genomes that showed clade-specific TSD length and enrichment of target site palindromes for TIR and LTR element insertions.
85. Nelson, M. G., Linheiro, R. S. & Bergman, C. M. McClintock: an integrated pipeline for detecting transposable element insertions in whole genome shotgun sequencing data. *G3* **7**, 2763–2778 (2017).
86. Kazazian, H. H. & Moran, J. V. The impact of L1 retrotransposons on the human genome. *Nat. Genet.* **19**, 19–24 (1998).
87. Goodier, J. L. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
88. Nakagome, M. et al. Transposon insertion finder (TIF): a novel program for detection of de novo transpositions of transposable elements. *BMC Bioinformatics* **15**, 71 (2014).
89. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
90. Wu, J. et al. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics* **15**, 795–715 (2014).
91. Platzer, A., Nizhynska, V. & Long, Q. TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology* **1**, 395–410 (2012).
92. Zhuang, J., Wang, J. & Theurkauf, W. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res.* **42**, 6826–6838 (2014).
93. Tubio, J. M. C. et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
This paper presents a method for somatic TE insertion from short sequencing reads and shows extensive L1-driven transposition and 3′ transduction in cancer genomes.
94. Helman, E. et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).
95. Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics* **16**, 768 (2015).
96. Doucet, T. T. & Kazazian, H. H. Long interspersed element sequencing (L1-Seq): a method to identify somatic LINE-1 insertions in the human genome. *Methods Mol. Biol.* **1400**, 79–93 (2016).
97. Tang, Z. et al. Human transposon insertion profiling: analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. *Proc. Natl Acad. Sci. USA* **114**, E733–E740 (2017).
98. Solyom, S. et al. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* **22**, 2328–2338 (2012).
99. Erwin, J. A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci.* **19**, 1583–1591 (2016).
100. Witherspoon, D. J. et al. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11**, 410 (2010).
101. Kvikstad, E. M., Piazza, P., Taylor, J. C. & Lunter, G. A high throughput screen for active human transposable elements. *BMC Genomics* **19**, 115 (2018).
102. Streva, V. A. et al. Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. *BMC Genomics* **16**, 220 (2015).
103. Disdero, E. & Filée, J. LoRTE: detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mob. DNA* **8**, 5 (2017).
104. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
105. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **14**, 125 (2018).
106. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
This study is a major effort to complete the human reference genome through long-read sequencing and a custom structural variant caller.
107. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
108. Ewing, A. D. Transposable element detection from whole genome sequence data. *Mob. DNA* **6**, 24 (2015).
109. Iskow, R. C. et al. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253–1261 (2010).
110. Rishishwar, L., Mariño-Ramírez, L. & Jordan, I. K. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinform.* **18**, 908–918 (2017).
111. Kofler, R. SimulaTE: simulating complex landscapes of transposable elements of populations. *Bioinformatics* **34**, 1439 (2018).
112. Navarro, F. C. & Galante, P. A. RCPedia: a database of retrocopied genes. *Bioinformatics* **29**, 1235–1237 (2013).
113. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
This article presents the RNA-seq differential expression software TEtranscripts, shown to be the most accurate at identifying reads from repetitive elements.
114. Lanciano, S. et al. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLOS Genet.* **13**, e1006630 (2017).
115. Sundaresan, V. & Freeling, M. An extrachromosomal form of the Mu transposons of maize. *Proc. Natl Acad. Sci. USA* **84**, 4924–4928 (1987).
116. Kamal, M., Xie, X. & Lander, E. S. A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl Acad. Sci. USA* **103**, 2740–2745 (2006).
117. Lowe, C. B., Bejerano, G. & Haussler, D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl Acad. Sci. USA* **104**, 8005–8010 (2007).
118. Chandrashekar, D. S., Dey, P. & Acharya, K. K. GREAM: a web server to short-list potentially important genomic repeat elements based on over-/under-representation in specific chromosomal locations, such as the gene neighborhoods, within or across 17 mammalian species. *PLOS One* **10**, e0133647 (2015).
This paper describes a tool that was developed to assess the impact of TEs on genes and biological pathways.
119. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, 583 (2014).
120. Han, B. W., Wang, W., Zamore, P. D. & Weng, Z. 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 (2015).
121. Luteijn, M. J. & Ketting, R. F. PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nat. Rev. Genet.* **14**, 523–534 (2013).
122. Lerat, E., Fablet, M., Modolo, L., Lopez-Maestre, H. & Vieira, C. 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**, e17 (2017).
123. Robberecht, C., Voet, T., Zamani Esteki, M., Nowakowska, B. A. & Vermeesch, J. R. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res.* **23**, 411–418 (2013).
124. He, D., Hormozdiari, F., Furlotte, N. & Eskin, E. Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics* **27**, 1513–1520 (2011).
125. Monlong, J. et al. Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Res.* **7**, 225 (2018).

126. Churakov, G. et al. A novel web-based TinT application and the chronology of the primate Alu retroposon activity. *BMC Evol. Biol.* **10**, 376 (2010).
127. Price, A. L., Eskin, E. & Pevzner, P. A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.* **14**, 2245–2252 (2004).
128. Jiang, C., Chen, C., Huang, Z., Liu, R. & Verdier, J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics* **16**, 72 (2015).
129. Daron, J. & Slotkin, R. K. EpiTEome: simultaneous detection of transposable element insertion sites and their DNA methylation levels. *Genome Biol.* **18**, 91 (2017).
130. Glusman, G. et al. A third approach to gene prediction suggests thousands of additional human transcribed regions. *PLOS Comput. Biol.* **2**, e18 (2006).
131. Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* **22**, R898–R899 (2012).
132. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
133. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
134. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
References 133 and 134 describe the aligners BowTie and BowTie2, which are capable of handling multi-mapped reads.
135. Thankaswamy-Kosalai, S., Sen, P. & Nookaew, I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* **109**, 186–191 (2017).
136. Kahles, A., Behr, J. & Rätsch, G. MMR: a tool for read multi-mapper resolution. *Bioinformatics* **32**, 770–772 (2016).
137. Wang, J., Huda, A., Lunyak, V. V. & Jordan, I. K. A. Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics* **26**, 2501–2508 (2010).
138. Chung, D. et al. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLOS Comput. Biol.* **7**, e1002111 (2011).
139. Wang, R. et al. LOcating non-unique matched tags (LONUT) to improve the detection of the enriched regions for ChIP-seq data. *PLOS One* **8**, e67788 (2013).
140. Nakato, R., Itoh, T. & Shirahige, K. DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. *Genes Cells* **18**, 589–601 (2013).
References 138–140 are examples of ChIP-seq peak callers developed to include multi-mapped reads in their analyses.
141. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
142. Anders, S., Pyl, P. T. & Huber, W. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
143. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
144. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
145. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500 (1985).
146. Eickbush, T. H. & Malik, H. S. in Mobile DNA II (eds. Craig N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 1111–1144 (American Society of Microbiology, Washington, DC, 2002).
147. Piégu, B., Bire, S., Arensburger, P. & Bigot, Y. A survey of transposable element classification systems — a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.* **86**, 90–109 (2015).
148. Llorens, C. et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* **39**, D70–D74 (2011).
149. Vassetzky, N. S. & Kramerov, D. A. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.* **41**, D83–D89 (2013).
150. Ma, B., Li, T., Xiang, Z. & He, N. MnTEdb, a collective resource for mulberry transposable elements. *Database* **2015**, bav004 (2015).
151. Shao, F., Wang, J., Xu, H. & Peng, Z. FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. *Database* **2018**, bax106 (2018).
152. Xu, H. E. et al. BmTEdb: a collective database of transposable elements in the silkworm genome. *Database* **2013**, bat055 (2013).
153. Li, S.-F., Zhang, G.-J., Yuan, J.-H., Deng, C.-L. & Gao, W.-J. Repetitive sequences and epigenetic modification: inseparable partners play important roles in the evolution of plant sex chromosomes. *Planta* **243**, 1083–1095 (2016).
154. Roberts, A. P. et al. Revised nomenclature for transposable genetic elements. *Plasmid* **60**, 167–173 (2008).
155. Nakagawa, S. & Takahashi, M. U. gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database* **2016**, baw087 (2016).
156. Paces, J., Pavlíček, A. & Paces, V. HERVd: database of human endogenous retroviruses. *Nucleic Acids Res.* **30**, 205–206 (2002).
157. Lappalainen, I. et al. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* **41**, D936–D941 (2013).
158. Rahman, R. et al. Unique transposon landscapes are pervasive across Drosophila melanogastergenomes. *Nucleic Acids Res.* **43**, 10655–10672 (2015).
159. Ye, C., Ji, G. & Liang, C. detectMITE: a novel approach to detect miniature inverted repeat transposable elements in genomes. *Sci. Rep.* **6**, 19688 (2016).
160. Keane, T. M., Wong, K. & Adams, D. J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**, 389–390 (2013).
161. Hormozdiari, F. et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350–i357 (2010).
162. Gilly, A. et al. TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics* **15**, 377 (2014).
163. Thung, D. T. et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol.* **15**, 488 (2014).
164. Quadrana, L. et al. The Arabidopsis thaliana mobilome and its impact at the species level. *eLife* **5**, e15716 (2016).
165. David, M., Mustafa, H. & Brudno, M. Detecting Alu insertions from high-throughput sequencing data. *Nucleic Acids Res.* **41**, e169 (2013).
166. Tica, J. et al. Next-generation sequencing-based detection of germline L1-mediated transductions. *BMC Genomics* **17**, 342 (2016).
167. Du, C., Caronna, J., He, L. & Dooner, H. K. Computational prediction and molecular confirmation of Helitron transposons in the maize genome. *BMC Genomics* **9**, 51 (2008).
168. Fiston-Lavier, A. S., Barron, M. G., Petrov, D. A. & Gonzalez, J. T-Lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res.* **43**, e22 (2015).
169. Kofler, R., Betancourt, A. J. & Schlötterer, C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLOS Genet.* **8**, e1002487 (2012).
170. Kofler, R. & Gómez-Sánchez, D. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.* **33**, 2759–2764 (2016).
171. Cridland, J. M., Macdonald, S. J., Long, A. D. & Thornton, K. R. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. *Mol. Biol. Evol.* **30**, 2311–2327 (2013).
172. Chen, J., Wrightsman, T. R., Wessler, S. R. & Stajich, J. E. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ* **5**, e2942 (2017).
173. Stuart, T. et al. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* **5**, e20777 (2016).

Acknowledgements
This work was supported by a grant from the Canadian Institute for Health Research (CIHR-MOP-115090). P.G.-P. is supported by the Programme de bourses de formation de doctorat du Fonds de Recherche Québec Santé (FRSQ-31874). G.B. is supported by the Fonds de Recherche Québec Santé (FRQS-25348). The authors also thank J.M.M. Monlong and the reviewers for very useful comments on the manuscript.

Author contributions
P.G.-P. and G.B. contributed to all aspects of the manuscript.

Competing interests
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information
Nature Reviews Genetics thanks E. Lerat, A. Smit and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

RELATED LINKS
NovoAlign: http://www.novocraft.com/products/novoalign
RepeatMasker: http://www.repeatmasker.org